РОССИЙСКАЯ АКАДЕМИЯ НАУК
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР ИМ. А. А. ДОРОДНИЦЫНА РАН

На правах рукописи

ВОРОНЦОВ КОНСТАНТИН ВЯЧЕСЛАВОВИЧ

КОМБИНАТОРНАЯ ТЕОРИЯ НАДЁЖНОСТИ ОБУЧЕНИЯ ПО ПРЕЦЕДЕНТАМ

05.13.17 — теоретические основы информатики

Диссертация на соискание ученой степени
доктора физико-математических наук

Научный консультант
чл.-корр. РАН К. В. Рудаков

Москва, 2010

Оглавление

Введение

Диссертационная работа посвящена проблемам обобщающей способности в задачах обучения по прецедентам. Предлагается комбинаторный подход, позволяющий получать точные оценки вероятности переобучения, учитывающие эффекты расслоения и связности в семействах алгоритмов.

Актуальность темы. Вопрос о качестве восстановления зависимостей по эмпирическим данным является фундаментальной проблемой *теории статистического обучения*[1] (statistical learning theory, SLT).

Основным объектом исследования в SLT является задача обучения по прецедентам: задана *обучающая выборка* пар «объект–ответ»; требуется восстановить функциональную зависимость ответов от объектов, т. е. построить алгоритм, способный выдавать адекватный ответ для произвольного объекта. К этому классу задач относятся задачи распознавания образов, классификации, восстановления регрессии, прогнозирования.

Основной задачей SLT является получение оценок вероятности ошибки построенного алгоритма на объектах, не входивших в обучающую выборку. Эта задача нетривиальна, поскольку частота ошибок на обучающей выборке, как правило, является смещённой (сильно заниженной) оценкой вероятности ошибки. Это явление называют *переобучением*. Способность алгоритмов восстанавливать неизвестную зависимость по конечной выборке данных называют *обобщающей способностью* (generalization ability).

Возникновение SLT связывают с появлением в начале 70-х годов статистической теории Вапника–Червоненкиса (далее VC-теория), которая получила широкую мировую известность и признание в середине 80-х [9, 10, 11, 8, 215]. В настоящее время SLT продолжает активно развиваться, постоянно появляются новые направления исследований и новые приложения.

Основным результатом VC-теории являются оценки, связывающие вероятность ошибки с длиной обучающей выборки и сложностью семейства функций, из которого выбирается искомый алгоритм. Согласно VC-теории, для получения надёжных алгоритмов необходимо ограничивать сложность семейства. Мерой сложности конечного семейства является его мощность. Однако на практике гораздо чаще используются

[1] Второе название — *теория вычислительного обучения* (computational learning theory, COLT). Различия между COLT и SLT, по мнению автора, незначительны и довольно условны. В частности, COLT включает в себя проблематику вычислительной эффективности алгоритмов обучения.

бесконечные семейства. Чтобы свести этот случай к конечному, вводится бинарная функция потерь. Тогда лишь конечное число алгоритмов оказываются попарно различимыми на выборке конечной длины. Зависимость максимального числа попарно различимых алгоритмов от длины выборки называется *функцией роста семейства*. В худшем случае она растёт экспоненциально, но если её рост ограничен сверху полиномом фиксированной степени, то оценки являются состоятельными — частота ошибок на обучающей выборке стремится к вероятности ошибки при стремлении длины выборки к бесконечности.

Основной проблемой VC-теории является сильная завышенность оценок вероятности ошибки. Попытка их практического применения приводит либо к требованию явно избыточного наращивания обучающей выборки, либо к переупрощению семейства алгоритмов. Наиболее интересные случаи — малых выборок и сложных семейств — находятся за границами применимости VC-теории. В частности, сложные алгоритмические композиции на практике могут обеспечивать высокое качество классификации, даже когда VC-оценка вероятности ошибки равна единице. Примерами таких конструкций являются корректные линейные и алгебраические композиции алгоритмов вычисления оценок [45, 46, 47, 48]. Нетривиальные оценки вероятности ошибки для таких композиций были получены В. Л. Матросовым в серии работ [63, 64, 65, 66, 67, 68]. Однако эти оценки также были сильно завышены, поскольку опирались на VC-теорию. Намного позже широкое распространение получили методы обучения линейных композиций — бустинг [132, 133] и бэггинг [114, 116]. Их статистические обоснования были получены в [196] с помощью техники, разработанной П. Бартлеттом в [101, 97]. Было показано, что верхние оценки вероятности ошибки не зависят от числа базовых алгоритмов в композиции, а только от сложности семейства базовых алгоритмов. Эти оценки опираются на усовершенствованный вариант VC-теории, но также не являются численно точными и дают лишь качественное обоснование линейных композиций, включая бустинг, многослойные нейронные сети и машины опорных векторов.

Основной причиной завышенности VC-оценок является их чрезмерная общность. Они справедливы для любой восстанавливаемой зависимости, любого метода обучения и любого распределения объектов в пространстве. Стало быть, они справедливы даже в «худших случаях», которые, как показывает практика, никогда не встречаются в реальных задачах. Очевидно также, что скалярная мера сложности семейства, не зависящая от решаемой задачи, содержит недостаточно информации о процессе обучения.

Дальнейшее развитие SLT шло по пути повышения точности оценок с учётом индивидуальных особенностей задач и методов обучения. Большое разнообразие исследований в SLT за последние 40 лет связано с неоднозначностью ответов на вопросы: какие именно характеристики задачи, семейства алгоритмов и метода обучения наиболее существенны, и в то же время достаточно удобны для практического оценивания и управления качеством алгоритма в процессе его обучения.

В идеале хотелось бы предсказывать вероятность ошибки примерно с той же точностью, с которой закон больших чисел предсказывает частоту выпадения орла

или решки. Однако проблемы переобучения и завышенности оценок обобщающей способности оказались гораздо более трудными, и до сих пор не имеют окончательного решения.

Основная трудность в том, что обучение — это оптимизационная процедура, которая способна аппроксимировать не только интересующую нас зависимость, но и ошибки измерения исходных данных, и погрешности модели. Величина смещения может зависеть от различных особенностей обучающей выборки и метода обучения; каких именно — до конца не ясно. Предлагалось учитывать сложность семейства алгоритмов (VC-теория), локальную сложность [199, 223, 166, 109, 110, 164, 103], устойчивость обучения [112, 113, 163], ширину зазора, разделяющего классы [156, 102, 91, 93], оценки скользящего контроля [151, 155, 144], априорную информацию о восстанавливаемой зависимости [88, 206].

Современные оценки основаны, главным образом, на теории эмпирических процессов [209, 159] и неравенствах концентрации вероятностной меры [180, 210, 172, 111, 90, 108]. Несмотря на развитость этих математических техник, они обладают рядом существенных недостатков:

— в процессе вывода верхних оценок практически невозможно проконтролировать, на каком именно шаге происходит основная потеря точности оценки; в результате трудно выделить истинные причины завышенности;
— автору не известны работы, в которых устранялись бы одновременно все причины завышенности классических VC-оценок; по всей видимости, сделать это с помощью известных техник очень трудно;
— наиболее точные на сегодняшний день результаты основаны на байесовском подходе [179, 164, 193], оставляющем значительный произвол при задании априорных распределений; задаются они, как правило, исходя из субъективных и довольно искусственных соображений, а анализ устойчивости оценок относительно априорных распределений практически никогда не производится.

Для устранения этих недостатков в данной работе предлагается слабая вероятностная аксиоматика и комбинаторный подход, позволяющий получать точные (не завышенные, не асимптотические) оценки вероятности переобучения.

Цель диссертационной работы — разработка нового математического аппарата для получения точных оценок вероятности переобучения.

Научная новизна. До сих пор вопрос о получении *точных* оценок (exact bounds) вероятности переобучения в SLT даже не ставился. Задача считалась безнадёжной, и обычно речь шла лишь о получении «слабо завышенных» оценок (tight bounds). Для получения точных оценок приходится отказываться от стандартного инструментария SLT — завышенных неравенств Маркова, Хёфдинга, Чернова, МакДиармида, Буля, и др. Комбинаторный подход требует радикального пересмотра всей теории, начиная с аксиоматики.

Впервые в SLT вводятся понятия локального эффективного коэффициента разнообразия, порождающих и запрещающих множеств объектов, *профилей* расслоения, связности, компактности, монотонности.

Методы исследования. Вместо завышенных функционалов равномерного отклонения, введённых в VC-теории и применяемых в SLT до сих пор, вводится более точный функционал *вероятности переобучения*, зависящий от задачи и метода обучения, и основанный на принципе полного скользящего контроля.

Обычно под *скользящим контролем* понимают среднюю частоту ошибок на контрольных данных, вычисленную по небольшому (например, случайному) подмножеству разбиений выборки на обучение и контроль. При *полном* скользящем контроле берётся множество *всех* разбиений. Непосредственное вычисление таких функционалов практически невозможно, поскольку число *всех* разбиений огромно. С другой стороны, удаётся показать, что для функционала вероятности переобучения справедливы те же верхние VC-оценки, что и для функционала равномерного отклонения, а предлагаемые в работе комбинаторные методы позволяют получать также и точные оценки.

Предлагаемая в данной работе комбинаторная теория надёжности эмпирических предсказаний опирается не на колмогоровскую теоретико-мерную аксиоматику, а на *слабую вероятностную аксиоматику*, основанную на единственном вероятностном допущении, что все разбиения конечной генеральной выборки на обучающую и контрольную части равновероятны. Этого допущения оказывается достаточно, чтобы получить аналог закона больших чисел, установить сходимость эмпирических распределений и воспроизвести основные результаты VC-теории. Кроме того, в слабой аксиоматике естественным образом строятся непараметрические статистические критерии и доверительные интервалы.

Применяемые в данной работе методы относятся скорее к области дискретной математики, в первую очередь комбинаторики, чем к математической статистике и теории вероятностей. В то же время, все комбинаторные результаты имеют прозрачный вероятностный смысл.

Хотя работа является теоретической, ход исследования в значительной степени определялся по результатам экспериментов на реальных и модельных задачах классификации. Эти эксперименты подробно описаны в главе 3.

Теоретическая значимость. В настоящее время в теории обобщающей способности наметилась стагнация. Ценой существенного усложнения математического аппарата удаётся добиться лишь незначительного повышения точности оценок. Интерес научного сообщества к проблематике оценок обобщающей способности заметно снизился в последние годы, сместившись к байесовской теории обучения и решению новых типов прикладных задач. Тем временем остаются открытыми фундаментальные проблемы — как преодолеть завышенность оценок, и как их использовать на практике для управления процессом обучения. Сложившаяся ситуация не раз повторялась в истории науки: очевидно, что для дальнейшего развития теории требуются радикально новые идеи и подходы. Данная работа является попыткой выхода из тупика.

Практическая значимость. Большинство оценок, полученных в данной работе, пока не нашли непосредственного практического применения, за исключением результатов главы 5. Точные оценки в большинстве случаев требуют определённой дора-

ботки и адаптации к прикладной задаче, поскольку они определяются через тонкие характеристики как самой задачи, так и применяемого к ней метода обучения. Ожидается, что одним из первых применений станет разработка новых методов поиска логических закономерностей и построения логических алгоритмов классификации.

Апробация работы. Результаты работы неоднократно докладывались на научных семинарах ВЦ РАН и на конференциях:

- всероссийская конференция «Математические методы распознавания образов» ММРО-7, 1995 г. [15];
- международная конференция «Интеллектуализация обработки информации» ИОИ-4, 2002 г. [18];
- всероссийская конференция «Математические методы распознавания образов» ММРО-11, 2003 г. [19];
- международная конференция «Интеллектуализация обработки информации» ИОИ-5, 2004 г. [23];
- всероссийская конференция «Математические методы распознавания образов» ММРО-12, 2005 г. [60];
- международная конференция «Интеллектуализация обработки информации» ИОИ-6, 2006 г. [29, 32];
- всероссийская конференция «Математические методы распознавания образов» ММРО-13, 2007 г. [25, 61, 54, 12, 81, 31];
- 7-й открытый немецко-российский семинар «Распознавание образов и понимание изображений», Эттлинген, Германия, 20–25 августа 2007 г. [219];
- ломоносовские чтения, МГУ, 17 апреля, 2008 г.;
- международная конференция «Интеллектуализация обработки информации» ИОИ-7, 2008 г. [85];
- международная конференция «Распознавание образов и анализ изображений: новые информационные технологии» РОАИ-9, Нижний Новгород, 2008 г. [220];
- международная конференция «Современные проблемы математики, механики и их приложений», посвященная 70-летию ректора МГУ академика В. А. Садовничего, Москва, 30 марта–2 апреля 2009 г.;
- семинар «Знания и онтологии ELSEWHERE 2009», ассоциированный с 17-й международной конференцией по понятийным структурам ICCS-17, Москва, Высшая школа экономики, 21–26 июля 2009 г.;
- всероссийская конференция «Математические методы распознавания образов» ММРО-14, 2009 г. [26, 53, 30].

Материалы данной диссертационной работы легли в основу спецкурса «Теория надёжности обучения по прецедентам», читаемого студентам старших курсов на факультете Вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова.

Полный текст диссертации доступен через персональную страницу автора на сайте ВЦ РАН: `http://www.ccas.ru/voron`.

Публикации по теме диссертации в изданиях списка ВАК: [76, 17, 22, 20, 21, 219, 221, 28]. Другие публикации по теме диссертации: [15, 18, 19, 24, 23, 60, 29, 32, 25, 54, 61, 12, 81, 31, 85, 220, 26, 53, 30].

Структура и объём работы. Работа состоит из оглавления, введения, пяти глав, списка основных обозначений, списка иллюстраций (34 пункта), списка таблиц (6 пунктов), списка литературы (224 пункта) и предметного указателя. Общий объём работы — 271 стр.

Краткое содержание работы по главам.

В главе 1 вводится *слабая вероятностная аксиоматика* и рассматриваются постановки задач эмпирического предсказания. Вводятся базовые технические приёмы: обращение оценок, переход от ненаблюдаемых оценок к наблюдаемым, эмпирическое оценивание вероятностей методом Монте-Карло. Обсуждается связь слабой вероятностной аксиоматики с классической колмогоровской аксиоматикой и основаниями теории вероятностей. В рамках слабой аксиоматики выводятся точные оценки надёжности эмпирических предсказаний для таких классических задач, как оценивание частоты события, оценивание функции распределения, доверительное оценивание. Большинство непараметрических статистических тестов также могут быть легко перенесены в слабую аксиоматику, что показывается на примере нескольких классических тестов. В рамках слабой аксиоматики выводятся также верхние оценки *вероятности переобучения*, аналогичные VC-оценкам. Предлагается новая оценка, учитывающая степень некорректности метода обучения, и показывается, что в случае корректности (отсутствия ошибок на обучающей выборке) вероятность переобучения может быть существенно меньше. Однако учёт корректности не устраняет ни одного из основных факторов завышенности VC-оценок. Анализ причин завышенности и получение точных оценок вероятности переобучения являются основной целью данной диссертационной работы.

В главе 2 рассматривается текущее состояние теории статистического обучения и оценок обобщающей способности.

В главе 3 описывается методика экспериментального количественного измерения факторов завышенности VC-оценок. В рамках VC-теории измерение функционала равномерной сходимости наталкивается на значительные трудности. Однако после перехода в слабую аксиоматику и замены его функционалом вероятности переобучения измерение становится возможным. Эксперимент с логическими алгоритмами классификации на реальных задачах из репозитория UCI показывает, что среди всех факторов завышенности наиболее существенны два — это игнорирование таких важных свойств семейства алгоритмов, как расслоение и связность. *Расслоение* возникает вследствие универсальности применяемых на практике семейств. Как правило, лишь ничтожная доля алгоритмов в семействе подходит для решения данной конкретной задачи. Именно эти алгоритмы имеют наиболее высокие шансы быть полученными в результате обучения. Распределение вероятностей на множестве алгоритмов существенно неравномерно, однако этот факт никак не учитывается классическими VC-оценками. *Связность* возникает вследствие непрерывности применяемых

на практике семейств. Как правило, для любого алгоритма в семействе находится большое число похожих на него алгоритмов. Однако классические VC-оценки ориентированы на «худший случай», когда все алгоритмы существенно различны, что почти невероятно встретить на практике. Второй эксперимент на модельных данных подтверждает необходимость совместного учёта эффектов расслоения и связности. Рассматривается простейшее семейство с расслоением и связностью — монотонная цепочка алгоритмов. Его естественные модификации, не обладающие либо расслоением, либо связностью, сильно переобучаются уже при нескольких десятках алгоритмов в семействе. Третий эксперимент проводится на двухэлементном семействе и показывает, что даже в этом простейшем случае появляется переобучение, а эффекты расслоения и сходства снижают вероятность переобучения.

В главе 4 предлагается несколько способов получения точных оценок вероятности переобучения. Первый способ основан на понятиях порождающих и запрещающих множеств объектов. Порождающее множество — это те объекты, которые обязательно должны присутствовать в обучающей выборке, чтобы данный алгоритм был выбран данным методом обучения. Запрещающее множество — это те объекты, которых не должно быть в обучающей выборке, чтобы данный алгоритм был выбран. Доказывается, что порождающие и запрещающие множества можно указать всегда, а коли они указаны, можно выписать точные формулы вероятности переобучения. Второй способ основан на разбиении генеральной выборки на блоки; соответствующие оценки эффективны только при малом числе алгоритмов в семействе. Третий способ основан на гипотезе, что множество векторов ошибок рассматриваемого семейства алгоритмов образует интервал булева куба. Четвёртый способ основан на рекуррентных формулах, позволяющих корректировать порождающие и запрещающие множества при добавлении в семейство ещё одного алгоритма. Доказано, что путём некоторого упрощения рекуррентной процедуры можно получать и верхние, и нижние оценки вероятности переобучения. При этом точность оценок можно обменивать на время вычислений. При самом простом варианте рекуррентной процедуры верхние оценки выписываются в явном виде. Они похожи на VC-оценки, но содержат «поправку на связность», экспоненциально убывающую с ростом размерности семейства. Вводятся новые понятия профиля расслоения и профиля связности семейства алгоритмов, и некоторые их свойства исследуются экспериментально.

В главе 5 рассматриваются оценки функционала полного скользящего контроля CCV, определяемого как средняя по всем разбиениям частота ошибок на контрольной выборке. Рассматриваются два практически важных частных случая — метод ближайшего соседа и монотонные классификаторы. В первом случае вводится понятие профиля компактности выборки, с его помощью выписывается точная формула CCV для метода ближайшего соседа. Предлагается метод отбора эталонных объектов, оптимизирующий CCV. Эксперименты показывают, что данный метод не переобучается. Во втором случае вводятся понятия верхнего и нижнего клина объекта, и на их основе определяется профиль монотонности выборки. С его помощью выписывается немного завышенная верхняя оценка CCV. Рассматриваются вопросы построения монотонных корректирующих операций путём оптимизации CCV.

Благодарности. Автор признателен своему учителю члену-корреспонденту РАН Константину Владимировичу Рудакову за внимание и интерес к работе, академику РАН Юрию Ивановичу Журавлёву за советы и поддержку, аспирантам и студентам Денису Кочедыкову, Андрею Ивахненко, Илье Решетняку, Александру Фрею, Павлу Ботову, Ивану Гузу, Максиму Иванову, Анастасии Зухба за плодотворные дискуссии, экспериментальную работу и дальнейшее развитие комбинаторного подхода.

Глава 1

Слабая вероятностная аксиоматика

В прикладных задачах анализа данных число наблюдений всегда конечно, тем не менее, широко используется понятие *вероятности*, подразумевающее предельный переход к бесконечной выборке. Известно, что при малых объёмах данных асимптотические методы теории вероятностей и математической статистики могут приводить к неточным или даже ошибочным выводам [72, 73]. Возникают вопросы: всегда ли обосновано использование инфинитарных (асимптотических) вероятностей в задачах анализа данных? Всегда ли понятие вероятности является инфинитарным?

Рассмотрим фундаментальную задачу теории вероятностей, тесно связанную с *законом больших чисел*: оценить вероятность большого отклонения частоты $\nu(S,X)$ события S на конечной выборке X от вероятности $P(S)$ данного события:

$$P_\varepsilon = \mathsf{P}\big\{|\nu(S,X) - P(S)| > \varepsilon\big\}. \tag{1.1}$$

Если вероятностная мера P неизвестна, то для вычисления вероятности события $P(S)$ необходимо провести бесконечное число наблюдений, что на практике невозможно. В результате оказывается, что вероятность большого отклонения P_ε непосредственно не может быть измерена в эксперименте как частота события $\big\{X\colon |\nu(S,X) - P(S)| > \varepsilon\big\}$, поскольку само наступление этого события не может быть точно идентифицировано.

Данная проблема не возникает, если с самого начала отказаться от употребления вероятности $P(S)$. Она определяется как предел частоты $\nu(S,X')$ события S на произвольной случайной выборке X' при $|X'| \to \infty$. В то же время, практический интерес представляет именно частота $\nu(S,X')$, как величина, непосредственно наблюдаемая в эксперименте. Изменим постановку задачи (1.1) и будем оценивать вероятность большого отклонения частот события S в двух различных выборках:

$$Q_\varepsilon = \mathsf{P}\big\{|\nu(S,X) - \nu(S,X')| > \varepsilon\big\}. \tag{1.2}$$

Если предполагать, что выборки X и X' независимы, то для определения вероятности Q_ε уже не нужно ни бесконечного числа испытаний, ни знания вероятностной

меры на исходном пространстве событий. Вероятность Q_ε является финитарной и может быть вычислена комбинаторными методами как доля разбиений объединённой выборки $X \cup X'$ на две подвыборки, при которых имеет место большое отклонение частот. Она может быть непосредственно измерена в эксперименте, так как идентификация события $\bigl\{X, X' \colon |\nu(S, X) - \nu(S, X')| > \varepsilon\bigr\}$ не вызывает затруднений.

Таким образом, вероятности $P(S)$ и P_ε в (1.1) имеют различную природу. Вероятность $P(S)$ принципиально инфинитарна — для её определения требуется либо знать вероятностную меру P, либо осуществить предельный переход $\nu(S, X') \to P(S)$ при $|X'| \to \infty$, что, как правило, невозможно сделать при решении практических задач. Вероятность P_ε также инфинитарна, но после замены $P(S)$ на частоту $\nu(S, X')$ она принимает финитарный вид Q_ε, допускающий и точное вычисление, и непосредственное эмпирическое измерение.

Приведённые соображения приводят к идее запретить на уровне аксиоматики использование инфинитарных вероятностей и «событий», которые не могут быть идентифицированы в эксперименте. Однако возможно ли при столь сильном ограничении построить содержательную теорию, включающую основные фундаментальные факты теории вероятностей, математической статистики, теории информации, теории статистического обучения, относящиеся к конечным выборкам?

Современная теория вероятностей возникла из стремления объединить в рамках единого формализма частотное понятие вероятности, берущее начало от азартных игр, и континуальное, идущее от геометрических задач, таких как задача Бюффона о вероятности попадания иглы в паркетную щель. В аксиоматике Колмогорова континуальное понятие берётся за основу как более общее. Ради этой общности в теорию вероятностей привносятся гипотезы сигма-аддитивности и измеримости — технические предположения из теории меры, имеющие довольно слабые эмпирические обоснования [3]. Однако далеко не во всех задачах, связанных со случайностью, определение вероятности как континуальной меры действительно необходимо.

Обратим внимание на высказывание А. Н. Колмогорова: «представляется важной задача освобождения всюду, где это возможно, от излишних вероятностных допущений. На независимой ценности чисто комбинаторного подхода к теории информации я неоднократно настаивал в своих лекциях» [58, стр. 252]. Один из вариантов комбинаторно-алгебраического построения теории информации предложен в книге В. Д. Гоппы [35]. Процитированное высказывание А. Н. Колмогорова в значительной степени относится и к математической статистике, поскольку она также изучает конечные выборки. Ученик А. Н. Колмогорова Ю. К. Беляев в предисловии к книге «Вероятностные методы выборочного контроля» пишет: «возникло глубокое убеждение, что в теории выборочных методов можно получить содержательные аналоги большинства основных утверждений теории вероятностей и математической статистики, которые к настоящему времени найдены в предположении взаимной независимости результатов измерений» [4, стр. 9]. Уместно привести ещё одно высказывание А. Н. Колмогорова: «Чистая математика благополучно развивается как по преимуществу наука о бесконечном... Весьма вероятно, что с развитием современной вычислительной техники будет понято, что в очень многих случаях разумно изучение

реальных явлений вести, избегая промежуточный этап их стилизации в духе представлений математики бесконечного и непрерывного, переходя прямо к дискретным моделям» [58, стр. 239].

В данной главе предлагается *слабая вероятностная аксиоматика*, в которой допускаются только финитарные вероятности. Понятие вероятности вводится без использования теории меры и без предельного перехода к выборкам бесконечной длины. Предельный переход вполне допустим и при необходимости может быть выполнен, однако он не закладывается в определение понятия вероятности. Единственное вероятностное предположение заключается в том, что объекты выборки становятся известны в случайном порядке, другими словами, что все перестановки выборки равновероятны, или что наблюдения в выборке независимы. Столь слабого вероятностного допущения оказывается достаточно, чтобы установить сходимость частот (аналог закона больших чисел), сходимость эмпирических распределений (критерий Колмогорова-Смирнова), получить многие ранговые и перестановочные критерии. Слабая аксиоматика полностью согласуется с колмогоровской, но её область применимости ограничена *задачами анализа данных*.

В данной главе с позиций слабой аксиоматики рассматриваются задачи эмпирического предсказания, проверки статистических гипотез, статистического обучения.

1.1 Основная аксиома

Пусть $\mathbb{X} = \{x_1, \ldots, x_L\}$ — фиксированное множество попарно различных объектов, называемое *генеральной выборкой*. Обозначим через S_L группу перестановок L элементов. Всевозможные перестановки элементов генеральной выборки будем обозначать через $\tau\mathbb{X}$, $\tau \in S_L$.

Аксиома 1.1 (о независимости элементов выборки). *Все $L!$ перестановок генеральной выборки $\tau\mathbb{X}$, $\tau \in S_L$, имеют одинаковые шансы реализоваться.*

Это *единственная аксиома* слабой вероятностной аксиоматики. Она позволяет определить понятие вероятности как «долю перестановок выборки».

Определение 1.1. *Пусть задан предикат $\psi\colon \mathbb{X}^L \to \{0,1\}$. Если $\psi(\tau\mathbb{X}) = 1$, то будем говорить, что событие ψ произошло на перестановке $\tau\mathbb{X}$. Вероятностью события ψ называется доля перестановок $\tau\mathbb{X}$, на которых произошло событие ψ:*

$$\mathsf{P}_\tau\, \psi(\tau\mathbb{X}) = \frac{1}{L!} \sum_{\tau \in S_L} \psi(\tau\mathbb{X}). \tag{1.3}$$

В слабой аксиоматике вероятность события зависит от состава объектов генеральной выборки $\mathbb{X}$, но не зависит от порядка их перечисления. Функция распределения и математическое ожидание также зависят от выборки.

Определение 1.2. *Пусть* $\xi\colon \mathbb{X}^L \to \mathbb{R}$ *— вещественная функция. Функцией распределения величины* ξ *на выборке* $\mathbb{X}$ *будем называть функцию* $F_\xi\colon \mathbb{R} \to [0,1]$ *вида*

$$F_\xi(z) = \mathsf{P}_\tau\big[\xi(\tau\mathbb{X}) \leqslant z\big]. \tag{1.4}$$

Определение 1.3. *Математическим ожиданием величины* $\xi\colon \mathbb{X}^L \to \mathbb{R}$ *на выборке* $\mathbb{X}$ *будем называть её среднее арифметическое по всем перестановкам* τ*:*

$$\mathsf{E}_\tau \xi(\tau\mathbb{X}) = \frac{1}{L!}\sum_{\tau\in S_L} \xi(\tau\mathbb{X}). \tag{1.5}$$

Заметим, что вероятность и матожидание формально определяются одинаково. Знаки P_τ и E_τ можно понимать как операцию среднего арифметического по всем перестановкам: $\mathsf{P}_\tau \equiv \mathsf{E}_\tau \equiv \frac{1}{L!}\sum\limits_{\tau\in S_L}$.

Вероятность как доля разбиений выборки. Рассмотрим важный частный случай, когда предикат ψ является функцией двух подвыборок: $X \subset \mathbb{X}$ длины ℓ и её дополнения $\bar{X} = \mathbb{X}\backslash X$ длины $k = L - \ell$, причём значение предиката $\psi(\mathbb{X}) = \varphi(X, \bar{X})$ не зависит от порядка элементов в подвыборках X и $\bar{X}$. Из аксиомы 1.1 следует, что все C_L^ℓ разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$ имеют равные шансы реализоваться. Следовательно, в данном случае вероятность можно определять не только как долю перестановок, но и как долю разбиений выборки $\mathbb{X}$:

$$\mathsf{P}_\tau\, \psi(\tau\mathbb{X}) = \mathsf{P}\, \varphi(X, \bar{X}) = \frac{1}{C_L^\ell}\sum_{(X,\bar{X})} \varphi(X, \bar{X}).$$

Рассмотрим более общий случай, когда предикат ψ является функцией q непересекающихся подвыборок, $\psi(\mathbb{X}) = \varphi(X^{\ell_1}, \ldots, X^{\ell_q})$, где $\ell_1, \ldots, \ell_q$ — длины подвыборок, $X^{\ell_1} \sqcup \cdots \sqcup X^{\ell_q} = \mathbb{X}$. Допустим, что значение предиката φ не зависит от порядка элементов в подвыборках. Тогда вероятность определяется и в этом случае как доля разбиений, при которых реализуется событие φ:

$$\mathsf{P}\, \varphi(X^{\ell_1}, \ldots, X^{\ell_q}) = \frac{\ell_1! \cdots \ell_q!}{L!} \sum_{(X^{\ell_1}, \ldots, X^{\ell_q})} \varphi(X^{\ell_1}, \ldots, X^{\ell_q}).$$

1.1.1 Задачи эмпирического предсказания

Задача эмпирического предсказания является одной из центральных в теории вероятностей и математической статистике. Она часто возникает в приложениях, связанных с прогнозированием и принятием решений. Неформально задача состоит в том, чтобы, получив выборку данных, предсказать определённые свойства аналогичных данных, пока ещё неизвестных, и заранее оценить точность предсказания.

Рассмотрим эксперимент, в котором реализуется одно из C_L^ℓ равновероятных разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$. После реализации разбиения наблюдателю сообщается подвыборка X. Не зная *скрытой подвыборки* $\bar{X}$, требуется предсказать значение $t = T(\bar{X}, X)$ заданной функции T, существенно зависящее от $\bar{X}$. Требуется

также оценить надёжность предсказания, то есть вероятность того, что предсказанное значение $\hat{t} = \hat{T}(X)$ будет не сильно отличаться от истинного значения t.

Рассмотрим два варианта формальной постановки задачи.

Задача 1.1. При заданной функции $T\colon \mathbb{X}^k \times \mathbb{X}^\ell \to R$ построить *предсказывающую функцию* $\hat{T}\colon \mathbb{X}^\ell \to R$, значение которой на *наблюдаемой подвыборке* $\hat{t} = \hat{T}(X)$ приближало бы неизвестное значение $t = T(\bar{X}, X)$, а также оценить надёжность предсказаний, указав невозрастающую *оценочную функцию* $\eta(\varepsilon)$ такую, что

$$\mathsf{P}\big[d\big(\hat{T}(X), T(\bar{X}, X)\big) > \varepsilon\big] \leqslant \eta(\varepsilon), \tag{1.6}$$

где $d\colon R \times R \to \mathbb{R}$ — заданная функция, характеризующая величину отклонения $d(\hat{t}, t)$ предсказанного значения $\hat{t}$ от неизвестного истинного значения t.

Обозначим эту задачу через $\mathscr{P}_1\big\langle R, d, T;\ \hat{T}\big\rangle$.

Параметр ε называется *точностью*, а величина $(1-\eta(\varepsilon))$ — *надёжностью* предсказания. Если в (1.6) достигается равенство, то $\eta(\varepsilon)$ называется *точной оценкой*.

Обычно предполагается, что $\varepsilon > 0$ и $0 < \eta < 1$. Если (1.6) выполняется при достаточно малых ε и η, то говорят, что в окрестности предсказываемого значения $t = T(\bar{X}, X)$ имеет место *концентрация вероятности* [172].

Для упрощения обозначений условимся далее опускать второй аргумент X функции $T(\bar{X}, X)$, если она зависит только от $\bar{X}$.

Как станет видно далее, во многих задачах в качестве предсказывающей функции $\hat{T}(X)$ выбирается $T(X)$. Тем не менее, роль функций T и $\hat{T}$ принципиально различна. Функция T задаётся в самой постановке задачи, тогда как предсказывающую функцию $\hat{T}$ наблюдатель имеет право выбирать по собственному усмотрению.

Задача $\mathscr{P}_1$ допускает следующее естественное обобщение.

Определение 1.4. *Семейство подмножеств $\Omega_\varepsilon(X) \subseteq R$ с параметром ε называется семейством (расширяющихся) вложенных подмножеств, если для любого $X \in \mathbb{X}^\ell$ и любых допустимых значений параметра $\varepsilon, \varepsilon'$ из $\varepsilon \leqslant \varepsilon'$ следует $\Omega_\varepsilon(X) \subseteq \Omega_{\varepsilon'}(X)$.*

Задача 1.2. При заданной функции $T\colon \mathbb{X}^k \times \mathbb{X}^\ell \to R$ построить семейство вложенных подмножеств $\Omega_\varepsilon(X) \subseteq R$ и невозрастающую оценочную функцию $\eta(\varepsilon)$, для которых выполнено неравенство

$$\mathsf{P}\big[T(\bar{X}, X) \notin \Omega_\varepsilon(X)\big] \leqslant \eta(\varepsilon).$$

Обозначим эту задачу через $\mathscr{P}_2\big\langle R, T;\ \Omega_\varepsilon\big\rangle$.

Задача $\mathscr{P}_1$ является частным случаем задачи $\mathscr{P}_2$, если в качестве семейства вложенных подмножеств взять $\Omega_\varepsilon(X) = \big\{t \in R \;\big|\; d(\hat{T}(X), t) \leqslant \varepsilon\big\}$.

Примеры задач эмпирического предсказания. Выбирая множество R, функцию T и семейство Ω_ε (или функции $\hat{T}$ и d вместо Ω_ε), можно получить многие классические задачи теории вероятностей, математической статистики, статистического обучения. Приведём основные постановки, рассматриваемые в данной работе.

Задача 1.3 (оценивание частоты события). Пусть $S \subseteq \mathbb{X}$ — некоторое множество объектов; будем называть его «событием». Введём функцию *частоты события* S на произвольной конечной выборке $U \subseteq \mathbb{X}$:

$$\nu(U) = \frac{|S \cap U|}{|U|}.$$

Требуется предсказать частоту события S на скрытой выборке $\bar{X}$ по его частоте на наблюдаемой выборке X и оценить надёжность предсказания:

$$\mathsf{P}\big[|\nu(\bar{X}) - \nu(X)| \geqslant \varepsilon\big] \leqslant \eta(\varepsilon); \tag{1.7}$$

Очевидно, данная задача есть $\mathscr{P}_1\big\langle \mathbb{R},\ |t-\hat{t}|,\ \nu(\bar{X});\ \nu(X)\big\rangle$.

Иногда (в тех случаях, когда S интерпретируется как «нежелательное событие») требуется получить одностороннюю верхнюю оценку:

$$\mathsf{P}\big[\nu(\bar{X}) - \nu(X) \geqslant \varepsilon\big] \leqslant \eta(\varepsilon). \tag{1.8}$$

Очевидно, данная задача есть $\mathscr{P}_1\big\langle \mathbb{R},\ (t-\hat{t}),\ \nu(\bar{X});\ \nu(X)\big\rangle$.

Задача 1.3 имеет фундаментальное значение для теории вероятностей и тесно связана с законом больших чисел и предельными теоремами. Она возникает и в практических приложениях, таких, как выборочный контроль качества [4].

В 1.2.2 приводятся точные оценки для (1.7) и (1.8).

Задача 1.3′ (оценивание частоты события на генеральной выборке). Требуется предсказать частоту события $S \subseteq \mathbb{X}$ на полной выборке $\mathbb{X}$ по его частоте на наблюдаемой выборке $X \subset \mathbb{X}$ и оценить надёжность предсказания:

$$\mathsf{P}\big[|\nu(\mathbb{X}) - \nu(X)| \geqslant \varepsilon\big] \leqslant \eta(\varepsilon); \tag{1.9}$$
$$\mathsf{P}\big[\ \nu(\mathbb{X}) - \nu(X)\ \geqslant \varepsilon\big] \leqslant \eta(\varepsilon). \tag{1.10}$$

Задача 1.3′ эквивалентна Задаче 1.3, поскольку $\nu(\mathbb{X}) = \frac{\ell}{L}\nu(X) + \frac{k}{L}\nu(\bar{X})$, откуда следует

$$\nu(\mathbb{X}) - \nu(X) = \tfrac{k}{L}\big(\nu(\bar{X}) - \nu(X)\big).$$

Задача 1.4 (построение доверительных интервалов). Пусть задана функция $\xi\colon \mathbb{X} \to \mathbb{R}$. Требуется построить по наблюдаемой выборке X семейство *доверительных интервалов* $\Omega_\varepsilon(X) = \big[\xi_\varepsilon^-(X), \xi_\varepsilon^+(X)\big]$, такое, что значение $\xi(\bar{x})$ на скрытом объекте $\bar{x}$ попадает в $\Omega_\varepsilon(X)$ с вероятностью не менее $1-\eta(\varepsilon)$:

$$\mathsf{P}\big[\xi(\bar{x}) \notin \Omega_\varepsilon(X)\big] \leqslant \eta(\varepsilon).$$

Очевидно, данная задача есть $\mathscr{P}_2\langle\mathbb{R},\ \xi(\bar{x});\ \Omega_\varepsilon(X)\rangle$ при единичной длине контрольной выборки, $k = |\bar{X}| = 1$.

В 1.4.1 приводятся точные оценки для данной задачи.

Задача 1.5 (оценивание функции распределения). Определим для произвольной функции $\xi\colon \mathbb{X} \to \mathbb{R}$ и произвольной конечной выборки $U \subseteq \mathbb{X}$ *эмпирическую функцию распределения* $F_\xi\colon \mathbb{R} \to [0,1]$ как долю объектов x выборки U, для которых значение $\xi(x)$ не превосходит z:

$$F_\xi(z,U) = \frac{1}{|U|}\sum_{x\in U}[\xi(x) \leqslant z].$$

Требуется предсказать максимальное отклонение функции распределения на скрытой выборке $F_\xi(z,\bar{X})$ от известной функции распределения на наблюдаемой выборке $F_\xi(z,X)$ и оценить надёжность предсказания:

$$\mathsf{P}\Big[\max_{z\in\mathbb{R}}|F_\xi(z,\bar{X}) - F_\xi(z,X)| > \varepsilon\Big] \leqslant \eta(\varepsilon). \tag{1.11}$$

Данная задача является частным случаем $\mathscr{P}_1$, если в качестве R взять множество всех неубывающих кусочно-постоянных функций, $R = \big\{F\colon \mathbb{R} \to [0,1]\big\}$, ввести на R равномерную (чебышевскую) метрику $d(\hat{t},t) = \max\limits_{z\in\mathbb{R}}|t(z) - \hat{t}(z)|$, и положить $T(\bar{X})(z) = F_\xi(z,\bar{X}),\ \hat{T}(X)(z) = F_\xi(z,X)$, где $z \in \mathbb{R}$.

Данная задача тесно связана с оцениванием скорости сходимости эмпирических распределений. На оценке (1.11) основан критерий Смирнова проверки гипотезы однородности (одинаковой распределённости) двух выборок [80, 5].

В 1.3.2 приводятся точные оценки для (1.11) и односторонних неравенств.

Задача 1.6 (оценивание вероятности переобучения). Задано множество A, элементы которого называются *алгоритмами*. Существует бинарная функция $I\colon A \times \mathbb{X} \to \{0,1\}$, называемая *индикатором ошибки*. Если $I(a,x) = 1$, то говорят, что алгоритм a допускает ошибку на объекте x.

Частотой ошибок алгоритма a на выборке $U \subseteq \mathbb{X}$ называется величина

$$\nu(a,U) = \frac{1}{|U|}\sum_{x\in U} I(a,x).$$

Методом обучения называется отображение $\mu\colon 2^{\mathbb{X}} \to A$, которое произвольной *обучающей выборке* $X \subset \mathbb{X}$ ставит в соответствие некоторый алгоритм $a = \mu X$ из A.

Переобученностью метода μ относительно пары выборок X и $\bar{X}$ называется отклонение частоты ошибок алгоритма $a = \mu X$ на скрытой *контрольной* выборке $\bar{X}$ от частоты его ошибок на наблюдаемой обучающей выборке X:

$$\delta_\mu(X,\bar{X}) = \nu(\mu X,\bar{X}) - \nu(\mu X, X).$$

Если $\delta_\mu(X,\bar{X}) \geqslant \varepsilon$ при некотором достаточно малом $\varepsilon \in (0,1)$, то говорят, что метод μ *переобучен* относительно $X,\bar{X}$.

Требуется оценить *вероятность переобучения*:

$$\mathsf{P}\big[\delta_\mu(X,\bar{X}) \geqslant \varepsilon\big] = \mathsf{P}\big[\nu(\mu X,\bar{X}) - \nu(\mu X, X) \geqslant \varepsilon\big] \leqslant \eta(\varepsilon). \tag{1.12}$$

Очевидно, данная задача есть $\mathscr{P}_1\big\langle \mathbb{R},\ (t-\hat{t}),\ \nu(\mu X,\bar{X});\ \nu(\mu X, X)\big\rangle$.

Переобучение является серьёзной проблемой в *задачах обучения по прецедентам*, включая задачи классификации, регрессии и прогнозирования [22]. Исследованию проблемы переобучения посвящена основная часть данной работы, начиная с параграфа 1.5 данной главы.

1.1.2 Обращение оценок

Пусть для функции $\varphi\colon \mathbb{X}^\ell \times \mathbb{X}^k \to \mathbb{R}$ найдена оценка вида

$$\mathsf{P}\big[\varphi(X,\bar{X}) > \varepsilon\big] \leqslant \eta(\varepsilon), \quad \varepsilon \in \mathbb{R}. \tag{1.13}$$

Задача *обращения оценки* заключается в том, чтобы по функции $\eta(\varepsilon)$ построить функцию $\varepsilon(\eta)$, при которой (1.13) переходит в эквивалентное утверждение:

$$\varphi(X,\bar{X}) \leqslant \varepsilon(\eta) \ \text{ с вероятностью, не меньшей } \ 1-\eta, \tag{1.14}$$

где η — произвольное число из $[0,1]$. Предполагается, что η достаточно близко к нулю.

Рассмотрим несколько возможных случаев.

Непрерывная верхняя оценка. Пусть для функции φ найдена оценка вида $\mathsf{P}\big[\varphi(X,\bar{X}) > \varepsilon\big] \leqslant \tilde{\eta}(\varepsilon)$, где $\tilde{\eta}(\varepsilon)$ — непрерывная строго убывающая функция.

Тогда существует *обратная* к ней функция $\varepsilon(\eta)$ такая, что $\tilde{\eta}(\varepsilon(\eta)) \equiv \eta$, и для любого $\eta \in [0,1]$ выполняется $\mathsf{P}\big[\varphi(X,\bar{X}) > \varepsilon(\eta)\big] \leqslant \tilde{\eta}(\varepsilon(\eta)) = \eta$.

Итак, (1.14) выполняется, если в качестве $\varepsilon(\eta)$ взять обратную функцию.

Заметим, что функция $\tilde{\eta}(\varepsilon)$ не может быть точной оценкой, иначе она была бы кусочно-постоянной и не могла бы быть строго убывающей.

Точная оценка, полунепрерывная справа. Допустим теперь, что для функции φ найдена точная оценка $\mathsf{P}\big[\varphi(X,\bar{X}) > \varepsilon\big] = \eta(\varepsilon)$. Тогда функция $\eta(\varepsilon)$ монотонно не возрастает, кусочно-постоянна, полунепрерывна справа и принимает конечное множество значений $H = \big\{\eta(\varepsilon)\colon \varepsilon \in \mathbb{R}\big\}$, см. рис. 1.1. Обратная к ней $\varepsilon(\eta)$ определена только при $\eta \in H$, но её можно доопределить при любом $\eta' \in \mathbb{R}$, причём двумя способами:

$$\varepsilon(\eta') = \min\{\varepsilon\colon \eta(\varepsilon) \leqslant \eta'\} = \sup\{\varepsilon\colon \eta(\varepsilon) > \eta'\}. \tag{1.15}$$

Функция $\varepsilon(\eta')$ также монотонно не возрастает, кусочно-постоянна, полунепрерывна справа. Следующая оценка справедлива при любом $\eta' \in \mathbb{R}$, обращаясь в равенство при $\eta' \in H$:

$$\mathsf{P}\big[\varphi(X,\bar{X}) > \varepsilon(\eta')\big] = \eta\big(\varepsilon(\eta')\big) = \eta\big(\min\{\varepsilon\colon \eta(\varepsilon) \leqslant \eta'\}\big) \leqslant \eta'.$$

Таким образом, если обратную функцию доопределять согласно (1.15), то выполняется (1.14).

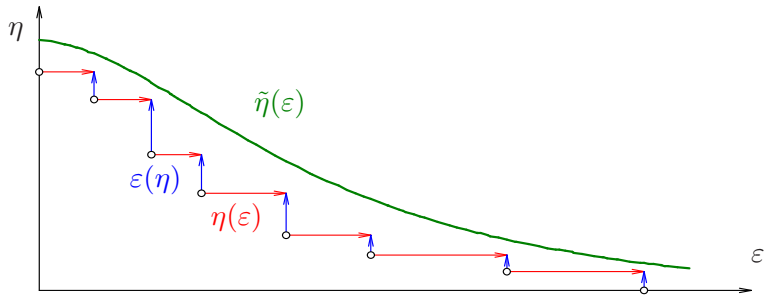


Рис. 1.1. Точная оценка, полунепрерывная *справа*, $\eta(\varepsilon)$ — красные горизонтальные линии, её обратная $\varepsilon(\eta)$ — синие вертикальные линии. Обе функции монотонно не возрастают, кусочно-постоянны, полунепрерывны *справа*. Строго убывающая непрерывная функция $\tilde\eta(\varepsilon)$ является завышенной верхней оценкой.

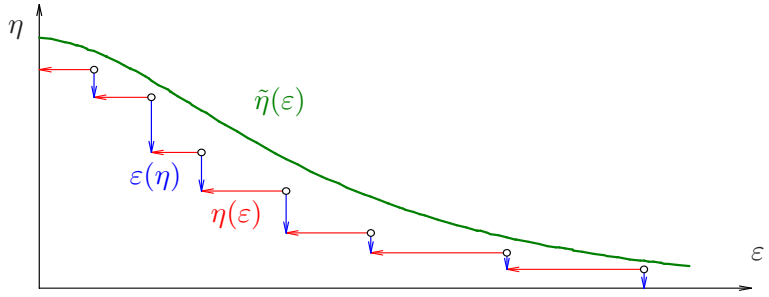


Рис. 1.2. Точная оценка, полунепрерывная *слева*, $\eta(\varepsilon)$ — красные горизонтальные линии, её обратная $\varepsilon(\eta)$ — синие вертикальные линии. Обе функции монотонно не возрастают, кусочно-постоянны, полунепрерывны *слева*. Строго убывающая непрерывная функция $\tilde\eta(\varepsilon)$ является завышенной верхней оценкой.

Точная оценка, полунепрерывная слева. Пусть теперь $\mathsf{P}\big[\varphi(X,\bar X) \geqslant \varepsilon\big] = \eta(\varepsilon)$. Тогда функция $\eta(\varepsilon)$ монотонно не возрастает, кусочно-постоянна, полунепрерывна слева и принимает конечное множество значений $H = \big\{\eta(\varepsilon)\colon \varepsilon \in \mathbb{R}\big\}$, см. рис. 1.2. Обратная к ней доопределяется одним из двух способов:

$$\varepsilon(\eta') = \max\{\varepsilon\colon \eta(\varepsilon) \geqslant \eta'\} = \inf\{\varepsilon\colon \eta(\varepsilon) < \eta'\}. \tag{1.16}$$

Функция $\varepsilon(\eta')$ также монотонно не возрастает, кусочно-постоянна, полунепрерывна слева. При любом $\eta' \in \mathbb{R}$ справедлива оценка:

$$\begin{aligned}\mathsf{P}\big[\varphi(X,\bar X) \geqslant \varepsilon(\eta')\big] &= \eta\big(\varepsilon(\eta')\big) = \eta\big(\max\{\varepsilon\colon \eta(\varepsilon) \geqslant \eta'\}\big) = \\ &= \begin{cases}\eta', & \eta' \in H;\\ \operatorname{next}_H(\eta'), & \eta' \notin H;\end{cases} \\ &\leqslant \operatorname{next}_H(\eta'),\end{aligned}$$

где $\operatorname{next}_H(\eta') = \min\big\{\eta \in H\colon \eta > \eta'\big\}$ — элемент множества H, следующий за η'.

Таким образом, если обратную функцию доопределять согласно (1.16), то вместо (1.14) выполняется

$\varphi(X,\bar X) < \varepsilon(\eta)$ с вероятностью, не меньшей $1 - \operatorname{next}_H(\eta)$.

Замечание 1.1. Из-за дополнительной операции next_H оценки, полунепрерывные слева, менее удобны, чем полунепрерывные справа. При больших длинах выборки L разностью $\mathrm{next}_H(\eta) - \eta$ можно пренебрегать как несущественной добавкой к уровню значимости. Однако при малых выборках она может быть существенной.

Замечание 1.2. Оценка, полунепрерывная справа, $\mathsf{P}\big[\varphi(X, \bar{X}) > \varepsilon\big] = \eta(\varepsilon)$, позволяет записать функцию распределения величины φ на выборке $\mathbb{X}$:

$$F_\varphi(\varepsilon) = \mathsf{P}\big[\varphi(X, \bar{X}) \leqslant \varepsilon\big] = 1 - \eta(\varepsilon).$$

Значение обратной функции $\varepsilon(\eta)$ называется *квантúлью порядка* $(1-\eta)$ для распределения величины φ.

Замечание 1.3. Непрерывные оценочные функции удобны тем, что их обращение, как правило, удаётся выполнить аналитически. Кусочно-постоянные функции задаются последовательностями точек, поэтому их обращение является вычислительной процедурой. Если мощность множества $\Phi = \{\varphi(X, \bar{X})\colon X \sqcup \bar{X} = \mathbb{X}\}$ не велика, то кусочно-постоянные функции $\eta(\varepsilon)$ и $\varepsilon(\eta)$ вычисляются довольно эффективно. Для этого множество Φ упорядочивается по возрастанию, и поиск минимального или максимального значения ε в (1.15) и (1.16) производится методом половинного деления, за $O(\log|\Phi|)$ операций. Дальнейшее повышение эффективности возможно за счёт вычисления только той относительной небольшой части множества Φ, которая соответствует достаточно малым значениям ε.

Доверительные интервалы предсказываемой величины. Пусть в задаче $\mathscr{P}_1$ функция отклонения имеет вид $d(\hat{t}, t) = t - \hat{t}$, $R = \mathbb{R}$ и имеется верхняя оценка

$$\mathsf{P}\big[T(\bar{X}, X) - \hat{T}(X) > \varepsilon\big] \leqslant \eta(\varepsilon). \tag{1.17}$$

Тогда справедлива верхняя оценка (доверительный полуинтервал) для предсказываемой величины $T(\bar{X}, X)$: для любого $\eta \in [0, 1]$ с вероятностью не менее $1 - \eta$

$$T(\bar{X}, X) \leqslant \hat{T}(X) + \varepsilon(\eta),$$

где $\varepsilon(\eta)$ — функция, обратная к $\eta(\varepsilon)$.

При функции отклонения вида $d(\hat{t}, t) = |t - \hat{t}|$ аналогичным образом выписывается двусторонняя оценка (доверительный интервал) для предсказываемой величины $T(\bar{X}, X)$: для любого $\eta \in [0, 1]$ с вероятностью не менее $1 - \eta$

$$\hat{T}(X) - \varepsilon(\eta) \leqslant T(\bar{X}, X) \leqslant \hat{T}(X) + \varepsilon(\eta).$$

Подчеркнём, что эти оценки справедливы не для всех разбиений, а только для достаточно большой их доли, то есть с вероятностью, близкой к 1. Задача оценивания надёжности эмпирических предсказаний как раз и состоит в том, чтобы находить условия, при которых значения η и ε одновременно достаточно малы.

1.1.3 Наблюдаемые и ненаблюдаемые оценки

При получении оценок вида (1.17) часто оказывается, что точная оценочная функция существенно зависит от всей выборки, $\eta(\varepsilon) = \eta(\varepsilon, \mathbb{X})$. Такие оценки называются *ненаблюдаемыми* (unobservable bound) [167, 164]. Их невозможно непосредственно использовать в задачах эмпирического предсказания, так как скрытая часть выборки в момент предсказания неизвестна.

Ниже рассматривается достаточно общий технический приём, позволяющий переходить от ненаблюдаемых оценок к *наблюдаемым* (observable bound).

Переход от ненаблюдаемой оценки к наблюдаемой. Рассмотрим случай, когда оценочная функция зависит от некоторой статистики $z(\mathbb{X})$ генеральной выборки: $\eta(\varepsilon, \mathbb{X}) = \eta(\varepsilon, z(\mathbb{X}))$. Функцию $z(\mathbb{X})$ невозможно вычислить, не зная скрытой части выборки. Возможны несколько путей решения этой проблемы.

1. Заменить точное выражение $\eta(\varepsilon, z(\mathbb{X}))$ его верхней оценкой $\max\limits_{z} \eta(\varepsilon, z)$, справедливой для любой выборки $\mathbb{X}$. Это «оценка худшего случая», и она может оказаться сильно завышенной.

2. Более тонкие результаты даёт оценивание значения $z(\mathbb{X})$ по значению этой же статистики z на наблюдаемой выборке $z(X)$. Допустим, в Задаче 1.1 получена верхняя оценка $\mathsf{P}\big[T - \hat{T} > \varepsilon\big] \leqslant \eta_T(\varepsilon, z(\mathbb{X}))$, и функция $\varepsilon_T(\eta, z)$ является обратной для $\eta_T(\varepsilon, z)$ при любом значении параметра z. Тогда с надёжностью не менее $1 - \eta_1$ справедлива верхняя оценка

$$T \leqslant \hat{T} + \varepsilon_T\big(\eta_1, z(\mathbb{X})\big).$$

Вторым шагом необходимо получить оценку для $z(\mathbb{X})$ через $z(X)$. Допустим, что функция $\varepsilon_T(\eta, z)$ монотонно не убывает по второму аргументу z, и что для z имеется оценка сверху:

$$\mathsf{P}\big[z(\mathbb{X}) - z(X) > \varepsilon\big] \leqslant \eta_z(\varepsilon).$$

Тогда с надёжностью не менее $1 - \eta_2$ справедлива оценка сверху для $z(\mathbb{X})$:

$$z(\mathbb{X}) \leqslant z(X) + \varepsilon_z(\eta_2),$$

где $\varepsilon_z(\eta_2)$ — функция, обратная для $\eta_z(\varepsilon)$. В итоге, с надёжностью не менее $1 - \eta_1 - \eta_2$ справедлива оценка сверху для T:

$$T \leqslant \hat{T} + \varepsilon_T\big(\eta_1, z(X) + \varepsilon_z(\eta_2)\big).$$

Теперь правая часть зависит только от наблюдаемой выборки. Её можно минимизировать по параметрам η_1 и η_2 при заданном $\eta = \eta_1 + \eta_2$, или для простоты положить $\eta_1 = \eta_2 = \eta/2$.

Замечание 1.4. Если статистика z выражается через статистику T, то можно обойтись без введения второго параметра η_2. Такой случай будет рассмотрен в 1.2.3.

Связь задач эмпирического предсказания и проверки гипотезы однородности. В математической статистике разработано большое количество критериев для выявления значимых отличий между двумя выборками X и $\bar{X}$, а в общем случае — между несколькими выборками [56].

В рамках сильной аксиоматики выдвигается *нулевая гипотеза* об однородности выборок: «две выборки X и $\bar{X}$ случайны, независимы и получены из одного распределения». Затем для заданной *статистики* $\varphi(X, \bar{X})$ при условии истинности нулевой гипотезы выводится функция распределения $F_\varphi(\varepsilon) = \mathsf{P}_{X,\bar{X}}\big[\varphi(X, \bar{X}) \leqslant \varepsilon\big]$. Нулевая гипотеза отвергается, если оказывается, что наблюдаемое значение $\varphi(X, \bar{X})$ попадает в *критическую область* маловероятных значений статистики φ — как правило, слишком больших или слишком малых.

В слабой аксиоматике проверка гипотез осуществляется аналогичным образом, с той лишь разницей, что высказывание «выборки получены из одного распределения» теперь некорректно. *Гипотеза однородности* формулируется так: «выборки X и $\bar{X}$ получены в результате реализации одного из C_L^ℓ равновероятных разбиений генеральной выборки $\mathbb{X}$». При условии истинности нулевой гипотезы выводится функция распределения $F_\varphi(\varepsilon) = \mathsf{P}\big[\varphi(X, \bar{X}) \leqslant \varepsilon\big]$. Нулевая гипотеза отвергается на *уровне значимости* η, если для заданной пары выборок $(X, \bar{X})$ выполняется условие $\varphi(X, \bar{X}) > \varepsilon(\eta)$, где $\varepsilon(\eta)$ — квантиль порядка $1 - \eta$ для распределения $F_\varphi(\varepsilon)$.

Нетрудно заметить, что одни и те же функции распределения $F_\varphi(\varepsilon)$ могут быть использованы как для оценивания надёжности эмпирических предсказаний, так и для проверки однородности. В частности, оценка (1.7) может быть использована для проверки однородности при произвольном заранее заданном событии S.

Существует принципиальное отличие между этими двумя важными классами задач анализа данных. При проверке однородности нет скрытой выборки; обе выборки являются наблюдаемыми. Поэтому оценочная функция $\eta(\varepsilon)$ имеет право зависеть (и, как правило, зависит) от всех объектов из $\mathbb{X}$, что не вызывает никаких затруднений. В задачах эмпирического предсказания ситуация существенно сложнее — оценочная функция не должна зависеть от скрытой части выборки. Это требует дополнительных усилий по переходу от ненаблюдаемой оценки к наблюдаемой, который может сопровождаться как вычислительными затратами, так и некоторой потерей точности оценки.

1.1.4 Эмпирическое оценивание вероятности

В слабой аксиоматике вероятность определяется как «доля *всех* разбиений выборки», и потому может быть легко измерена эмпирически как «доля разбиений из *выбранного подмножества* разбиений». Если выбор разбиений осуществляется случайным образом, то можно говорить о применении *метода Монте-Карло* для эмпирического оценивания вероятности.

Пусть задан предикат $\varphi\colon \mathbb{X}^\ell \times \mathbb{X}^k \to \{0,1\}$. Вероятность $Q = \mathsf{P}\,\varphi(X,\bar{X})$ есть среднее значение φ по множеству всех C_L^ℓ разбиений:

$$Q = \frac{1}{C_L^\ell} \sum_{(X,\bar{X})} \varphi(X,\bar{X}).$$

Непосредственное вычисление величины Q по этой формуле практически осуществимо только при небольших значениях ℓ или k. В типичных случаях число разбиений C_L^ℓ огромно. Рассмотрим приближённую оценку Q как среднее по некоторому подмножеству разбиений N, не слишком большому, чтобы сумма вычислялась за приемлемое время:

$$\hat{Q}_N = \hat{\mathsf{P}}\varphi(X,\bar{X}) = \frac{1}{|N|} \sum_{(X,\bar{X})\in N} \varphi(X,\bar{X}).$$

Метод Монте-Карло. Выберем случайное подмножество разбиений N из равномерного распределения на множестве всех $C_{C_L^\ell}^{|N|}$ подмножеств мощности $|N|$. Тогда задача оценивания точности приближения $\hat{Q}_N \approx Q$ сводится к Задаче 1.3′ об оценивании частоты события на генеральной выборке, только теперь в качестве генеральной выборки рассматривается множество всех разбиений.

В параграфе 1.2 для функционала (1.10) в Задаче 1.3′ будет получена точная оценка. Сейчас предположим, что имеется верхняя оценка

$$\mathsf{P}_N\big\{Q - \hat{Q}_N > \tilde{\varepsilon}\big\} \leqslant \tilde{\eta}(\tilde{\varepsilon}).$$

Обращая эту оценку, получаем, что с надёжностью не менее $1-\tilde{\eta}$ выполняется неравенство $Q \leqslant \hat{Q}_N + \tilde{\varepsilon}(\tilde{\eta})$, где $\tilde{\varepsilon}(\tilde{\eta})$ — обратная функция для $\tilde{\eta}(\tilde{\varepsilon})$.

Конкретизируем вид предиката $\varphi(X,\bar{X})$. Рассмотрим задачу эмпирического предсказания величины $T(\bar{X},X)$, в которой $\varphi(X,\bar{X}) = [T(\bar{X},X) - \hat{T}(X) > \varepsilon]$. Тогда величины $Q(\varepsilon)$ и $\hat{Q}_{N'}(\varepsilon)$ являются кусочно-постоянными невозрастающими функциями параметра ε. С надёжностью не менее $1-\tilde{\eta}$ выполняется неравенство

$$\mathsf{P}\big[T - \hat{T} > \varepsilon\big] = Q(\varepsilon) \leqslant \hat{Q}_N(\varepsilon) + \tilde{\varepsilon}(\tilde{\eta}).$$

Ещё раз применяя обращение, заключаем, что с надёжностью не менее $(1-\tilde{\eta}-\eta)$ справедлива верхняя оценка для $T(\bar{X},X)$:

$$T(\bar{X},X) \leqslant \hat{T}(X) + \varepsilon(\eta), \tag{1.18}$$

где $\varepsilon(\eta)$ — обратная функция для $\eta(\varepsilon) = \hat{Q}_N(\varepsilon) + \tilde{\varepsilon}(\tilde{\eta})$.

Таким образом, вычисляя значения $T(\bar{X},X)$ и $\hat{T}(X)$ по относительно небольшому подмножеству разбиений, $(\bar{X},X) \in N$, можно получить верхнюю границу, которую $T(\bar{X},X)$ не превосходит для любого разбиения $(\bar{X},X)$, с заданной надёжностью.

Описанный способ эмпирического оценивания вероятности имеет несколько существенных недостатков. Во-первых, он даёт лишь приближённые оценки.

Во-вторых, он требует знания генеральной выборки $\mathbb{X}$, и потому не может быть использован непосредственно для эмпирического предсказания. В-третьих, он не позволяет получать оценки в аналитическом виде. Наконец, он может потребовать большого объёма вычислений.

Таким образом, область применимости эмпирического оценивания довольно ограничена. На практике его можно использовать для предварительного экспериментального исследования зависимости Q от некоторых параметров задачи (например, от длины выборки).

В задачах обучения по прецедентам эмпирическое оценивание называют *скользящим контролем* (cross-validation) и используют для оценивания качества метода обучения μ, а не отдельного алгоритма $a = \mu X$, полученного в результате обучения. Скользящий контроль незаменим в тех случаях, когда теоретические верхние оценки вероятности Q не известны или сильно завышены. В главе 3 данной работы эмпирическое оценивание применяется для анализа точности теоретических оценок и понимания причин их завышенности.

1.1.5 Замечания и интерпретации

О связи с сильной вероятностной аксиоматикой. Классическая теоретико-мерная аксиоматика А. Н. Колмогорова (будем называть её сильной) основана на понятии вероятностного пространства $\langle \mathscr{X}, \Omega, \mathsf{P} \rangle$, где $\mathscr{X}$ — множество допустимых объектов, Ω — аддитивная σ-алгебра событий на $\mathscr{X}$, P — вероятностная мера, определённая на событиях из Ω. Во многих задачах статистического анализа данных предполагается, что исходные данные $\mathbb{X} = \{x_1, \ldots, x_L\}$ представляют собой *простую выборку*, то есть конечное множество объектов, выбранных из множества $\mathscr{X}$ случайно и независимо согласно вероятностной мере P. В реальных приложениях множество $\mathscr{X}$, как правило, бесконечно, а мера P — неизвестна.

В слабой аксиоматике множество $\mathscr{X}$ не вводится. Рассматривается только конечное множество объектов — *генеральная выборка* $\mathbb{X}$. Оно может включать в себя как объекты, наблюдавшиеся ранее, так и скрытые объекты, которые станут известны в будущем. Вероятностным пространством является конечное множество всех перестановок генеральной выборки $\mathbb{X}$, на котором задаётся равномерное распределение. Таким образом, случайными полагаются не сами объекты, а лишь порядок их появления, что соответствует предположению о независимости объектов выборки в сильной аксиоматике.

Следующая теорема утверждает, что для перевода оценки из слабой аксиоматики в сильную достаточно применить к ней операцию математического ожидания.

Теорема 1.1. *Пусть в слабой аксиоматике найдено значение вероятности*

$$\mathsf{P}_\tau \, \psi(\tau \mathbb{X}) = f(\mathbb{X}). \tag{1.19}$$

Тогда в сильной аксиоматике выполняется равенство

$$\mathsf{P}_{\mathbb{X}} \, \psi(\mathbb{X}) = \mathsf{E}_{\mathbb{X}} f(\mathbb{X}). \tag{1.20}$$

Доказательство. В силу независимости наблюдений в выборке $\mathbb{X}$ для произвольной перестановки τ справедливо равенство $\mathsf{P}_{\mathbb{X}}\,\psi(\mathbb{X}) = \mathsf{P}_{\mathbb{X}}\,\psi(\tau\mathbb{X})$. Следовательно,

$$\mathsf{P}_{\mathbb{X}}\,\psi(\mathbb{X}) = \mathsf{E}_{\mathbb{X}}\psi(\tau\mathbb{X}) = \mathsf{E}_{\mathbb{X}}\,\mathsf{P}_{\tau}\,\psi(\tau\mathbb{X}) = \mathsf{E}_{\mathbb{X}}f(\mathbb{X}),$$

что и требовалось доказать. ■

В случаях, когда оценка $f(\mathbb{X})$ не зависит от выборки $\mathbb{X}$, конечный результат — оценка в правой части (1.19) и (1.20) — будет одинаков в обеих аксиоматиках.

Если же оценка имеет вид $f(\mathbb{X}) = f(S(\mathbb{X}))$, где S — некоторая функция (статистика) полной выборки, то возможны несколько вариантов дальнейших действий.

1. *«What-if анализ»*: значение статистики $S = S(\mathbb{X})$ интерпретируется как априори задаваемый параметр, и окончательный результат представляется в виде зависимости оценки f от значения S.

2. Строится *оценка худшего случая* $f(\mathbb{X}) \leqslant \max\limits_{S} f(S)$, которая не зависит от выборки, но может оказаться сильно завышенной.

3. Строится *доверительный интервал* для ненаблюдаемого значения статистики $S(\mathbb{X})$ по значению той же статистики на наблюдаемой выборке $S(X)$, см. 1.1.3. Затем доверительный интервал для $S(\mathbb{X})$ переводится в доверительный интервал для $f(S)$. Данный подход даёт наиболее точные результаты.

Во всех перечисленных случаях вид оценки не меняется при переходе от слабой аксиоматики к сильной. Фактически, этот переход связан только с заменой функционала и его неформальных интерпретаций, но никак не влияет на получаемую оценку. Поэтому далее этот переход будет опускаться, и все результаты будут формулироваться в рамках слабой аксиоматики.

Об асимптотических оценках. Бесконечно длинные выборки не реализуются в практических задачах, просто потому, что конечна память компьютеров и время, отпущенное исследователям на эксперименты. В классической вероятностной аксиоматике данное обстоятельство не принимается во внимание, о чём свидетельствует привычность записи

$$P(S) = \lim_{|X|\to\infty} \nu(S, X),$$

где $P(S)$ — вероятность события S, $\nu(S, X)$ — частота события S в выборке X.

В слабой аксиоматике запись $|X| \to \infty$ запрещена, и понятие вероятности события $P(S)$ не определено. Мы не вправе предполагать, что выборка реальных объектов может быть сколь угодно длинной. Тем не менее, было бы нелепо отказываться от преимуществ и богатого математического аппарата асимптотического анализа.

Простой компромисс заключается в том, чтобы разрешить асимптотический анализ получаемых *численных оценок*, рассматривая его лишь как способ приближённых вычислений. Например, получив в слабой аксиоматике оценку, зависящую от длины выборки, $\mathsf{P}_{\tau}\,\psi(\tau\mathbb{X}) = f(L)$, мы можем исследовать асимптотическое поведение числовой функции $f(L)$ при $L \to \infty$. Очевидно, при этом нет необходимости предполагать существование выборки сколь угодно большой длины.

О частотных подходах в основаниях теории вероятностей. Предлагаемая в данной работе слабая вероятностная аксиоматика отличается не только от теоретико-мерной аксиоматики Колмогорова (являясь её специальным частным случаем), но и от известных частотных подходов к определению понятия вероятности.

Частотный подход фон Мизеса [218] является инфинитарным. Его основная цель — определение фундаментального понятия «вероятности» как предела частоты. Его основная проблема, которую до сих пор не удаётся разрешить до конца — необходимость строгой формализации понятия *иррегулярной* (т. е. бесконечной случайной) последовательности [86, 82]. В отличие от подхода фон Мизеса, в слабой аксиоматике рассматриваются только конечные последовательности.

А. Н. Колмогоров был убеждён, что «частотный подход, основанный на понятии *предельной частоты* при стремящемся к бесконечности числе испытаний, не позволяет обосновать применимость результатов теории вероятностей к практическим задачам, в которых мы имеем дело с конечным числом испытаний» [58, стр. 205]. Начиная с 1965 г., Колмогоров развивал *финитарную теорию алгоритмической случайности* [57]. В этом подходе конечная последовательность считается случайной, если длина её кратчайшего описания, называемая также *колмогоровской сложностью*, не сильно отличается от максимально возможного значения, равного $\log_2 |A|$. Здесь A — это конечное множество всевозможных последовательностей. Например, $|A| = C_L^\ell$ для бернуллиевских последовательностей — двоичных последовательностей длины L, состоящих из ℓ единиц и $k = L - \ell$ нулей. Предполагается, что при разумных определениях множества A доля простых (значит, неслучайных) последовательностей крайне мала, и подавляющее большинство последовательностей являются сложными (значит, случайными). Затем строится *сложностная модель* [14]: на конечном множестве всех случайных последовательностей $A' \subset A$ вводится некоторое естественное распределение вероятностей, как правило, равномерное. В ряде важных частных случаев сложностные модели и стандартные *статистические модели* приводят к одинаковым результатам, см. ссылки на работы Е. А. Асарина в [14]. С помощью сложностных моделей возможно делать эмпирические предсказания, оценивать доверительные интервалы, проверять статистические гипотезы, и при этом вообще не вводить понятие вероятности.

Слабая аксиоматика наиболее близка к финитарной теории Колмогорова. Заметим, что между множеством всех C_L^ℓ разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$ и множеством всех бернуллиевских последовательностей $A = (a_1, \ldots, a_L)$ существует взаимнооднозначное соответствие: $a_i = [x_i \in \bar{X}]$, $i = 1, \ldots, L$.

Однако имеется и принципиальное отличие: в слабой аксиоматике равномерное распределение вводится не на множестве случайных последовательностей A', а на множестве всех последовательностей A. Таким образом, в нашей модели неслучайное по Колмогорову разбиение имеет точно такие же шансы реализоваться, как и случайное. Если доля неслучайных разбиений крайне мала, то шансы получить какое-либо из неслучайных разбиений также крайне малы, хотя и не равны нулю. В этом случае количественные результаты, полученные в слабой аксиоматике и в финитарной теории Колмогорова, должны быть очень близки.

Вопросы о том, давать ли неслучайным разбиениям равные шансы или нулевые, а также, при какой величине *дефекта случайности* [14] разбиение должно считаться неслучайным, решаются априори, за пределами математической теории.

В отличие от подходов фон Мизеса и Колмогорова, в слабой аксиоматике не предпринимается никаких попыток определить понятие случайной последовательности. Проблема в том, что все известные критерии случайности неконструктивны и слишком тяжелы для практической реализации. Например, длину кратчайшего описания возможно оценить только приближённо, применяя какой-либо алгоритм сжатия данных. Простой способ обойти эту проблему — считать, что почти все разбиения выборки, за исключением малой доли η, случайны. При этом ничего не утверждается о том, *какие именно* разбиения неслучайны. Априори задаётся только величина η — *уровень значимости* или *надёжность*, с которой будут справедливы статистические выводы. Если доля неслучайных разбиений $\frac{|A\setminus A'|}{|A|}$ мала, то выводы, справедливые для большинства разбиений из A, будут также справедливы и для большинства случайных (по Колмогорову) разбиений из A'. Качество выводов на оставшейся малой доле разбиений может быть сколь угодно плохим, однако *именно эти разбиения можно считать неслучайными для данной задачи.*

Таким образом, слабая аксиоматика фактически представляет собой компромиссное упрощение финитарной теории Колмогорова, не требующее оценивания колмогоровских сложностей и дефектов случайности, и тем самым лучше приспособленное для непосредственного практического применения в задачах анализа данных.

Об уровнях значимости. Обычно уровень значимости устанавливается экспертом, исходя из субъективных представлений о том, какой должна быть «вероятность маловероятного события». При такой интерпретации возникает искушение установить уровень значимости поменьше, особенно в тех приложениях, где требуется высокая надёжность. В слабой аксиоматике понимание, почему этого не стоит делать, возникает благодаря естественной интерпретация уровня значимости как *доли разбиений (в общем случае — перестановок) выборки, которые нельзя считать случайными.*

В Задаче 1.3 о предсказании частоты события S неслучайными являются, в частности, те разбиения, при формировании которых используется информация о самом событии S. Например, когда в скрытую выборку преднамеренно включаются только элементы из S. Очевидно, это приводит к нарушению основной Аксиомы 1.1, поскольку разбиения перестают быть равновероятными.

В Задаче 1.6 (обучения по прецедентам) неслучайными можно считать разбиения, при формировании которых используется информация об алгоритмах. Например, фиксируется некоторый алгоритм $a_0 \in A$, и в скрытую (контрольную) выборку преднамеренно включаются все объекты, на которых данный алгоритм $a_0 \in A$ допускает ошибку. Другой пример: пусть $\mathbb{X}$ — это точки линейного векторного пространства, заранее разделённого некоторой гиперплоскостью на два полупространства; в наблюдаемую (обучающую) выборку преднамеренно включаются только объекты из первого полупространства, а в скрытую — только из второго. В подобных случаях алгоритм, лучший на обучающей выборке, может оказаться сколь угодно плохим

на контрольной выборке в силу того, что две выборки, взятые из разных областей пространства, могут практически не нести информации друг о друге.

Заметим, что в примере с гиперплоскостью доля неслучайных разбиений возрастает с ростом размерности пространства $\mathbb{X}$. Стало быть, понятие «неслучайного разбиения» может зависеть от особенностей конкретной задачи. Возможно, именно по этой причине в частотной теории фон Мизеса так и не удалось выработать окончательного определения иррегулярной последовательности.

В общем случае неизвестно, какие именно разбиения являются неслучайными, однако можно считать, что их доля невелика, и априори оценивать её уровнем значимости η. Можно следовать общепринятой практике и назначать уровень значимости эвристически, в частности, полагать $\eta = 0.05$. Более обоснованно уровень значимости можно было бы определять путём комбинаторного подсчёта доли разбиений выборки, которые в данной конкретной задаче нельзя считать случайными. Однако при попытке практической реализации этой идеи возникают те же трудности, что и в подходах фон Мизеса и Колмогорова.

Нет никакого смысла устанавливать уровень значимости существенно ниже доли неслучайных разбиений. Волне допустимо, чтобы функция $\hat{T}(X)$ давала неточные предсказания именно тогда, когда разбиение не случайно (преднамеренно) выбрано неудачно. Таким образом, искусственное занижение уровня значимости η понижает точность ε эмпирических предсказаний.

О трансдукции. Предсказание некоторого свойства выборки на основании свойств другой выборки называется *трансдукцией* или переходом от частного к частному. Принято считать, что трансдукция более примитивна и ограничена, чем *индукция* — переход от частного к общему. В нашем случае это не совсем так. Допустим, в правой части (1.6) удалось получить оценку $\eta(\varepsilon)$, не зависящую от генеральной выборки $\mathbb{X}$. Тогда с помощью техники обращения выводится оценка и для $T(\bar{X}, X)$. Поскольку она будет справедлива *для любой* скрытой выборки $\bar{X}$, трансдукция в данном случае становится не менее общей, чем индукция.

Заметим, что в машинном обучении под *трансдуктивным обучением* (transductive learning) имеется в виду совсем другая постановка задачи — там это специальный тип задач классификации, в которых объекты контрольной выборки $\bar{X}$ известны ещё до решения задачи, а неизвестными являются только ответы на этих объектах [215].

О скользящем контроле. В Задаче 1.6 (обучения по прецедентам) эмпирическое оценивание $\hat{Q}_N \approx Q$ по подмножеству разбиений N принято называть *скользящим контролем* или *кросс-проверкой* (cross-validation, CV). В зависимости от способа формирования подмножества разбиений N различают несколько разновидностей скользящего контроля [157].

Если берётся одно случайное разбиение, $|N| = 1$, говорят об *оценке по отдельной тестовой выборке* (hold-out estimate).

Если берутся все разбиения при длине контрольной выборки $k = 1$, то говорят об *оценке с отделением объектов по одному* (leave-one-out estimate, LOO).

Если используются все разбиения с контрольной выборкой фиксированной, но не обязательно единичной, длины, то говорят об *оценке полного скользящего контроля* (complete cross-validation) [182].

Если производится случайная независимая выборка L объектов из $\mathbb{X}$ с возвращениями при фиксированной длине контроля k, то говорят о *бутстреп-оценке* (bootstrap estimate) [87].

Если множество разбиений N образуется q непересекающимися контрольными выборками, в объединении дающими генеральную выборку $\mathbb{X}$, то говорят о *q-кратном скользящем контроле* (q-fold cross-validation).

Если множество разбиений N образуется t случайными разбиениями на q непересекающихся контрольных выборок, каждое их которых в объединении даёт генеральную выборку $\mathbb{X}$, то говорят о *$t\times q$-кратном скользящем контроле* ($t\times q$-fold cross-validation).

Современные методики скользящего контроля, такие как $t\times q$-fold CV, стремятся уменьшить дисперсию оценки и вычислительные затраты, выбирая разбиения «более равномерно», а не просто случайно и независимо, как в методе Монте-Карло.

В прикладной статистике скользящий контроль активно применяется, начиная с работ Б. Эфрона [127, 87], и относится к «нетрадиционным методам» многомерного статистического анализа. Теоретических исследований скользящего контроля относительно немного. Показательно, что в современном издании энциклопедического словаря «Вероятность и математическая статистика» [13] даже нет статьи о скользящем контроле.

В машинном обучении скользящий контроль признан де факто стандартной методикой эмпирического оценивания *обобщающей способности* (generalization ability), сравнения и выбора методов обучения [157].

В данной работе *полный скользящий контроль* принимается, по сути дела, за определение обобщающей способности. Более того, в слабой аксиоматике этот же принцип закладывается в само понятие вероятности. По всей видимости, скользящий контроль имеет фундаментальное значение, которое в настоящее время ещё до конца не осознано исследователями.

Преимущества слабой аксиоматики проявляются в задачах анализа данных. Ещё раз резюмируем основные из них.

1. В задачах анализа данных выборки всегда конечны, будь то уже известные наблюдаемые данные или скрытые данные, которые станут известны в будущем. В некоторых практических задачах число предсказаний k настолько мало́, что оценивать *вероятность ошибки*, которая есть предел частоты ошибок при $k \to \infty$, просто некорректно. В слабой аксиоматике рассматриваются только *статистики* — функции от конечных выборок.

2. Некоторые предположения сильной аксиоматики имеют недостаточное эмпирическое обоснование. Для проверки случайности, независимости и одинаковой распределённости выработаны специальные статистические тесты. Однако гипотеза о существовании адекватной σ-аддитивной меры P на множестве объектов $\mathscr{X}$

эмпирической проверке не поддаётся [3]. Слабая аксиоматика обходится без теории меры, предъявляя существенно более скромные требования к пространству объектов и исходным данным. Об объектах вне конечной выборки $\mathbb{X}$, которых мы никогда не увидим в эксперименте, вообще не делается никаких предположений.

3. Вероятности вида (1.3) легко измеряются эмпирически по подмножеству разбиений, в частности, методом Монте-Карло. Поэтому результаты, полученные в слабой аксиоматике, всегда могут быть проверены экспериментально. В задачах статистического обучения такая проверка реализуется скользящим контролем.

Недостатки слабой аксиоматики связаны, главным образом, с наложением запретов на применение многих классических методов и приёмов теории вероятностей, а отчасти и на употребление самого понятия «вероятность».

1. Нельзя сказать «*вероятность ошибки*», но можно сказать «вероятность того, что число ошибок на скрытой выборке превысит n». К этому трудно привыкать.

2. В отличие от асимптотических оценок, точные оценки часто представляют собой сложные комбинаторные формулы, требующие значительных объёмов вычислений. Упрощение или асимптотический анализ этих формул требует дополнительных математических усилий. Сильная аксиоматика допускает бóльшую свободу в выборе математических техник, поскольку в любой момент можно воспользоваться одним из многочисленных асимптотических приёмов (интегрирование по неизвестной вероятностной мере, применение неравенств Маркова, Чебышёва, изопериметрических неравенств и др.). Слабая аксиоматика диктует более жёсткую схему вывода оценок: сначала оценка получается в виде точной комбинаторной формулы; затем, при необходимости, решаются задачи её упрощения, эффективного вычисления или асимптотического приближения.

3. Многие континуальные вероятностные модели в физике, биологии, экономике и других естественных науках существенно опираются на асимптотическое понятие вероятности. Привлечение слабой аксиоматики в эти области вряд ли целесообразно, и приведёт лишь к усложнению математического аппарата.

О границах применимости слабой аксиоматики. Подсчёт доли разбиений (или перестановок) выборки, как технический приём, широко применяется в теории вероятностей, математической статистике, теории статистического обучения. Полученные с его помощью результаты легко переносятся в слабую аксиоматику. Возникает интригующий вопрос: насколько существенную часть этих математических теорий можно построить, пользуясь *только этим приёмом*, то есть в рамках слабой аксиоматики? Цель данной главы — показать примеры решения классических задач в слабой аксиоматике и наработать математическую технику, необходимую для рассмотрения более сложных задач статистического обучения в последующих главах.

1.2 Задача оценивания частоты события

Рассмотрим Задачу 1.3 о предсказании частоты события $S \subseteq \mathbb{X}$. Будем обозначать через $n(U) = |S \cap U|$ *число элементов* события S на произвольной конечной выборке $U \subseteq \mathbb{X}$.

Лемма 1.2. *Если $n(\mathbb{X}) = m$, то число элементов события S в наблюдаемой подвыборке $n(X)$ и в скрытой подвыборке $n(\bar{X})$ подчиняются гипергеометрическому распределению:*

$$\mathsf{P}\big[n(X) = s\big] = \mathsf{P}\big[n(\bar{X}) = m - s\big] = h_L^{\ell,m}(s) = \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}, \tag{1.21}$$

где s принимает значения от $s_0 = \max\{0, m-k\}$ до $s_1 = \min\{\ell, m\}$.

Доказательство. Отобрать s элементов события S в наблюдаемую подвыборку можно C_m^s различными способами. Для каждого из этих способов имеется $C_{L-m}^{\ell-s}$ способов сформировать оставшуюся часть наблюдаемой подвыборки из объектов, не принадлежащих S. Значит, $C_m^s C_{L-m}^{\ell-s}$ — число разбиений, при которых s элементов множества S попадают в наблюдаемую подвыборку, остальные $(m-s)$ — в скрытую. Их доля в общем числе разбиений C_L^ℓ как раз и составляет $h_L^{\ell,m}(s)$. ■

Замечание 1.5. Если не выполняется одно из условий $0 \leqslant s \leqslant m,\ 0 \leqslant \ell - s \leqslant L - m$, или, что то же самое, не выполняется условие $s_0 \leqslant s \leqslant s_1$, то соответствующее число разбиений равно нулю. В этом случае будем полагать $h_L^{\ell,m}(s) = 0$.

1.2.1 Свойства гипергеометрического распределения

Гипергеометрическое распределение носит фундаментальный характер и возникает, как мы увидим далее, в большинстве задач эмпирического предсказания. В данном параграфе перечисляются в справочном порядке основные свойства гипергеометрического распределения [4, 5].

1. При фиксированных L и ℓ функция $h_L^{\ell,m}(s) = \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ определена на множестве пар целых чисел (m, s): $0 \leqslant m \leqslant L,\ \max\{0, m-k\} = s_0 \leqslant s \leqslant s_1 = \min\{\ell, m\}$. Это множество имеет форму параллелограмма, рис. 1.3. Вне этой области принято полагать $h_L^{\ell,m}(s) = 0$.

2. Введём следующие обозначения для сумм крайних левых и крайних правых членов гипергеометрического распределения:

$$H_L^{\ell,m}(z) = \sum_{s=s_0}^{\lfloor z \rfloor} h_L^{\ell,m}(s); \qquad \bar{H}_L^{\ell,m}(z) = \sum_{s=\lceil z \rceil}^{s_1} h_L^{\ell,m}(s). \tag{1.22}$$

Справедлива формула полной вероятности:

$$\sum_{s=s_0}^{s_1} h_L^{\ell,m}(s) = H_L^{\ell,m}(s_1) = \bar{H}_L^{\ell,m}(s_0) = 1.$$

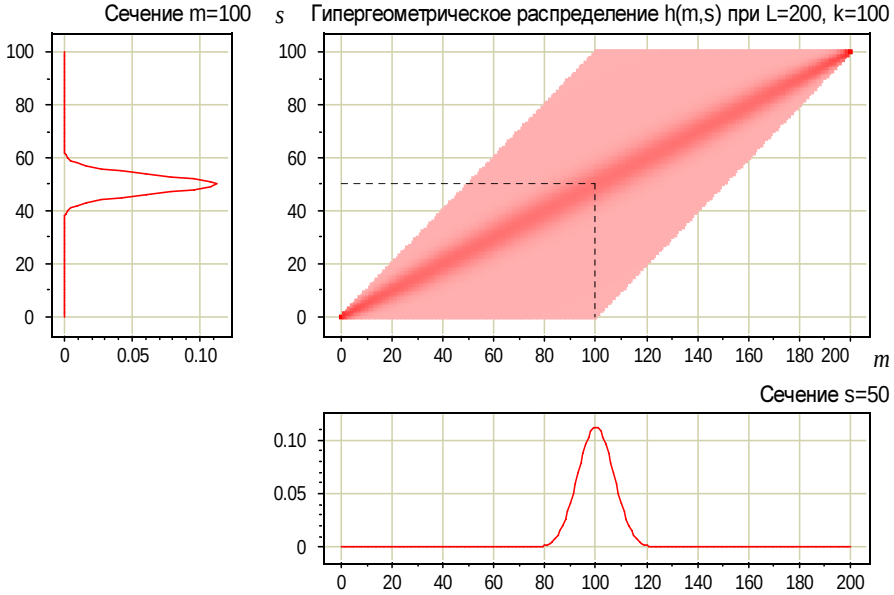


Рис. 1.3. Область определения гипергеометрической функции $h_L^{\ell,m}(s)$ при $L = 200$, $\ell = k = 100$, $m = 30$.

При фиксированных L, ℓ и m функция $h(s) = h_L^{\ell,m}(s)$ является одномерным дискретным распределением. Для примера на рис. 1.3 слева показана функция $h(s)$ при фиксированном $m = 100$. Функция $h'(m) = h_L^{\ell,m}(s)$ распределением, вообще говоря, не является, так как не удовлетворяет условию нормировки $\sum_m h'(m) \neq 1$. На рис. 1.3 снизу показана функция $h'(m)$ при фиксированном $s = 50$.

3. Параметры ℓ и m можно переставлять местами: $h_L^{\ell,m}(s) = h_L^{m,\ell}(s)$.
4. Параметры m и s можно заменять разностями: $h_L^{\ell,m}(s) = h_L^{\ell,L-m}(\ell - s)$.
5. Справедливы тождества:

$$h_L^{\ell,m}(s) = h_L^{\ell,L-m}(\ell - s) = h_L^{m,\ell}(s) = h_L^{m,k}(m - s) = h_L^{k,m}(m - s).$$

6. Отсюда вытекают тождества для функций H и $\bar{H}$:

$$H_L^{\ell,m}(s) = \sum_{j=s_0}^{s} h_L^{\ell,m}(j) = \sum_{j=s_0}^{s} h_L^{k,m}(m - j) = \bar{H}_L^{k,m}(m - s).$$

7. Распределение $h(s)$ является унимодальным (имеет форму пика). Максимальное значение достигается при $s^* = \frac{(m+1)(\ell+1)}{L+2}$, с точностью до округления.

8. Таблица гипергеометрического распределения содержит ℓk ненулевых значений. Её можно эффективно вычислить с помощью рекуррентных соотношений:

$$\begin{aligned}
h_L^{\ell,0}(0) &= 1; \\
h_L^{\ell,m+1}(s) &= h_L^{\ell,m}(s)\,\frac{m+1}{m+1-s}\cdot\frac{k-m+s}{L-m}; \\
h_L^{\ell,m}(s+1) &= h_L^{\ell,m}(s)\,\frac{m-s}{s+1}\cdot\frac{\ell-s}{k-m+s+1}; \\
h_L^{\ell,m}(s-1) &= h_L^{\ell,m}(s)\,\frac{s}{m-s+1}\cdot\frac{k-m+s}{\ell-s+1}.
\end{aligned} \tag{1.23}$$

Чтобы избежать накопления вычислительных погрешностей, значения $h_L^{\ell,m}(s)$ вычисляются последовательно для всех $m = 0,\ldots,L$. Для каждого m первым вычисляется максимальное значение или близкое к максимальному (достаточно взять $s = s_{\max}$), затем меньшие значения вычисляются через бóльшие.

9. Матожидание величины s есть

$$\lambda = \sum_{s=s_0}^{s_1} s h_L^{\ell,m}(s) = \frac{\ell m}{L}.$$

10. Дисперсия величины s есть

$$\sigma^2 = \sum_{s=s_0}^{s_1} (s-\lambda)^2 h_L^{\ell,m}(s) = \lambda \frac{k(L-m)}{L(L-1)}.$$

11. При больши́х значениях параметров L, ℓ, m предельными распределениями для $h(s) = h_L^{\ell,m}(s)$ могут быть только распределения одного из четырёх типов:

— при $\lambda \to \infty$ *нормальное распределение* $h(s) \to \frac{1}{\sqrt{2\pi}} \exp\bigl(-\frac{(s-\lambda)^2}{2\sigma^2}\bigr)$;
— если λ имеет конечный предел:
— при $\frac{m}{L} \to p$ *биномиальное распределение* $h(s) \to C_\ell^s\, p^s (1-p)^{\ell-s}$;
— при $\frac{\ell}{L} \to p$ биномиальное распределение $h(s) \to C_m^s\, p^s (1-p)^{m-s}$;
— при $\frac{m}{L} \to 0$ или $\frac{\ell}{L} \to 0$ *распределение Пуассона* $h(s) \to e^{-\lambda}\lambda^s / s!$;
— при $\lambda \to 0$ вырожденное распределение $s = 0$.

12. Гипергеометрическое распределение довольно точно приближается с помощью аппроксимации Моленара:

$$h(s) \approx C_\ell^s\, \tilde{p}^s (1-\tilde{p})^{\ell-s}, \quad \tilde{p} = \frac{2m-s}{2L-\ell+1}.$$

1.2.2 Закон больших чисел в слабой аксиоматике

Продолжим рассмотрение Задачи 1.3 о предсказании частоты события $S \subseteq \mathbb{X}$. Введём сокращённые обозначения для частот события S на выборках X и $\bar{X}$:

$$\nu = \frac{n(X)}{\ell}, \quad \bar{\nu} = \frac{n(\bar{X})}{k}.$$

Теорема 1.3. *Пусть $n(\mathbb{X}) = m$. Для любого $\varepsilon \in [0,1)$ справедливы точные оценки:*

$$\mathsf{P}\bigl[\nu \leqslant \varepsilon\bigr] = H_L^{\ell,m}\bigl(\lfloor \varepsilon\ell \rfloor\bigr); \tag{1.24}$$

$$\mathsf{P}\bigl[\bar{\nu} \geqslant \varepsilon\bigr] = H_L^{\ell,m}\bigl(\lfloor m - \varepsilon k \rfloor\bigr); \tag{1.25}$$

$$\mathsf{P}\bigl[\bar{\nu} - \nu \geqslant \varepsilon\bigr] = H_L^{\ell,m}\bigl(s_m^-(\varepsilon)\bigr), \qquad s_m^-(\varepsilon) = \bigl\lfloor \tfrac{\ell}{L}(m - \varepsilon k) \bigr\rfloor; \tag{1.26}$$

$$\mathsf{P}\bigl[|\bar{\nu} - \nu| \geqslant \varepsilon\bigr] = H_L^{\ell,m}\bigl(s_m^-(\varepsilon)\bigr) + \bar{H}_L^{\ell,m}\bigl(s_m^+(\varepsilon)\bigr), \qquad s_m^+(\varepsilon) = \bigl\lceil \tfrac{\ell}{L}(m + \varepsilon k) \bigr\rceil. \tag{1.27}$$

Доказательство. Первые два неравенства являются непосредственным следствием (1.21). Третье следует из первого, если заметить, что $\bar{\nu} - \nu \geqslant \varepsilon$ равносильно $\frac{m-s}{k} - \frac{s}{\ell} \geqslant \varepsilon$, откуда элементарными преобразованиями получаем $s \leqslant \frac{\ell}{L}(m - \varepsilon k)$.

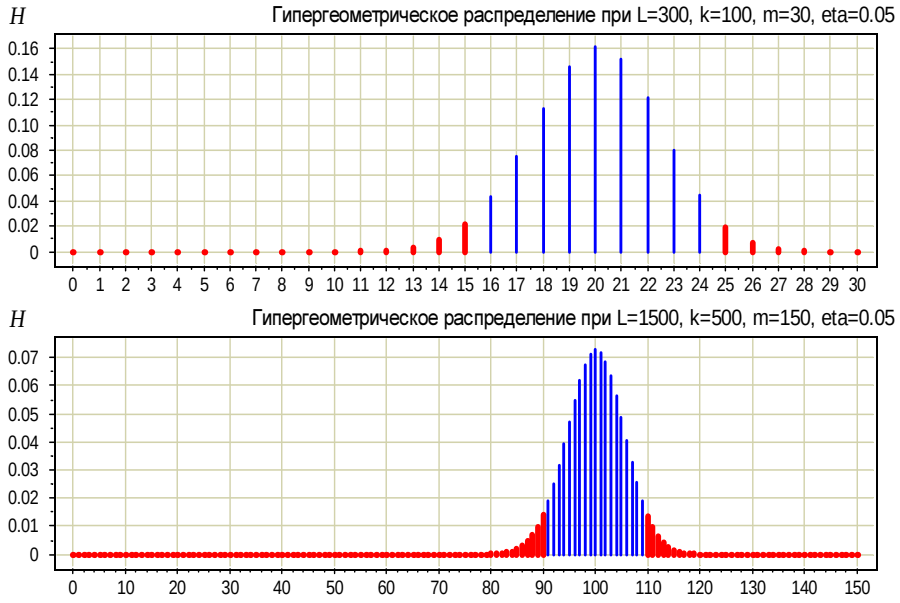


Рис. 1.4. Зависимость относительной ширины гипергеометрического пика $h_L^{\ell,m}(s)$ от длины выборки L. Верхний график получен при $L = 300$, $\ell = 200$, $m = 30$. Выделены крайние левые, $[s_0, s_m^-(\varepsilon)] = [0, 15]$, и крайние правые $[s_m^-(\varepsilon), s_1] = [25, 30]$, члены распределения, соответствующие значению надёжности $\eta = 0.05$. Нижний график получен при значениях L, ℓ, m, пропорционально увеличенных в 5 раз.

Двусторонняя оценка (1.27) доказывается аналогично, если представить множество разбиений в виде объединения двух непересекающихся подмножеств:

$$\mathsf{P}\big[|\bar\nu - \nu| \geqslant \varepsilon\big] = \mathsf{P}\big[\bar\nu - \nu \geqslant \varepsilon\big] + \mathsf{P}\big[\nu - \bar\nu \geqslant \varepsilon\big] = H_L^{\ell,m}\left(s_m^-(\varepsilon)\right) + \bar H_L^{\ell,m}\left(s_m^+(\varepsilon)\right).$$

Теорема доказана. ■

Замечание 1.6. В условии теоремы под $\lfloor z \rfloor$ понимается целая часть действительного числа z, то есть наибольшее целое, *меньшее или равное* z. Аналогично, $\lceil z \rceil$ — наименьшее целое, *большее или равное* z. Если в левой части поменять нестрогие неравенства на строгие, то все оценки останутся в силе с одной оговоркой: $\lfloor z \rfloor$ надо будет понимать как наибольшее целое, *меньшее* z; оно отличается от функции целой части только тем, что $\lfloor z \rfloor = z - 1$ при целых z. Соответственно, и $\lceil z \rceil$ надо будет понимать как наименьшее целое, *большее* z, тогда $\lceil z \rceil = z + 1$ при целых z.

О законе больших чисел. При пропорциональном увеличении L, ℓ и m относительная ширина гипергеометрического пика уменьшается, см. рис. 1.4. В пределе при $L, \ell, m \to \infty$ это позволяет сколь угодно точно предсказывать частоту события S в скрытой выборке $\bar\nu$ по его частоте на наблюдаемой выборке ν. Равенство (1.27) даёт точную оценку скорости сходимости частот события в двух выборках.

Классический закон больших чисел утверждает сходимость частоты события к её вероятности. Однако в слабой аксиоматике понятие «вероятности события S» не определено. Поэтому (1.27) можно интерпретировать как аналог *закона больших чисел* в слабой аксиоматике. Основанием для такой интерпретации служит тот факт,

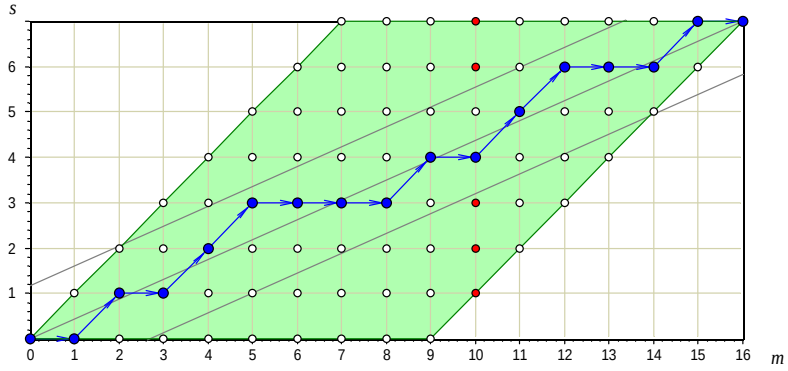


Рис. 1.5. Траектория, соответствующая бинарной последовательности $\{b_i\}_{i=0}^{L} =$ $= 0101100010110010$, при $L = 16$, $\ell = 7$, $\varepsilon = 0.3$. Проведены линии $s = \frac{\ell}{L}m$ и $s = \frac{\ell}{L}(m \pm \varepsilon k)$. При $m = 10$ выделены точки выше верхней линии, $s \geqslant s_m^+(\varepsilon)$, и ниже нижней линии, $s \leqslant s_m^-(\varepsilon)$.

что два функционала — (а) вероятность большого отклонения частот события в двух выборках и (б) вероятность большого отклонения частоты события от его вероятности — оцениваются сверху друг через друга [8]. По сути, эти две оценки отличаются не принципиально.

Классические неравенства Чебышёва, Чернова, Бернштейна, Хёффдинга [172] позволяют количественно оценивать скорость сходимости в законе больших чисел. Однако все они являются асимптотическими и дают несколько завышенные оценки вероятности большого отклонения. Выражение (1.27) является точной (не завышенной, не асимптотической) оценкой, и потому его можно считать наиболее точным выражением закона больших чисел.

Геометрическая интерпретация соотношений (1.21), (1.26) и (1.27). Рассмотрим прямоугольную сетку $\{0, \ldots, L\} \times \{0, \ldots, \ell\}$, см. рис. 1.5. Положим $b_i = [x_i \in X]$, то есть $b_i = 1$ означает, что при разбиении $\mathbb{X} = X \sqcup \bar{X}$ объект x_i попадает в наблюдаемую часть выборки. Договоримся отображать выборку в виде траектории, проходящей по узлам сетки из точки $(0, 0)$ в точку (L, ℓ) согласно правилу: если $b_i = 1$, то смещаемся на единицу вправо-вверх; если $b_i = 0$, то смещаемся на единицу вправо. Все такие траектории не выходят за пределы параллелограмма, выделенного на рис. 1.5. Множество всех таких траекторий изоморфно множеству разбиений выборки $\mathbb{X} = X \sqcup \bar{X}$, и оба они изоморфны множеству L-мерных бинарных векторов $(b_1, \ldots, b_L)$, содержащих ровно ℓ единиц. Поэтому для подсчёта числа разбиений, удовлетворяющих некоторому свойству, достаточно найти число соответствующих траекторий.

Чтобы вывести (1.21), пронумеруем объекты выборки так, чтобы первые m объектов принадлежали множеству S. Тогда задача сведётся к подсчёту доли траекторий, проходящих через точку (m, s). Назовём такие траектории *допустимыми*. Число всех возможных траекторий на отрезке от $(0, 0)$ до (m, s) равно C_m^s, и для каждой траектории существует $C_{L-m}^{\ell-s}$ вариантов её продолжения от (m, s) до (L, ℓ).

Следовательно, число допустимых траекторий равно $C_m^s C_{L-m}^{\ell-s}$. Разделив на общее число возможных траекторий C_L^ℓ, получаем требуемое $h_L^{\ell,m}(s)$.

Чтобы вывести (1.26), необходимо подсчитать число траекторий, проходящих через любую точку (m, s), лежащую ниже диагонали на $\varepsilon\frac{\ell k}{L}$ или более. Для этого суммируется число траекторий $C_m^s C_{L-m}^{\ell-s}$ по всем $s = s_0, \ldots, s_m^-(\varepsilon)$.

Для вывода двусторонней оценки (1.27) подсчитывается число траекторий, отстоящих от диагонали на $\varepsilon\frac{\ell k}{L}$ или более. В этом случае суммирование идёт по всем $s = s_0, \ldots, s_m^-(\varepsilon)$, затем по всем $s = s_m^+(\varepsilon), \ldots, s_1$.

Задача выборочного контроля качества является классическим примером прикладной задачи, в которой оценки Теоремы 1.3 применяются непосредственно [4]. Пусть $\mathbb{X}$ — множество изделий, $S \subset \mathbb{X}$ — подмножество дефектных изделий. Изготовлена партия изделий $\mathbb{X}$, из них m оказались дефектными. Число m неизвестно. Проверить всю партию поштучно не представляется возможным. Поэтому делается *выборочный контроль качества*: случайно, независимо, без возвращений выбирается подмножество $X \subset \mathbb{X}$, что равносильно случайному равномерному выбору разбиения $X \sqcup \bar{X} = \mathbb{X}$. Зная долю дефектов в наблюдаемой подвыборке ν, требуется предсказать долю дефектов в скрытой подвыборке $\bar{\nu}$. Если при заданной точности ε и надёжности η имеет место оценка $\mathsf{P}[\bar{\nu} \geqslant \varepsilon] < \eta$, то партия $\mathbb{X}$ принимается, иначе она целиком бракуется. Параметры ε и η подбираются из экономических соображений — с учётом стоимости контроля одного изделия и величины потерь от использования дефектного изделия.

1.2.3 Проблема неизвестного m и наблюдаемые оценки

Оценочные функции (1.24)–(1.27) зависят от числа элементов m события S в генеральной выборке $\mathbb{X}$, которое невозможно узнать, пока неизвестна скрытая часть данных. Таким образом, оценки (1.24)–(1.27) являются ненаблюдаемыми.

Прежде чем применить описанный в 1.1.3 (стр. 23) переход от ненаблюдаемой оценки к наблюдаемой, покажем, что известные альтернативные подходы либо не дают достаточно точных оценок, либо требуют привлечения субъективной дополнительной информации.

Верхняя оценка. Простейшее решение проблемы неизвестного m заключается в том, чтобы взять максимум по m и получить вместо точной оценки завышенную верхнюю оценку:

$$\mathsf{P}\big[\bar{\nu} - \nu \geqslant \varepsilon\big] \leqslant \max_{m=0,\ldots,L} H_L^{\ell,m}\big(s_m^-(\varepsilon)\big) \equiv \Gamma_L^\ell(\varepsilon). \tag{1.28}$$

Здесь максимум достаточно взять по всем m от $\lceil \varepsilon k \rceil$ до $\lfloor L - \varepsilon\ell \rfloor$, так как при остальных значениях m левая часть неравенства равна нулю.

К сожалению, (1.28) — довольно грубая оценка при малых m, см. рис. 1.6. По этой причине данный подход не приемлем для задач выборочного контроля качества, обучения по прецедентам, и других случаев, когда именно малые значения m представляют большой практический интерес.

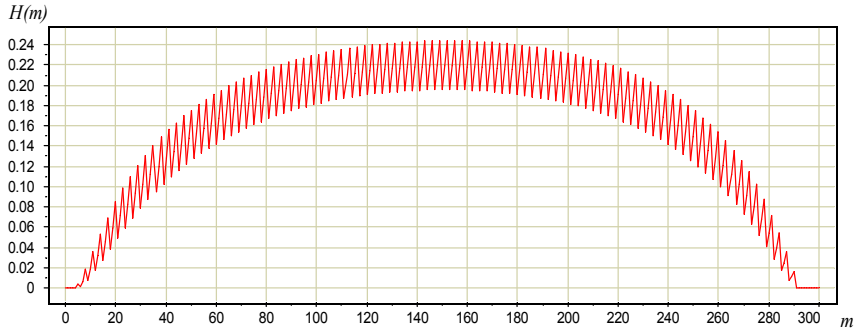


Рис. 1.6. График зависимости $H(m) = H_L^{\ell,m}\left(s_m^-(\varepsilon)\right)$ от m при $L = 300,\ \ell = 200,$ $\varepsilon = 0.05$.

Известна верхняя оценка «хвоста» гипергеометрического распределения [120], с помощью которой можно оценить сверху правую часть (1.28): для любого $\varepsilon > 0$

$$\Gamma_L^\ell(\varepsilon) \leqslant \exp\left(-2\varepsilon^2\frac{\ell k^2}{L^2}\right).$$

Эта оценка ещё более грубая (на рис. 1.6 ей соответствовала бы горизонтальная линия с ординатой 0.89, но она не показана). Асимптотически эта оценка сходится к нулю при одновременном стремлении ℓ и k к бесконечности, что ещё раз подтверждает связь точных оценок (1.26) и (1.27) с *законом больших чисел*.

О байесовском оценивании. В задаче выборочного контроля качества m — это неизвестное число бракованных изделий в общей партии из L штук изделий. Возникают вопросы: является ли m случайной величиной или неслучайным параметром; и если это случайная величина, то каково её распределение? Окончательного ответа на них в [4] не даётся. Вместо этого предлагается несколько альтернативных подходов, приводящих, вообще говоря, к разным результатам.

Пусть m — случайная величина с заданным априорным распределением $p(m)$. Тогда, зная число s элементов события S в наблюдаемой выборке и зная распределение $p(s|m) = H_L^{\ell,m}(s)$, можно оценить апостериорное распределение $p(m|s)$ по формуле Байеса. В [4] рассмотрено несколько вариантов задания априорного распределения: равномерное, биномиальное, гипергеометрическое, и даже смеси биномиальных или гипергеометрических распределений.

В слабой аксиоматике, чтобы применить байесовский подход, необходимо определить вероятность $p(m)$ через долю разбиений выборки. Для этого придётся ввести расширенную генеральную выборку $X^{\mathbb{L}}$, в которой M элементов принадлежат событию S, и предположить, что выборка $\mathbb{X}$ равновероятна среди всех $C_{\mathbb{L}}^L$ разбиений расширенной выборки $X^{\mathbb{L}}$. Тогда величина m будет подчиняться гипергеометрическому распределению $p(m) = h_{\mathbb{L}}^{L,M}(m)$ с неизвестным параметром M. Однако вопрос «чему равно M для выборки $X^{\mathbb{L}}$», ничем не отличается от исходного вопроса «чему равно m для выборки $\mathbb{X}$». Таким образом, априорная вероятность $p(m)$ не имеет частотной интерпретации, и её следует понимать как *субъективную вероятность*.

Недостаток байесовского оценивания в том, что результат зависит от субъективной априорной информации $p(m)$, что нежелательно в большинстве приложений.

Неизвестную величину m желательно рассматривать как неслучайный параметр, оцениваемый по случайной наблюдаемой выборке X.

Переход от ненаблюдаемой оценки к наблюдаемой не требует привлечения дополнительной информации и приводит к точным верхним и нижним оценкам для $n(\mathbb{X})$ и $n(\bar{X})$, вычисляемым по наблюдаемому значению числа ошибок $s = n(X)$.

Теорема 1.4. *Если $s = n(X)$ — число элементов события S в наблюдаемой выборке, то для числа элементов события S в полной выборке с вероятностью $(1-\eta)$ справедлива верхняя оценка:*

$$n(\mathbb{X}) \leqslant \max\bigl\{m = m_0, \ldots, L \;\bigm|\; H_L^{\ell,m}(s) \geqslant \eta\bigr\}, \quad \text{где } \; m_0 = \bigl\lceil s\tfrac{L+2}{\ell+1} - 1\bigr\rceil. \tag{1.29}$$

Доказательство. Рассмотрим одностороннюю точную оценку (1.25), обозначив правую её часть через $H(\varepsilon, m)$, где $m = n(\mathbb{X})$ — неизвестная величина:

$$\mathsf{P}\bigl[\bar{\nu} \geqslant \varepsilon\bigr] = H_L^{\ell,m}(\lfloor m - \varepsilon k\rfloor) = H(\varepsilon, m). \tag{1.30}$$

Тогда с вероятностью $(1-\eta)$ справедлива оценка сверху $\bar{\nu} < E(\eta, m)$, где $E(\eta, m)$ — обратная функция от $H(\varepsilon, m)$. Обращение производится по первому аргументу при каждом значении второго аргумента m, который выступает в роли параметра. Поскольку функция $E(\eta, m)$ не возрастает по первому аргументу, из оценки $\bar{\nu} < E(\eta, m)$ следует неравенство $H(\bar{\nu}, m) \geqslant \eta$. Подставляя $\bar{\nu} = \frac{m-s}{k}$ в функцию $H(\bar{\nu}, m)$, определяемую согласно (1.30), получаем, что с вероятностью $(1-\eta)$ справедливо неравенство $H_L^{\ell,m}(s) \geqslant \eta$. Чтобы разрешить данное неравенство относительно m при фиксированном s, найдём максимальное значение m, при котором оно выполнено. При максимальном значении m значение s должно находится левее точки максимума гипергеометрического распределения $s^* = \frac{(m+1)(\ell+1)}{L+2}$, Следовательно, $s(L+2) < (m+1)(\ell+1)$. Поэтому для нахождения максимального значения m достаточно перебрать значения m, не меньшие $s\frac{L+2}{\ell+1} - 1$. Теорема доказана. ■

Замечание 1.7. Аналогично оценивается частота на скрытой выборке $\bar{\nu}$ по частоте на наблюдаемой выборке ν: с вероятностью $(1-\eta)$ выполнено неравенство

$$\bar{\nu} \leqslant \tfrac{1}{k}\max\bigl\{t = t_0, \ldots, k \;\bigm|\; H_L^{\ell,\nu\ell+t}(\nu\ell) \geqslant \eta\bigr\}, \quad t_0 = \bigl\lceil s\tfrac{k+1}{\ell+1} - 1\bigr\rceil. \tag{1.31}$$

Замечание 1.8. Аналогично строятся и нижние оценки: с вероятностью $(1-\eta)$

$$n(\mathbb{X}) \geqslant \min\bigl\{m = 0, \ldots, m_0 \;\bigm|\; \bar{H}_L^{\ell,m}(s) \geqslant \eta\bigr\};$$
$$\bar{\nu} \geqslant \tfrac{1}{k}\min\bigl\{t = 0, \ldots, t_0 \;\bigm|\; \bar{H}_L^{\ell,\nu\ell+t}(\nu\ell) \geqslant \eta\bigr\},$$

Вычисление полученных верхних и нижних оценок с использованием рекуррентных соотношений (1.23) требует порядка $O\bigl(n(X)n(\bar{X})\bigr)$ операций.

На рис. 1.7 показаны верхние и нижние оценки числа элементов события S в скрытой выборке $t = n(\bar{X})$ в зависимости от их числа в наблюдаемой выборке $s = n(X)$. Толстые ступенчатые линии — граничные области, в которых выполняется равенство $H_L^{\ell,s+t}(s) = \eta$. Точная верхняя оценка совпадает с верхней границей

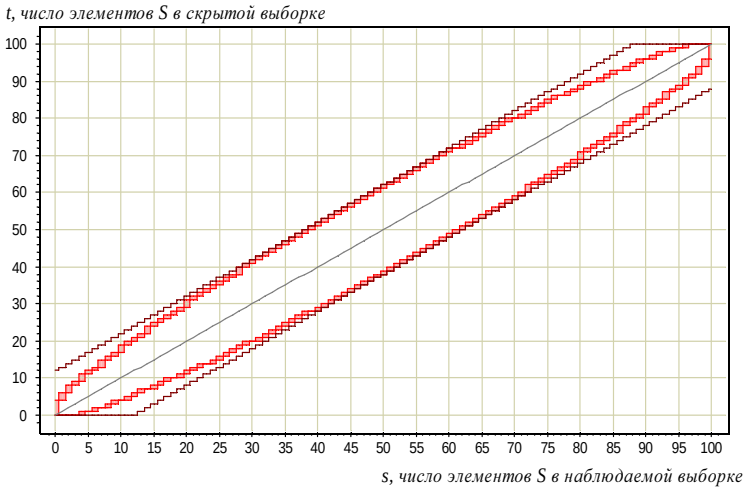


Рис. 1.7. Точные верхние и нижние оценки числа $t = n(\bar{X})$ элементов события S в скрытой выборке в зависимости от их числа $s = n(X)$ в наблюдаемой выборке. Условия эксперимента: $L = 200,\ \ell = k = 100,\ \eta = 0.05$.

верхней области, точная нижняя — с нижней границей нижней области. Вместе они определяют $1 - 2\eta = 90\%$-й доверительный интервал для числа t при каждом значении s. Тонкие ступенчатые линии — это оценки по наихудшему m, вычисленные согласно (1.28). Их точность падает по мере приближения m к 0 или к L.

О вероятности нуль-события. В работе С. И. Гурова [38] ставится задача точечного оценивания вероятности p события, ни разу не наблюдавшегося в заданной выборке длины ℓ. Рассматриваются следующие типы оценок (предполагается, что коэффициент доверия $\eta \in (0, 1)$ задаётся априори и достаточно близок к единице):

1. Оценка максимального правдоподобия вырождена и равна нулю.

2. Верхняя граница доверительного интервала согласно классическому частотному подходу равна $\hat{p} = 1 - \sqrt[\ell]{1 - \eta}$ и представляется сильно завышенной.

3. Байесовские оценки по математическому ожиданию $\left(\hat{p}_B = \frac{1}{\ell+2}\right)$ или по медиане $\left(\hat{p} = 1 - \sqrt[\ell]{0.5}\,\right)$ апостериорного распределения представляются завышенными при больших ℓ.

4. Наконец, предлагается оценка $\hat{p}_0(\ell) = 1 - \sqrt[\ell]{\eta}$, получаемая несколькими способами, в том числе путём замены наблюдавшейся выборки некоторой гипотетической выборкой, в которой событие произошло хотя бы один раз. Замена делается таким образом, чтобы для данной пары выборок принималась гипотеза однородности. Затем к новой выборке применяется оценка максимального правдоподобия. В качестве критерия однородности используется *точный тест Фишера*, основанный на гипергеометрическом распределении.

Основные выводы [38] следующие: при малых выборках (ℓ от 4 до 20–30 объектов) рекомендуется использовать байесовскую оценку $\hat{p}_B$; при больших выборках (ℓ более 20–30 объектов) — оценку $\hat{p}_0(\ell)$. Констатируется наличие резкого скачка оценки при переходе от «малой» выборки к «большой».

В описанном подходе вызывает неудовлетворённость как большое количество оценок, дающих противоречивые результаты, так и их эвристический характер. Нет оснований полагать, что проведённый анализ полон, то есть все разумные способы оценивания исчерпаны. Но, главное, вызывает сомнение правомерность самого понятия «вероятности редкого события», а также возможность и полезность его точечного оценивания в условиях малых выборок.

В рамках слабой аксиоматики задача точечного оценивания вероятности нуль-события неправомерна. Вместо неё можно ставить задачу интервального оценивания частоты нуль-события на произвольной скрытой выборке заданной длины k. Это частный случай задачи эмпирического предсказания 1.3 (стр. 18) при условии, что $\nu(X) = 0$. Данная задача имеет точное решение (1.31), в которое достаточно подставить $\nu = 0$. В частности, по графику на рис. 1.7 видно, что при $s = 0$ и длине наблюдаемой выборки $\ell = 100$ число событий в скрытой выборке длины $k = 100$ не превзойдёт 4 с вероятностью $1 - \eta = 95\%$.

Заметим, что для получения этой оценки, так же, как и в [38], использовалось гипергеометрическое распределение. Однако в слабой аксиоматике решение является точным, получается более коротким и при том единственным способом, не требующим ни асимптотических приближений, ни изобретения эвристических приёмов.

1.3 Задача оценивания функции распределения

Рассмотрим Задачу 1.5 об оценивании функции распределения. Для произвольной функции $\xi\colon \mathbb{X} \to \mathbb{R}$ и произвольного разбиения $X \sqcup \bar{X} = \mathbb{X}$ определим одностороннее и двустороннее *равномерное отклонение эмпирических функций распределения*:

$$D^+(X,\bar{X}) = \max_{z\in\mathbb{R}}\big(F_\xi(z,\bar{X}) - F_\xi(z,X)\big);$$
$$D^-(X,\bar{X}) = \max_{z\in\mathbb{R}}\big(F_\xi(z,X) - F_\xi(z,\bar{X})\big);$$
$$D(X,\bar{X}) = \max_{z\in\mathbb{R}}\big|F_\xi(z,\bar{X}) - F_\xi(z,X)\big|.$$

В сильной аксиоматике имеет место следующая теорема [80].

Теорема 1.5 (Н. В. Смирнов). *Если $X, \bar{X} \subseteq \mathscr{X}$ — случайные, независимые, одинаково распределённые выборки; $\xi\colon \mathscr{X} \to \mathbb{R}$ — случайная величина с непрерывным распределением, то справедливы асимптотические оценки*

$$\lim_{\ell,k\to\infty} \mathsf{P}\big\{D^\pm(X,\bar{X}) \geqslant \varepsilon\big\} = \exp\left(-2\varepsilon^2\tfrac{\ell k}{\ell+k}\right); \tag{1.32}$$

$$\lim_{\ell,k\to\infty} \mathsf{P}\big\{D(X,\bar{X}) \geqslant \varepsilon\big\} = 2\sum_{i=1}^{\infty}(-1)^{i+1}\exp\left(-2\varepsilon^2\tfrac{\ell k}{\ell+k}i^2\right); \tag{1.33}$$

Заметим, что правая часть (1.33) представима также в виде $1 - K\left(\varepsilon\sqrt{\frac{\ell k}{\ell+k}}\right)$, где $K(z) = \sum\limits_{i=-\infty}^{\infty}(-1)^i e^{-2z^2i^2}$ — *функция распределения Колмогорова*.

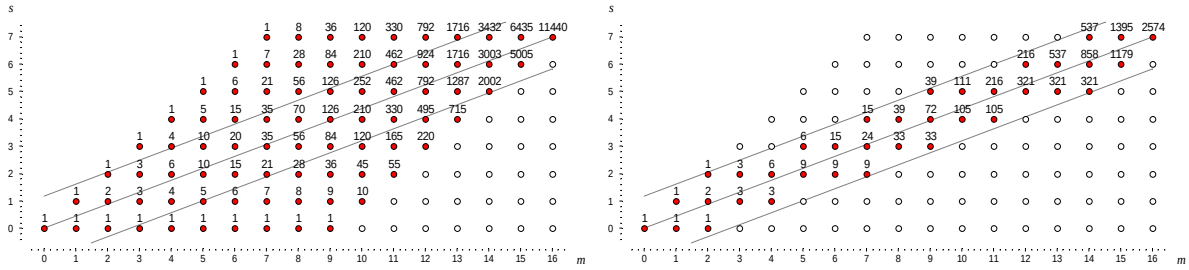


Рис. 1.8. Треугольники Паскаля: классический $C_m^s = G_m^s[0, m]$ и усечённый $G_m^s\big[g_m^-(\varepsilon), g_m^+(\varepsilon)\big]$, при $L = 16$, $\ell = 7$, $\varepsilon = 0.3$.

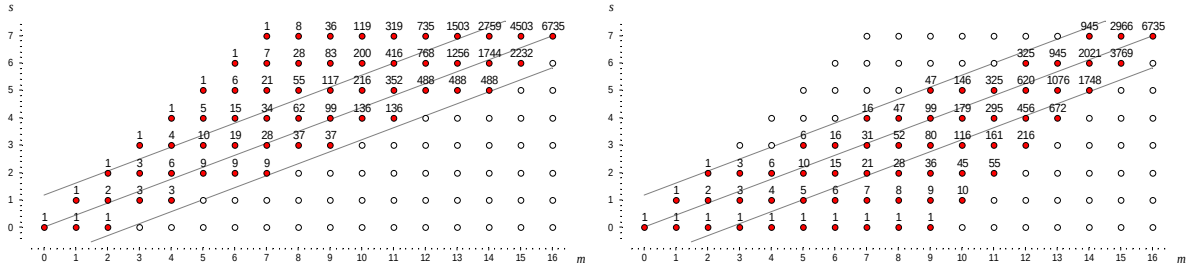


Рис. 1.9. Треугольники Паскаля: усечённый слева $G_m^s\big[g_m^-(\varepsilon), m\big]$ и усечённый справа $G_m^s\big[0, g_m^+(\varepsilon)\big]$, при $L = 16$, $\ell = 7$, $\varepsilon = 0.3$.

Известны и неасимптотические точные оценки, но они имеют достаточно громоздкий вид [33]. Мы покажем, что точные оценки могут быть выражены более элегантно через усечённый треугольник Паскаля [104]. Доказательство проводится в рамках слабой аксиоматики и имеет прозрачный геометрический смысл.

1.3.1 Усечённый треугольник Паскаля

Пусть $g_m^-,\ g_m^+,\ m = 0, \ldots, L$ — две неубывающие последовательности, удовлетворяющие условию $0 \leqslant g_m^- \leqslant g_m^+ \leqslant m$.

Определение 1.5. *Усечённым треугольником Паскаля с нижней границей g_m^- и верхней границей g_m^+ будем называть целочисленную функцию $G_m^s = G_m^s[g_m^-, g_m^+]$, определённую рекуррентными соотношениями $G_0^s = [s = 0]$ и*

$$G_m^s = \big(G_{m-1}^s + G_{m-1}^{s-1}\big)\big[g_m^- \leqslant s \leqslant g_m^+\big], \quad m \in \mathbb{N}, \quad s \in \mathbb{Z}. \tag{1.34}$$

Усечённый треугольник Паскаля вычисляется по тому же рекуррентному правилу, что и классический треугольник Паскаля, если в нём обнулить все элементы, лежащие за пределами границ $[g_m^-, g_m^+]$. «Неусечённый» треугольник Паскаля $G_m^s[0, m]$ совпадает с классическим и даёт биномиальные коэффициенты C_m^s.

При начальном условии $G_0^s = [s = 0]$ ненулевыми могут быть только элементы G_m^s при $0 \leqslant s \leqslant m$. Другие начальные условия приводят к неклассическим обобщениям треугольника Паскаля, которые в данной работе не рассматриваются.

Определим для произвольных $\varepsilon > 0$, $m = 0, 1, 2, \ldots$, следующие функции (в дальнейшем аргумент ε будем опускать):

$g_m^+(\varepsilon) = \frac{\ell}{L}(m + \varepsilon k)$;
$g_m^-(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$.

На рис. 1.8 и 1.9 приведены четыре возможных варианта усечённых треугольников Паскаля с такими границами. В отличие от принятого способа изображения здесь они «положены на бок» путём поворота на 90° против часовой стрелки.

1.3.2 Теорема Смирнова в слабой аксиоматике

Теорема 1.6. *В слабой аксиоматике для произвольной конечной выборки $\mathbb{X}$ и произвольной функции $\xi\colon \mathbb{X} \to \mathbb{R}$, значения которой попарно различны на элементах выборки $\mathbb{X}$, справедливы точные оценки:*

$$\mathsf{P}\big[D^+(X, \bar{X}) \leqslant \varepsilon\big] = G_L^\ell\big[0, g_L^+(\varepsilon)\big]/C_L^\ell; \tag{1.35}$$
$$\mathsf{P}\big[D^-(X, \bar{X}) \leqslant \varepsilon\big] = G_L^\ell\big[g_L^-(\varepsilon), L\big]/C_L^\ell; \tag{1.36}$$
$$\mathsf{P}\big[D(X, \bar{X}) \leqslant \varepsilon\big] = G_L^\ell\big[g_L^-(\varepsilon), g_L^+(\varepsilon)\big]/C_L^\ell. \tag{1.37}$$

Доказательство. 1. Составим вариационный ряд значений функции $\xi(x)$ на элементах выборки: $\xi(x^{(1)}) < \xi(x^{(2)}) < \cdots < \xi(x^{(L)})$. Здесь все неравенства строгие в силу условия попарной различности.

Обозначим $b_i = b_i(X) = [x^{(i)} \in X]$. Бинарная последовательность $b_1, \ldots, b_L$ содержит ровно ℓ единиц и k нулей.

Воспользуемся определением функции распределения:

$$D(X, \bar{X}) = \max_{z\in\mathbb{R}} \big|F_\xi(z, \bar{X}) - F_\xi(z, X)\big| = \\ = \max_{z\in\mathbb{R}} \left|\frac{1}{k}\sum_{i=1}^{L}[x_i \in \bar{X}][\xi(x_i) < z] - \frac{1}{\ell}\sum_{i=1}^{L}[x_i \in X][\xi(x_i) < z]\right|. \tag{1.38}$$

Изменим порядок слагаемых в суммах, теперь суммируя их в порядке возрастания значений $\xi(x_i)$. Это равносильно тому, что в данной формуле все вхождения x_i заменятся на $x^{(i)}$. Тогда можно убрать сомножитель $\big[\xi(x^{(i)}) < z\big]$, заменив верхний предел суммирования на $m = \max\big\{i\colon \xi(x^{(i)}) < z\big\}$, и максимум брать не по действительному параметру z, а по целочисленному параметру m:

$$D(X, \bar{X}) = \max_{m=1..L} \left|\frac{1}{k}\sum_{i=1}^{m}\underbrace{[x^{(i)} \in \bar{X}]}_{1-b_i} - \frac{1}{\ell}\sum_{i=1}^{m}\underbrace{[x^{(i)} \in X]}_{b_i}\right| = \\ = \max_{m=1..L}\left|\frac{m}{k} - \frac{\ell+k}{\ell k}\sum_{i=1}^{m} b_i\right| = \frac{L}{\ell k}\max_{m=1..L}\left|B_m - \frac{m\ell}{L}\right|,$$

где $B_m = B_m(X, \bar{X}) = b_1 + \cdots + b_m$.

Таким образом, равномерное отклонение эмпирических распределений на выборках X и $\bar{X}$ выражается через равномерное отклонение числа единиц в первых m членах последовательности B_m от «ожидаемого» числа единиц $m\ell/L$:

Теперь запишем долю разбиений выборки $\mathbb{X}$, при которых равномерное отклонение эмпирических распределений не превышает пороговую точность ε:

$$\begin{aligned}
&\mathsf{P}\big[D(X,\bar{X}) \leqslant \varepsilon\big] = \\
&= \mathsf{P}\left[\max_{m=1..L} \left|B_m - \frac{m\ell}{L}\right| \leqslant \frac{\varepsilon\ell k}{L}\right] = \\
&= \mathsf{P}\left[\max_{m=1..L} \Big(-B_m + \underbrace{\big(\tfrac{m\ell}{L} - \tfrac{\varepsilon\ell k}{L}\big)}_{g_m^-(\varepsilon)}\Big) \leqslant 0\right]\left[\max_{m=1..L} \Big(B_m - \underbrace{\big(\tfrac{m\ell}{L} + \tfrac{\varepsilon\ell k}{L}\big)}_{g_m^+(\varepsilon)}\Big) \leqslant 0\right] = \\
&= \frac{1}{C_L^\ell} \sum_{(X,\bar{X})} \prod_{m=1}^{L} \big[g_m^-(\varepsilon) \leqslant B_m(X,\bar{X}) \leqslant g_m^+(\varepsilon)\big]. \qquad (1.39)
\end{aligned}$$

Последнее равенство следует из тождества $\big[\max\limits_m A_m \leqslant 0\big] = \prod\limits_m \big[A_m \leqslant 0\big]$.

2. Рассмотрим подвыборку $X^m = \{x^{(1)}, \ldots, x^{(m)}\}$, состоящую из первых m членов вариационного ряда. Возьмём максимальное (по включению) подмножество N разбиений $(X,\bar{X})$, удовлетворяющих двум условиям:

1) они индуцируют попарно различные разбиения подвыборки X^m;

2) ровно s объектов из X^m попадают в X.

Очевидно, число этих разбиений $|N| = C_m^s$. Представим множество разбиений N в виде объединения непересекающихся подмножеств $N_0 = \{(X,\bar{X}) \in N\colon b_m(X) = 0\}$ и $N_1 = \{(X,\bar{X}) \in N\colon b_m(X) = 1\}$. Очевидно, $|N_0| = C_{m-1}^s$, $|N_1| = C_{m-1}^{s-1}$.

Нас будет интересовать выражение $H_m^s = \sum\limits_{(X,\bar{X})} \prod\limits_{r=1}^{m} \big[g_r^- \leqslant B_r(X,\bar{X}) \leqslant g_r^+\big]$, поскольку правая часть (1.39) есть ни что иное, как H_L^ℓ / C_L^ℓ. Разобьём в этом выражении сумму по N на две суммы — по N_0 и по N_1, и ещё заметим, что $B_m(X,\bar{X}) = s$ для всех $(X,\bar{X}) \in N$:

$$\begin{aligned}
H_m^s &= \underbrace{\sum_{(X,\bar{X})\in N_0} \prod_{r=1}^{m-1} \big[g_r^- \leqslant B_r(X,\bar{X}) \leqslant g_r^+\big]}_{H_{m-1}^s} \big[g_m^- \leqslant s \leqslant g_m^+\big] + \\
&+ \underbrace{\sum_{(X,\bar{X})\in N_1} \prod_{r=1}^{m-1} \big[g_r^- \leqslant B_r(X,\bar{X}) \leqslant g_r^+\big]}_{H_{m-1}^{s-1}} \big[g_m^- \leqslant s \leqslant g_m^+\big] = \\
&= \big(H_{m-1}^s + H_{m-1}^{s-1}\big)\big[g_m^- \leqslant s \leqslant g_m^+\big].
\end{aligned}$$

Таким образом, получена рекуррентная формула для H_m^s, формально совпадающая с формулой усечённого треугольника Паскаля (1.34). Осталось только проверить граничные случаи.

При $m = 1$ и фиксированном $s \in \{0, 1\}$ имеется только одно разбиение, $|N| = 1$, следовательно, $H_1^s = \big[g_m^- \leqslant s \leqslant g_m^+\big]$, что совпадает с (1.34).

При $s = 0$ и произвольном $m = 1, \ldots, k$ имеется только одно разбиение, $|N| = 1$, причём ни один объект из X^m не попадает в X. Это означает, что $B_r = 0$ при всех $r = 1, \ldots, m$. Но тогда $H_m^0 = \prod_{r=1}^m \big[g_r^- \leqslant s \leqslant g_r^+\big]$, что, опять-таки, совпадает с (1.34).

Заметим также, что при $s = 0$ запись $G_{m-1}^{s-1} = 0$ по определению корректна, в то же время $H_{m-1}^{s-1} = 0$, поскольку $N_1 = \varnothing$. Аналогично, при $s = m$ имеем $N_0 = \varnothing$, следовательно, $H_{m-1}^s = 0 = G_{m-1}^s$.

3. Односторонние оценки (1.35) и (1.36) выводится аналогично. Различие в том, что для них выражение под знаком произведения в (1.39) принимает вид, соответственно, либо $\big[0 \leqslant B_m \leqslant g_m^+(\varepsilon)\big]$, либо $\big[g_m^-(\varepsilon) \leqslant B_m \leqslant m\big]$. Изменяется только форма границы в усечённом треугольнике Паскаля, соответственно, либо нижней $g_m^-(\varepsilon) = 0$, либо верхней $g_m^+(\varepsilon) = m$, и все дальнейшие рассуждения остаются в силе. ■

Геометрическая интерпретация. Вторую часть доказательства (после формулы (1.39)) можно провести гораздо короче и нагляднее, пользуясь следующими геометрическими соображениями.

Каждое разбиение $X \sqcup \bar{X} = \mathbb{X}$ взаимно однозначно соответствует бинарному вектору $b = (b_1, \ldots, b_L)$, состоящему из ℓ единиц и k нулей, и, в то же время, некоторой траектории, проходящей из точки $(0, 0)$ в точку (L, ℓ) согласно правилу: если $b_i = 1$, то сместиться вправо и вверх на 1; если $b_i = 0$, то сместиться вправо на 1, см. рис. 1.5. Очевидно, траектория состоит из всех точек $(m, B_m)_{m=0}^L$. Выполнение совокупности условий $\big[g_m^- \leqslant B_m(X, \bar{X}) \leqslant g_m^+\big]$ при всех $m = 1, \ldots, L$ означает, что траектория не может проходить ниже границы g_m^- или выше границы g_m^+. На рис. 1.5 эти границы показаны линиями. Согласно (1.39) функционал $\mathsf{P}\big[D(X, \bar{X}) \leqslant \varepsilon\big]$ в точности равен доле таких траекторий. Будем называть их *допустимыми*. Обозначим через H_m^s число допустимых траекторий, проходящих из точки $(0, 0)$ в точку (m, s). Допустимая траектория может придти в (m, s) либо из $(m-1, s-1)$, либо из $(m-1, s)$. Отсюда следует рекуррентная формула для числа допустимых траекторий: $H_m^s = H_{m-1}^s + H_{m-1}^{s-1}$. Однако, если $s \notin \big[g_m^-, g_m^+\big]$, то все такие траектории уже не будут допустимыми, поэтому окончательная формула принимает вид $H_m^s = \big(H_{m-1}^s + H_{m-1}^{s-1}\big)\big[g_m^- \leqslant s \leqslant g_m^+\big]$, что совпадает с определением усечённого треугольника Паскаля: $H_m^s \equiv G_m^s$.

Практическое вычисление по рекуррентным соотношениям (1.34) сталкивается с проблемой переполнения: значения G_m^s выходят за пределы разрядной сетки современных компьютеров при L порядка нескольких сотен. Проблема снимается, если вывести рекуррентную формулу для отношений $\varphi_m^s = G_m^s / C_m^s$, которые принимают значения из отрезка $[0, 1]$. Применив тождества $C_m^s = \frac{m}{m-s} C_m^{s-1} = \frac{m}{s} C_{m-1}^{s-1}$, получим:

$$\varphi_m^s = \frac{m - s}{m} \varphi_{m-1}^s + \frac{s}{m} \varphi_{m-1}^{s-1}.$$

Усечённый треугольник Паскаля оказывается полезной концепцией не только при выводе точного выражения для критерия Смирнова, но во многих задачах, связанных со случайными блужданиями при ограничениях. Упомянем только выборочный контроль качества [4] и анализ выживаемости [104].

1.3.3 Обобщение на случай вариационного ряда со связками

В Теореме 1.5 (Смирнова) требование непрерывности функции распределения является существенным. В сильной аксиоматике оно гарантирует, что с вероятностью 1 вариационный ряд $\xi(x^{(1)}) < \xi(x^{(2)}) < \cdots < \xi(x^{(L)})$ не содержит одинаковых элементов. Это нужно для того, чтобы ранжировка была определена единственным образом. Когда условие непрерывности нарушается, равенства (1.32) и (1.33) могут не выполняться [5].

В Теореме 1.6 требование различности всех элементов вариационного ряда формулировалось в явном виде. Покажем, оставаясь в рамках слабой аксиоматики, что отказ от этого требования не сильно меняет вид результата — в Теореме 1.6 изменяются только границы усечения $[g_m^-, g_m^+]$. Следующая теорема обобщает критерий Смирнова на случай дискретных распределений.

Теорема 1.7. *Пусть $\xi\colon \mathbb{X} \to \mathbb{R}$ — произвольная функция, $\mathbb{X}$ — произвольная конечная выборка, вариационный ряд значений $\xi(x_i)$ состоит из H связок:*

$$\underbrace{\xi(x^{(1)}) = \cdots = \xi(x^{(i_1)})}_{\text{1-я связка}} < \underbrace{\xi(x^{(i_1+1)}) = \cdots = \xi(x^{(i_2)})}_{\text{2-я связка}} < \cdots < \underbrace{\xi(x^{(i_{H-1}+1)}) = \cdots = \xi(x^{(i_H)})}_{\text{H-я связка}}.$$

Тогда в слабой аксиоматике справедливы точные оценки (1.35), (1.36), (1.37), *если взять границы усечённого треугольника Паскаля $[\tilde{g}_m^-, \tilde{g}_m^+]$:*

$$\tilde{g}_m^+(\varepsilon) = \min\bigl\{g_{i_{h-1}}^+(\varepsilon) + m - i_{h-1},\ g_{i_h}^+(\varepsilon)\bigr\};$$
$$\tilde{g}_m^-(\varepsilon) = \max\bigl\{g_{i_{h-1}}^-(\varepsilon),\ g_{i_h}^-(\varepsilon) + m - i_h\bigr\};$$

для всех $m = i_{h-1}{+}1, \ldots, i_h$, где h пробегает значения от 1 до H, $i_0 = 0$, $i_H = L$.

Доказательство в целом аналогично доказательству Теоремы 1.6, поэтому остановимся только на различиях.

Доказательство. Рассмотрим выражение (1.38). Как и прежде, изменим порядок слагаемых, просуммировав их в порядке возрастания значений $\xi(x_i)$:

$$D(X, \bar{X}) = \max_{z\in\mathbb{R}} \left| \frac{1}{k}\sum_{i=1}^{L}(1-b_i)\bigl[\xi(x^{(i)}) < z\bigr] - \frac{1}{\ell}\sum_{i=1}^{L} b_i\bigl[\xi(x^{(i)}) < z\bigr] \right|.$$

Максимум достаточно брать не по всем $z \in \mathbb{R}$, а лишь по конечному множеству значений, которые функция ξ принимает на выборке, $z \in \bigl\{\xi(x^{(1)}), \ldots, \xi(x^{(H)})\bigr\}$. Уберём сомножитель $\bigl[\xi(x^{(i)}) < z\bigr]$, заменив верхний предел суммирования L на $m = \max\bigl\{i\colon \xi(x^{(i)}) < z\bigr\}$. Заметим, что все объекты одной связки либо вместе входят, либо вместе не входят в сумму по i. Поэтому число m может принимать значения только из множества $I_H = \{i_1, \ldots, i_H\}$:

$$D(X, \bar{X}) \;=\; \max_{m\in I_H} \left| \frac{1}{k}\sum_{i=1}^{m}(1-b_i) - \frac{1}{\ell}\sum_{i=1}^{m} b_i \right| \;=\; \frac{L}{\ell k} \max_{m\in I_H} \left| B_m - \frac{m\ell}{L} \right|.$$

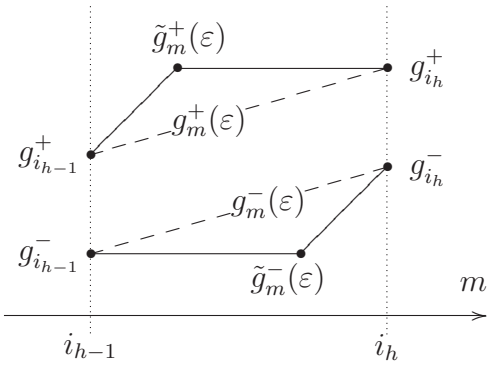


Рис. 1.10. Верхние $\tilde g^+_m(\varepsilon)$ и нижние $\tilde g^+_m(\varepsilon)$ границы усечённого треугольника Паскаля на отрезке $m = \{i_{h-1}, \ldots, i_h\}$, соответствующем h-й связке.

Аналогично (1.39), получаем:

$$\mathsf{P}\big[D(X, \bar X) \leqslant \varepsilon\big] = \mathsf{P} \prod_{m \in I_H} \big[g^-_m(\varepsilon) \leqslant B_m \leqslant g^+_m(\varepsilon)\big].$$

Единственное отличие от (1.39) заключается в том, что прохождение допустимых траекторий $\big(m, B_m\big)_{m=0}^L$ ограничено сверху $g^+_m(\varepsilon)$ и снизу $g^-_m(\varepsilon)$ не во всех точках $m = 0, \ldots, L$, а только в точках $m \in I_H$, соответствующих концам связок.

Рассмотрим допустимые траектории на отрезке $m = \{i_{h-1}, \ldots, i_h\}$, соответствующем h-й связке, см. рис. 1.10.

Между точками верхней границы $\big(i_{h-1}, \lfloor g^+_{i_{h-1}} \rfloor\big)$ и $\big(i_h, \lfloor g^+_{i_h} \rfloor\big)$ допустимая траектория может идти произвольным образом, следовательно, её путь ограничен сверху горизонтальной прямой $B_m \leqslant g^+_{i_h}$ и наклонной прямой $B_m \leqslant g^+_{i_{h-1}} + (m - i_{h-1})$.

Между точками нижней границы $\big(i_{h-1}, \lceil g^-_{i_{h-1}} \rceil\big)$ и $\big(i_h, \lceil g^-_{i_h} \rceil\big)$ допустимая траектория может идти произвольным образом, следовательно, её путь ограничен снизу горизонтальной прямой $B_m \geqslant g^-_{i_{h-1}}$ и наклонной прямой $B_m \geqslant g^-_{i_h} + (m - i_h)$.

Таким образом, получены границы $[\tilde g^-_m, \tilde g^+_m]$ усечённого треугольника Паскаля.

Теорема доказана. ■

Замечание 1.9. Если все связки одноэлементные, $\{i_1, \ldots, i_H\} \equiv \{1, \ldots, L\}$, то $\tilde g^+_m(\varepsilon) = g^+_m(\varepsilon)$, $\tilde g^-_m(\varepsilon) = g^-_m(\varepsilon)$, и Теорема 1.7 переходит в Теорему 1.6.

Замечание 1.10. Полученные оценки являются точными, но ненаблюдаемыми. Модифицированные границы $[\tilde g^-_m, \tilde g^+_m]$ существенно зависят от последовательности $i_1, \ldots, i_H$, которая строится по всей генеральной выборке $\mathbb{X}$; её невозможно знать, имея лишь наблюдаемую выборку X. Это означает, что Теорему 1.7 можно применять для проверки гипотезы однородности, однако непосредственно она не годится для эмпирического предсказания.

1.4 Некоторые непараметрические критерии и доверительные оценки

Цель данного параграфа — продемонстрировать возможности слабой аксиоматики на примере решения стандартных задач математической статистики.

Рассматриваются аналоги известных непараметрических статистических тестов и доверительных оценок. В слабой аксиоматике изменяется методология получения оценок. Сначала выводятся точные комбинаторные формулы и результат формулируется в терминах конечных выборок. Затем, при необходимости, к полученным оценкам применяется предельный переход $L \to \infty$. Его можно понимать лишь как способ получения асимптотических оценок, не подразумевая существования каких-то реальных выборок, длина которых устремляется в бесконечность. Асимптотические оценки, получаемые таким способом, во всех случаях совпадают с классическими.

1.4.1 Доверительное оценивание

Рассмотрим Задачу 1.4 (стр. 18) о доверительном оценивании.

Задана функция $\xi\colon \mathbb{X} \to \mathbb{R}$, значения которой попарно различны на элементах выборки $\mathbb{X}$. Требуется построить по наблюдаемой выборке X семейство вложенных *доверительных интервалов* $\Omega_\varepsilon(X) = \big[\xi_\varepsilon^-(X), \xi_\varepsilon^+(X)\big]$ такое, что для произвольного скрытого объекта $\bar{x}$ выполняется $\xi(\bar{x}) \in \Omega_\varepsilon(X)$ с вероятностью не менее $1 - \eta(\varepsilon)$.

Вариационным рядом функции ξ на выборке $U = \{u_1, \ldots, u_t\} \subseteq \mathbb{X}$ называется последовательность значений $\xi(u_1), \ldots, \xi(u_t)$, упорядоченная по возрастанию. Обозначим s-е значение вариационного ряда ξ на U через $\xi_U^{(s)}$, тогда $\xi_U^{(1)} < \cdots < \xi_U^{(t)}$.

Теорема 1.8. *Определим семейство вложенных отрезков* $\Omega_\varepsilon(X) = \big[\xi_X^{(\ell-\varepsilon+1)}, \xi_X^{(\varepsilon)}\big]$, *где* $\varepsilon = \lceil \ell/2 \rceil, \ldots, \ell$. *Тогда справедлива точная оценка:*

$$\mathsf{P}\big[\xi(\bar{x}) \notin \Omega_\varepsilon(X)\big] = 2\big(1 - \tfrac{\varepsilon}{L}\big), \quad \varepsilon = \lceil \ell/2 \rceil, \ldots, \ell. \tag{1.40}$$

Доказательство. Всего имеется $C_L^1 = L$ разбиений. Величина $\xi(\bar{x})$ превосходит $\xi_X^{(\varepsilon)}$ на тех разбиениях, при которых правее $\xi(\bar{x})$ в вариационном ряду находится менее $L - \varepsilon$ объектов. Таких разбиений ровно $L - \varepsilon$. Аналогично, $\xi(\bar{x}) < \xi_X^{(L-\varepsilon+1)}$ на тех разбиениях, при которых левее $\xi(\bar{x})$ находится менее $L - \varepsilon$ объектов. Таких разбиений также ровно $L - \varepsilon$. Итак, доля разбиений, при которых значение $\xi(\bar{x})$ попадает вне отрезка $\Omega_\varepsilon(X)$, составляет $2(L - \varepsilon)/L$. ■

Аналогично, справедлива точная верхняя оценка:

$$\mathsf{P}\big[\xi(\bar{x}) > \xi_X^{(\varepsilon)}\big] = 1 - \tfrac{\varepsilon}{L}, \quad \varepsilon = \lceil \ell/2 \rceil, \ldots, \ell. \tag{1.41}$$

Полагая в (1.40) $\varepsilon = \ell$, заключаем, что скрытое значение $\xi(\bar{x})$ выходит за пределы диапазона наблюдаемых значений $\big[\xi_X^{(1)}, \xi_X^{(\ell)}\big]$ с вероятностью $\frac{2}{L}$. Для предсказания $\xi(\bar{x})$ с надёжностью η достаточно иметь $\frac{1}{\eta} - 1$ объектов в случае односторонней оценки, и примерно вдвое больше, $\frac{2}{\eta} - 1$, для двусторонней. В частности, 19 объектов достаточно для получения верхней оценки с надёжностью 0.95.

1.4.2 Доверительные интервалы для квантилей

Доверительные интервалы для квантилей в сильной аксиоматике [41]. Пусть ξ — случайная величина с непрерывной строго возрастающей функцией распределения $F(x)$. Решение уравнения $F(x) = \alpha$ при $\alpha \in (0,1)$ существует и единственно, обозначается ξ_α и называется *α-квантúлью распределения* или квантилью порядка α. В частности, $\xi_{\frac{1}{2}}$ есть *медиана распределения*.

Пусть $X = \{\xi_1, \ldots, \xi_\ell\}$ — выборка ℓ независимых случайных величин из распределения F. Построим её вариационный ряд $\xi_X^{(1)} < \cdots < \xi_X^{(\ell)}$. С вероятностью 1 он не имеет связок. Отрезок $\left[\xi_X^{(r)}, \xi_X^{(s)}\right]$ является доверительным интервалом для α-квантили с доверительной вероятностью

$$\mathsf{P}\left[\xi_X^{(r)} \leqslant \xi_\alpha \leqslant \xi_X^{(s)}\right] = \sum_{t=r}^{s} C_\ell^t \alpha^t (1-\alpha)^{\ell-t}. \tag{1.42}$$

В частности, отрезок $\left[\xi_X^{(r+1)}, \xi_X^{(\ell-r)}\right]$ является доверительным интервалом для медианы ($\alpha = \frac{1}{2}$) с доверительной вероятностью $1 - 2\sum\limits_{t=0}^{r} C_\ell^t 2^{-\ell}$.

Доверительные интервалы для квантилей в слабой аксиоматике. Постановку задачи придётся менять, поскольку понятия α-квантили и функции распределения $F(x)$ определяются через теоретико-мерную вероятность. Адекватной заменой квантили ξ_α в слабой аксиоматике является m-й член вариационного ряда генеральной выборки $\mathbb{X}$ при $m = \alpha L$. Доверительное оценивание этой величины является ещё одним примером *задачи эмпирического предсказания*.

Допустим, что числовая генеральная выборка $\mathbb{X} = \{\xi_1, \ldots, \xi_L\}$, состоит из попарно различных значений. Построим её вариационный ряд $\xi_{\mathbb{X}}^{(1)} < \cdots < \xi_{\mathbb{X}}^{(L)}$.

Следующая теорема утверждает, что, имея наблюдаемую выборку X, можно предсказывать значение $M = \xi_{\mathbb{X}}^{(m)}$ и оценивать точность предсказаний.

Теорема 1.9. *Отрезок $\left[\xi_X^{(r)}, \xi_X^{(s)}\right]$ является доверительным интервалом для величины $M = \xi_{\mathbb{X}}^{(m)}$ с доверительной вероятностью*

$$\mathsf{P}\left[\xi_X^{(r)} \leqslant M \leqslant \xi_X^{(s)}\right] = \sum_{t=r}^{s} h_L^{\ell,m}(t), \tag{1.43}$$

где вероятность P понимается в соответствии со слабой аксиоматикой, как доля разбиений генеральной выборки.

Доказательство.

Во-первых, заметим, что $\mathsf{P}\left[\xi_X^{(r)} \leqslant M \leqslant \xi_X^{(s)}\right] = \mathsf{P}\left[\xi_X^{(r)} \leqslant M\right] - \mathsf{P}\left[\xi_X^{(s)} < M\right]$.

Рассмотрим событие $S_m = \left\{\xi_{\mathbb{X}}^{(1)}, \ldots, \xi_{\mathbb{X}}^{(m)}\right\} \subseteq \mathbb{X}$.

Условие $\mathsf{P}\left[\xi_X^{(r)} \leqslant M\right]$ равносильно тому, что в выборку X попадает не менее r элементов события S_m. Согласно Лемме 1.2 и определению «правого хвоста» гипергеометрического распределения (1.22) отсюда следует

$$\mathsf{P}\left[\xi_X^{(r)} \leqslant M\right] = \bar{H}_L^{\ell,m}(r).$$

Аналогично,

$$\mathsf{P}\big[\xi_X^{(s)} < M\big] = \bar{H}_L^{\ell,m}(s+1).$$

Таким образом,

$$\mathsf{P}\big[\xi_X^{(r)} \leqslant M \leqslant \xi_X^{(s)}\big] = \bar{H}_L^{\ell,m}(r) - \bar{H}_L^{\ell,m}(s+1) = \\ = \sum_{t=r}^{s_1} h_L^{\ell,m}(t) - \sum_{t=s+1}^{s_1} h_L^{\ell,m}(t) = \sum_{t=r}^{s} h_L^{\ell,m}(t).$$

Теорема доказана. ■

Предельный переход. Доверительная вероятность в слабой аксиоматике (1.43) выражается через гипергеометрическое распределение. Покажем, что при стремлении длины скрытой выборки к бесконечности оно стремится к биномиальному распределению, что приводит к классической доверительной вероятности (1.42).

При $L \to \infty$, фиксированной длине наблюдаемой выборки ℓ, фиксированных r и s и сохранении постоянного отношения $\alpha = \frac{m}{L}$ имеем:

$$h_L^{\ell,m}(t) = C_\ell^t \frac{C_k^{m-t}}{C_L^m} = C_\ell^t \frac{m!}{(m-t)!}\frac{k!}{L!}\frac{(L-m)!}{(k-m+t)!} \to C_\ell^t \frac{m^t (k-m)^{\ell-t}}{L^\ell} = \\ = C_\ell^t \Big(\frac{m}{L}\Big)^t \Big(\frac{k}{L} - \frac{m}{L}\Big)^{\ell-t} \to C_\ell^t \alpha^t (1-\alpha)^{\ell-t}.$$

Таким образом, перенос доверительных оценок в слабую аксиоматику потребовал переформулировать постановку задачи. В слабой аксиоматике доверительный интервал зависит от длины скрытой выборки, то есть от того, сколько ещё наблюдений предстоит сделать. В асимптотике, при стремлении этого числа в бесконечность, получается классический доверительный интервал в сильной аксиоматике.

1.4.3 Критерий знаков

Классический критерий знаков проверяет гипотезу H_0 о том, что в выборке бинарных величин $X = \{b_i\}_{i=1}^{\ell}$, $b_i \in \{0,1\}$, единицы появляются с вероятностью $p = \frac{1}{2}$, или, другими словами, что выборка подчиняется биномиальному распределению с параметром $p = \frac{1}{2}$.

Хрестоматийный пример применения критерия знаков — проверка симметричности монеты по последовательности выпадения «орлов» и «решек».

В практических задачах анализа данных критерий знаков применяют для проверки равенства нулю медианы вещественной случайной величины ξ. Для этого переходят к бинарной выборке $b_i = [\xi_i > 0]$. Другое применение — проверка *гипотезы сдвига* $\xi' = \xi + \delta$ в двух связных выборках, где δ — величина сдвига. Для этого переходят к бинарной выборке $b_i = [\xi'_i > \xi_i + \delta]$. В частности, при $\delta = 0$ проверяется отсутствие сдвига, или, как ещё говорят, *отсутствие эффекта обработки*.

Для выполнения теста вычисляется статистика $T(X) = \sum\limits_{i=1}^{\ell} b_i$. При условии, что гипотеза H_0 верна, статистика T подчиняется *биномиальному распределению*:

$$\mathsf{P}\big[T(X) = t\big] = C_\ell^t p^t (1-p)^{\ell-t} = C_\ell^t \, 2^{-\ell}.$$

Гипотеза H_0 отвергается на *уровне значимости* α, если значение статистики $T(X)$ попадает в критическую область — «хвосты» биномиального распределения,

$$\frac{1}{2^\ell} \sum_{t=0}^{T(X)} C_\ell^t \;\notin\; \big[\tfrac{\alpha}{2},\, 1 - \tfrac{\alpha}{2}\big]. \tag{1.44}$$

Критерий знаков в слабой аксиоматике строится по сути так же, только формулировка гипотезы H_0 уже не может опираться на понятие «вероятность единицы» и биномиальное распределение.

Теперь гипотеза H_0 формулируется так: «наблюдаемая выборка X получена в результате случайного разбиения $X \sqcup \bar{X} = \mathbb{X}$ генеральной выборки $\mathbb{X} = \{b_i\}_{i=1}^{L}$, содержащей ровно половину единиц, $m = \sum\limits_{i=1}^{L} b_i = \frac{L}{2}$», где L — чётное число.

Если гипотеза H_0 верна, то *основная аксиома* 1.1 также верна, и можно пользоваться слабой аксиоматикой. Применяя Лемму 1.2 к событию $S = \big\{b_i \in \mathbb{X}\colon\, b_i = 1\big\}$, приходим к выводу, что статистика $T(X)$ подчиняется *гипергеометрическому распределению*:

$$\mathsf{P}\big[T(X) = t\big] = h_L^{\ell,m}(t) = C_\ell^t \frac{C_{L-\ell}^{m-t}}{C_L^m}, \quad m = \frac{L}{2},$$

где вероятность P понимается в соответствии со слабой аксиоматикой, как доля разбиений генеральной выборки.

Далее стандартная логика проверки статистических гипотез переносится в слабую аксиоматику без изменений. Гипотеза H_0 отвергается на уровне значимости α, если значение $T(X)$ попадает в *критическую область*:

$$H_L^{\ell,m}\big(T(X)\big) \;\notin\; \big[\tfrac{\alpha}{2},\, 1 - \tfrac{\alpha}{2}\big],$$

Предельный переход. Параметр L был введён искусственно в результате мысленного достраивания наблюдаемой выборки X до генеральной выборки $\mathbb{X}$. Устремляя длину выборки $\mathbb{X}$ в бесконечность при $m = \frac{L}{2}$ и фиксированных ℓ, t, получаем всё то же биномиальное распределение, приводящее к классическому критерию (1.44):

$$\mathsf{P}\big[T(X) = t\big] = C_\ell^t \, \frac{m!}{(m-t)!} \frac{m!}{(m-\ell+t)!} \frac{k!}{L!} \to C_\ell^t \, \frac{m^t m^{\ell-t}}{L^\ell} = C_\ell^t \, \frac{m^\ell}{(2m)^\ell} = C_\ell^t \, 2^{-\ell}.$$

Таким образом, перенос критерия знаков в слабую аксиоматику потребовал лишь переформулировки нулевой гипотезы, не изменив результат по существу.

1.4.4 Критерий Уилкоксона–Манна–Уитни

Классический критерий Уилкоксона–Манна–Уитни — это непараметрический статистический критерий, обычно используемый для проверки *гипотезы однородности* — предположения, что две выборки $X = (x_1, \ldots, x_\ell)$ и $\bar{X} = (x'_1, \ldots, x'_k)$ взяты из одного распределения. Предполагается, что величины x, x' измеряются в количественной или порядковой шкале.

Строго говоря, нулевой гипотезой в данном критерии является более слабое предположение $H_0\colon \mathsf{P}(x < x') = \frac{1}{2}$. Поэтому он считается не достаточно мощным для проверки гипотезы однородности.

Для простоты в качестве альтернативы рассмотрим $H_1\colon \mathsf{P}(x < x') > \frac{1}{2}$ (проверка двусторонней альтернативы несколько сложнее).

То, что данный тест не всегда позволяет обнаружить неоднородность выборок, легко показать с помощью контрпримера. Возьмём две выборки $X, \bar{X}$ равной длины. Пусть X сосредоточена на отрезке $[1, 2]$, половина выборки $\bar{X}$ сосредоточена на отрезке $[0, 1]$, вторая половина — на $[2, 3]$. Тогда гипотеза H_0 справедлива, а гипотеза однородности, очевидно, нет.

Для выполнения теста вычисляется статистика

$$U(X, \bar{X}) = \sum_{i=1}^{\ell} \sum_{j=1}^{k} \big[x_i < x'_j\big].$$

Обычно используется более экономный способ вычисления U. Строится вариационный ряд $x^{(1)} < \cdots < x^{(L)}$ объединённой выборки $\mathbb{X} = X \sqcup \bar{X}$. Для простоты будем предполагать, что вариационный ряд не содержит связок. Далее находятся ранги $r(x_i) \in \{1, \ldots, L\}$ элементов первой выборки в общем вариационном ряду, и U выражается через их суммарный ранг:

$$U(X, \bar{X}) = \ell k + \frac{\ell(\ell+1)}{2} - \sum_{i=1}^{\ell} r(x_i).$$

Таким образом, данный критерий является примером *рангового критерия*.

Распределение статистики U асимптотически нормально, причём асимптотикой рекомендуется пользоваться уже при $\ell, k > 8$ [72, 5, 56]:

$$\tilde{U} = \frac{U - \frac{1}{2}\ell k}{\sqrt{\frac{1}{12}\ell k(\ell + k + 1)}} \sim \mathcal{N}(0, 1).$$

Если $\tilde{U} > \Phi_{1-\alpha}$, где $\Phi_{1-\alpha}$ — $(1-\alpha)$-квантиль нормального распределения, то гипотеза H_0 отвергается в пользу альтернативы H_1 при уровне значимости α.

Критерий Уилкоксона–Манна–Уитни в слабой аксиоматике. Аналогом гипотезы однородности в слабой аксиоматике является *основная аксиома* 1.1. Данный критерий позволяет проверить, могла ли пара выборок $X, \bar{X}$ быть получена в результате случайного разбиения исходной выборки $\mathbb{X}$ на две части. Соображения

о недостаточной мощности этого критерия и контрпример остаются в силе. Поэтому хотелось бы более точно сформулировать гипотезу H_0. Однако выдвигать условие $H_0\colon \mathsf{P}[x < x'] = \frac{1}{2}$ уже нельзя, поскольку в данном выражении знак вероятности P употреблён некорректно с точки зрения слабой аксиоматики.

Первая интерпретация гипотезы H_0. Допустим, что из подвыборок $X, \bar{X}$ случайно и равновероятно извлекается по одному объекту, x и x' соответственно. Тогда P надо понимать в смысле усреднения не только по всем разбиениям выборки $\mathbb{X}$, но и по всем объектам в подвыборках $X, \bar{X}$. В условиях аксиомы 1.1 вероятность $\mathsf{P}[x < x']$ действительно равна $\frac{1}{2}$:

$$\mathsf{P}[x < x'] = \mathsf{P}\,\frac{1}{\ell}\sum_{i=1}^{L}\big[x_i \in X\big]\frac{1}{k}\sum_{j=1}^{L}\big[x_j \in \bar{X}\big]\big[x_i < x_j\big] =$$

$$= \frac{1}{\ell k}\sum_{i=1}^{L}\sum_{j=1}^{L}\big[x_i < x_j\big]\underbrace{\mathsf{P}\big[x_i \in X\big]\big[x_j \in \bar{X}\big]}_{C_{L-2}^{\ell-1}/C_L^{\ell}} = \frac{1}{\ell k}\cdot\frac{L(L-1)}{2}\cdot\frac{C_{L-2}^{\ell-1}}{C_L^{\ell}} = \frac{1}{2}.$$

Вторая интерпретация гипотезы H_0 не связана искусственным введением дополнительной рандомизации. Выписанное выше выражение можно также трактовать как математическое ожидание $\frac{1}{\ell k}\mathsf{E}U(X, \bar{X}) = \frac{1}{2}$. Таким образом, гипотеза H_0 всего лишь утверждает, что отклонение статистики $U(X, \bar{X})$ от её ожидаемого значения не противоречит основной аксиоме 1.1.

В слабой аксиоматике легко выводится точная рекуррентная формула для распределения статистики U. Аналогичны образом она была получена и в исходных работах [222, 173].

Теорема 1.10. *Пусть все значения в выборке $\mathbb{X}$ попарно различны. Распределение статистики $U(X, \bar{X})$,*

$$P_{\ell,k}(u) = \mathsf{P}\big[U(X, \bar{X}) = u\big], \quad u \in \{0, \ldots, \ell k\},$$

не зависит от элементов выборок и вычисляется по рекуррентной формуле

$$P_{\ell,k}(u) = \tfrac{\ell}{L}P_{\ell-1,k}(u) + \tfrac{k}{L}P_{\ell,k-1}(u - \ell),$$

при начальных условиях

$$P_{\ell,0}(u) = P_{0,k}(u) = [u{=}0];$$
$$P_{\ell,k}(u) = 0 \;\; \textit{при} \;\; u \notin \{0, \ldots, \ell k\}.$$

Доказательство. Построим вариационный ряд элементов выборки $x^{(1)} < \cdots < x^{(L)}$. Он не содержит связок в силу условия попарной различности.

Обозначим $b_i = b_i(X) = [x^{(i)} \in X]$. Бинарная последовательность $b_1, \ldots, b_L$ содержит ровно ℓ единиц и k нулей. Тогда

$$U(X, \bar{X}) = \sum_{i=1}^{L}\sum_{j=1}^{L}[i < j]\, b_i\, \bar{b}_j,$$

где $\bar{b}_j = 1 - b_j$ — отрицание бинарной величины. Значение статистики U не зависит от значений элементов выборки $\mathbb{X}$, а только от разбиения $(X, \bar{X})$ и параметров длины подвыборок ℓ, k. Чтобы найти рекуррентную формулу для $U_{\ell,k} = U(X, \bar{X})$, запишем

$$U_{\ell,k} = \sum_{i=1}^{L-1} \sum_{j=1}^{L-1} [i < j]\, b_i\, \bar{b}_j + \bar{b}_L \sum_{i=1}^{L-1} b_i.$$

Если $b_L = 1$, то первое слагаемое в этой сумме есть $U_{\ell-1,k}$, а второе равно нулю. Если $b_L = 0$, то первое слагаемое в этой сумме есть $U_{\ell,k-1}$, а второе $\sum\limits_{i=1}^{L-1} b_i = \ell$.

Таким образом, $U_{\ell,k} = b_L U_{\ell-1,k} + \bar{b}_L (U_{\ell,k-1} + \ell)$.

Аналогичное рекуррентное соотношение выполняется и для вероятностей:

$$P_{\ell,k}(u) = \mathsf{P}\big[x^{(L)} \in X\big]\big[U_{\ell-1,k} = u\big] + \mathsf{P}\big[x^{(L)} \in \bar{X}\big]\big[U_{\ell,k-1} + \ell = u\big] = \\ = \tfrac{\ell}{L} P_{\ell-1,k}(u) + \tfrac{k}{L} P_{\ell,k-1}(u - \ell).$$

Начальные условия для $P_{\ell,k}(u)$ являются очевидным следствием начальных условий для U: $U_{\ell,0} = U_{0,k} = 0$ и ограничений на значения аргумента $u \in \{0, \ldots, \ell k\}$.

Теорема доказана. ■

Для проверки гипотезы H_0 строится распределение Уилкоксона

$$W(U) = \sum_{u=0}^{U} P_{\ell,k}(u).$$

Если $U > W_{1-\alpha}$, где $W_{1-\alpha}$ — $(1-\alpha)$-квантиль распределения Уилкоксона, то гипотеза H_0 отвергается в пользу альтернативы H_1 при уровне значимости α.

По всей видимости, не только критерий Уилкоксона–Манна–Уитни, но и более широкий класс ранговых критериев переносится в слабую аксиоматику без каких-либо дополнительных усилий.

1.5 Задача оценивания вероятности переобучения

В данном параграфе рассматривается Задача 1.6 об оценивании вероятности переобучения. В рамках слабой аксиоматики воспроизводятся основные результаты VC-теории — статистической теории восстановления зависимостей по эмпирическим данным, предложенной В. Н. Вапником и А. Я. Червоненкисом [10, 11, 8, 215]. В 1.5.5 вводится новое понятие *степени некорректности* метода обучения [22] и анализируется её влияние на вероятность переобучения. В 1.5.6 анализируются основные причины завышенности VC-оценок.

1.5.1 Основные понятия и определения

Пусть $\mathbb{X}$ — конечная *об*ъектов, A — множество *алгоритмов*, $I\colon A \times \mathbb{X} \to \{0, 1\}$ — бинарная функция, называемая *индикатором ошибки*:

$I(a, x) = \big[$алгоритм a допускает ошибку на объекте $x\big]$.

В *задачах обучения по прецедентам* с каждым объектом $x \in \mathbb{X}$ связан *правильный ответ* $y(x) \in \mathbb{Y}$, который известен для объектов наблюдаемой выборки и неизвестен для объектов скрытой выборки. Функция $y\colon \mathbb{X} \to \mathbb{Y}$ называется также *целевой зависимостью* (target function). Под «алгоритмами» понимаются функции того же вида $a\colon \mathbb{X} \to \mathbb{Y}$, предсказывающие правильный ответ $y(x)$ для произвольного объекта x. Индикатор ошибки характеризует точность предсказанного ответа и определяется исходя из природы множества допустимых ответов $\mathbb{Y}$.

Например, в задачах классификации множество $\mathbb{Y}$ конечно, поэтому индикатор ошибки чаще всего задаётся в виде

$$I(a,x) = \big[a(x) \neq y(x)\big], \quad \forall x \in \mathbb{X},\ \forall a \in A.$$

В задачах восстановления регрессии и прогнозирования, как правило, $\mathbb{Y} = \mathbb{R}$, и величину ошибки принято характеризовать вещественной *функцией потерь* (loss function), например, квадратичной: $\mathscr{L}(a,x) = \big(a(x) - y(x)\big)^2$. Тем не менее, можно определять и бинарные функции потерь вида

$$I(a,x) = \big[|a(x) - y(x)| \geqslant \delta(x)\big], \quad \forall x \in \mathbb{X},\ \forall a \in A,$$

где $\delta(x)$ — пороговый уровень, выше которого отклонение считается ошибкой. Использование бинарной функции потерь делает метод восстановления регрессии *робастным*, то есть нечувствительным к выбросам — редким, но большим отклонениям измеренного ответа y_i от неизвестного правильного ответа $y(x_i)$.

Для целей данного исследования нет необходимости конкретизировать понятие «алгоритма». Достаточно считать алгоритмы элементами некоторого абстрактного множества A, предполагая лишь, что существует способ определить, допускает ли алгоритм a ошибку на объекте x. Такое понимание «алгоритма» с одной стороны расширяет класс рассматриваемых задач, но с другой стороны ограничивает его теми задачами, в которых не столь важна величина отклонения $|a(x) - y(x)|$.

Бинарный вектор-столбец $\vec{a} = \big(I(a,x_i)\big)_{i=1}^{L}$ будем называть *вектором ошибок* алгоритма a. Совокупность всех попарно различных векторов ошибок, порождаемых алгоритмами $a \in A$, образует *матрицу ошибок* размера $L \times D$. Строки этой матрицы соответствуют объектам, столбцы — алгоритмам. Число столбцов D может быть меньше $|A|$, так как различные алгоритмы могут порождать одинаковые векторы ошибок. Множество алгоритмов A вполне может быть и бесконечными, однако число D попарно различных векторов ошибок всегда конечно и не превышает 2^D. В дальнейшем часто будет предполагаться, что A — это конечное множество, состоящее только из алгоритмов с попарно различными векторами ошибок.

Число ошибок алгоритма a на выборке $U \subseteq \mathbb{X}$ есть

$$n(a,U) = \sum_{x \in U} I(a,x).$$

Частотой ошибок алгоритма a на выборке $U \subseteq \mathbb{X}$ называется величина

$$\nu(a,U) = \frac{n(a,U)}{|U|} = \frac{1}{|U|} \sum_{x \in U} I(a,x).$$

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	
x_1	1	1	1	**0**	0	1	1	1	
x_2	0	0	0	**0**	1	1	0	0	
x_3	0	1	1	**0**	0	0	0	0	X — наблюдаемая обучающая выборка
x_4	1	0	1	**0**	1	0	1	0	
x_5	0	0	1	**0**	1	1	0	0	
x_6	0	0	0	1	1	1	0	0	
x_7	1	0	0	1	1	1	0	0	
x_8	0	0	0	1	0	0	0	1	$\bar{X}$ — скрытая контрольная выборка
x_9	0	1	1	1	1	1	0	0	
x_{10}	0	1	1	1	1	1	0	0	

Рис. 1.11. Пример матрицы ошибок, $L = 10,\ D = 8,\ \ell = k = 5$. Показано одно из C_{10}^5 разбиений выборки на наблюдаемую и скрытую подвыборки. Метод минимизации эмпирического риска выбирает алгоритм a_4 и является переобученным относительно данной пары выборок, причём при любом $\varepsilon \in (0, 1)$.

Методом обучения[1] называется отображение $\mu\colon \mathbb{X}^\ell \to A$, которое произвольной *обучающей выборке* $X \subset \mathbb{X}$ ставит в соответствие некоторый алгоритм $a = \mu X$ из A. Говорят также, что метод μ *восстанавливает* неизвестную *целевую зависимость* $y(x)$ по эмпирическим данным X [8].

Частота ошибок на обучающей выборке $\nu(\mu X, X)$ называется также *эмпирическим риском*. Обозначим через $A(X)$ подмножество алгоритмов a, на которых число ошибок $n(a, X)$ минимально. Метод μ называется методом *минимизации эмпирического риска*, если для любого $X \subset \mathbb{X}$

$$\mu X \in A(X) = \operatorname{Arg}\min_{a \in A} n(a, X) = \big\{a \in A\colon n(a, X) \leqslant n(a', X),\ \forall a' \in A\big\}. \tag{1.45}$$

Отклонением частоты ошибок алгоритма a на выборках X и $\bar{X}$ будем назвать разность $\delta(a, X, \bar{X}) = \nu(a, \bar{X}) - \nu(a, X)$.

Переобученностью метода μ относительно пары выборок X и $\bar{X}$ называется отклонение частоты ошибок алгоритма $a = \mu X$:

$$\delta_\mu(X, \bar{X}) = \delta(\mu X, X, \bar{X}) = \nu(\mu X, \bar{X}) - \nu(\mu X, X).$$

Если $\delta_\mu(X, \bar{X}) \geqslant \varepsilon$ при некотором достаточно малом $\varepsilon \in (0, 1)$, то говорят, что метод μ *переобучен* относительно пары выборок $X, \bar{X}$.

Пример 1.1. Матрица ошибок на рис. 1.11 разбита на обучающую и контрольную выборки так, что алгоритм a_4, минимизирующий эмпирический риск, допускает ошибки на всех объектах контрольной выборки. Это и есть переобучение.

---

[1] В англоязычной литературе метод обучения принято называть *алгоритмом обучения* (learning algorithm) [143], а алгоритм $a\colon \mathbb{X} \to \mathbb{Y}$ — классификатором (classifier), гипотезой (hypothesis), решающей функцией (decision function), либо просто функцией (function). Термин «алгоритм» как отображение из множества объектов во множество ответов употребляется в работах научной школы академика Ю. И. Журавлёва [48]. Термины «метод» и «алгоритм», обозначающие процедуру построения функции a по выборке данных употребляются в отечественной литературе вперемежку [11, 1, 51, 50].

Чтобы строить методы обучения μ, которые минимизировали бы не эмпирический риск, а вероятность переобучения, необходимо иметь достаточно точные её количественные оценки.

Основная постановка задачи — оценить *вероятность переобучения*:

$$Q_\varepsilon \equiv Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\delta_\mu(X, \bar{X}) \geqslant \varepsilon\big] = \\ = \mathsf{P}\big[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon\big]. \tag{1.46}$$

Наряду с данной задачей будем рассматривать задачу оценивания *вероятности большой частоты ошибок* на скрытой контрольной выборке:

$$R_\varepsilon \equiv R_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\nu(\mu X, \bar{X}) \geqslant \varepsilon\big]. \tag{1.47}$$

Замечания о природе переобучения. Неформально, переобучение — это чрезмерно точная подгонка алгоритма a под конкретную обучающую выборку X в ущерб его *обобщающей способности* (generalization ability). Ожидается, что метод μ обобщит частные факты, содержащиеся в выборке эмпирических данных X, обнаружит некие общие закономерности и построит алгоритм $a(x)$, приближающий неизвестную целевую зависимость $y(x)$ на всём множестве $\mathbb{X}$. Хотелось бы, чтобы алгоритм a ошибался как можно реже на новых данных $\bar{X} = \mathbb{X} \setminus X$, скрытых в момент обучения. Однако методу μ приходится выбирать алгоритм a из A, опираясь на *неполную информацию* $X \subset \mathbb{X}$. Поэтому выбор с некоторой вероятностью оказывается неверным. Таким образом, переобучение носит фундаментальный характер и связано с неполнотой информации в момент принятия решения. Однако величина переобученности может оказаться и несущественной, поэтому важной задачей является её количественное оценивание.

Ещё одно объяснение даёт следующий мысленный эксперимент. Пусть задано конечное множество из D алгоритмов, которые допускают на генеральной выборке $\mathbb{X}$ одно и то же число ошибок m, независимо друг от друга. Число ошибок любого из этих алгоритмов на обучающей выборке X подчиняется одному и тому же гипергеометрическому распределению. Выбирая алгоритм с минимальным числом ошибок s на обучающей выборке, мы фактически находим минимум из D независимых одинаково распределённых случайных величин. Математическое ожидание минимума уменьшается с ростом числа D. Следовательно, переобученность $\delta = \frac{m-s}{k} - \frac{s}{\ell} = \frac{m}{k} - s\frac{L}{\ell k}$ увеличивается с ростом D. Эти рассуждения остаются в силе и в общем случае, когда алгоритмы не являются независимыми (имеются схожие алгоритмы) и допускают различное число ошибок (имеется расслоение множества алгоритмов по уровням числа ошибок m). Тогда оценивать вероятность переобучения становится гораздо сложнее. Решению данной проблемы в основном и посвящено предлагаемое диссертационное исследование.

1.5.2 Простой частный случай: один алгоритм

Рассмотрим сначала простейший случай, когда метод μ по любой выборке $X \subset \mathbb{X}$ строит один и тот же алгоритм $a = \mu X$. Фактически это означает, что никакого обу-

чения нет. Для фиксированного алгоритма a требуется оценить отклонение частоты его ошибок на скрытой выборке от частоты ошибок на наблюдаемой выборке. Эта задача сводится к оцениванию частоты фиксированного события (см. стр. 33)

$$S_a = \big\{x_i \in \mathbb{X} \colon I(a, x_i) = 1\big\}.$$

Следующая теорема есть очевидная переформулировка Теоремы 1.3.

Теорема 1.11. *Пусть алгоритм a допускает m ошибок на генеральной выборке: $n(a, \mathbb{X}) = m$. Тогда для любого $\varepsilon \in [0, 1)$ справедливы точные оценки:*

$$Q_\varepsilon = \mathsf{P}\big[\nu(a, \bar{X}) - \nu(a, X) \geqslant \varepsilon\big] = H_L^{\ell,m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right); \tag{1.48}$$
$$R_\varepsilon = \mathsf{P}\big[\nu(a, \bar{X}) \geqslant \varepsilon\big] = H_L^{\ell,m}\big(m - \varepsilon k\big).$$

Далее будем оценивать как вероятность переобучения (функционал Q_ε), так и вероятность большой частоты ошибок на контроле (функционал R_ε). Оценки R_ε отличаются бóльшим числом слагаемых в гипергеометрическом распределении, следовательно, являются более грубыми. Это естественная плата за отказ от использования доступной информации о частоте ошибок на наблюдаемой выборке.

1.5.3 Коэффициенты разнообразия и профиль расслоения

Чтобы обобщить Теорему 1.11 на случай произвольного метода обучения μ, потребуется ввести несколько новых обозначений и понятий.

$\vec{A} = \big\{\vec{a} \colon a \in A\big\}$ — множество векторов ошибок, порождаемых заданным множеством алгоритмов A. Мощность $|\vec{A}|$ конечна, не превышает мощности множества A и не превышает 2^L — числа различных булевых векторов длины L.

$\Delta(A, \mathbb{X}) = |\vec{A}|$ — *коэффициент разнообразия* (shattering coefficient)[2] (shatter coefficient) множества алгоритмов A на выборке $\mathbb{X}$. В задачах классификации на два класса коэффициент разнообразия равен числу различных *дихотомий* (способов разделить выборку $\mathbb{X}$ на два класса), реализуемых всевозможными алгоритмами из A.

$A_L^\ell \equiv A_L^\ell(\mu, \mathbb{X}) = \big\{\mu X \colon X \subset \mathbb{X},\ |X| = \ell\big\}$ — множество алгоритмов, индуцируемых методом обучения μ на всевозможных обучающих подвыборках X. Мощность $|A_L^\ell|$ конечна и не превышает C_L^ℓ — числа различных разбиений $X \sqcup \bar{X} = \mathbb{X}$.

$\Delta_L^\ell \equiv \Delta_L^\ell(\mu, \mathbb{X}) = \Delta\big(A_L^\ell(\mu, \mathbb{X}), \mathbb{X}\big)$ — *локальный коэффициент разнообразия* (local shatter coefficient) метода μ на выборке $\mathbb{X}$. Локальный коэффициент разнообразия

[2] В исходных работах В. Н. Вапника и А. Я. Червоненкиса [10, 11, 8] коэффициент разнообразия назывался *индексом системы событий*. Алгоритм a индуцирует событие $S_a = \big\{x \in \mathbb{X} \;\big|\; I(a, x) = 1\big\}$. Семейство A индуцирует систему событий $S = \{S_a | \, a \in A\}$. Индекс системы событий S есть число различных подмножеств вида $S_a \cap \mathbb{X}$, где a пробегает всё множество A, что равносильно определению через $|\vec{A}|$. В англоязычных работах прижился термин shattering — число разбиений всеми возможными способами, буквальный перевод — «вдребезги». Другой вариант перевода термина *shattering* на русский язык — «дробление» [74].

не превосходит C_L^ℓ. Он может оказаться и строго меньше C_L^ℓ, поскольку метод μ может строить по различным выборкам совпадающие алгоритмы; кроме того, различные алгоритмы могут порождать одинаковые векторы ошибок.

$\Delta^A(L) = \max\limits_{\mathbb{X}} \Delta(A, \mathbb{X})$ — *глобальный коэффициент разнообразия* (global shatter coefficient), называемый также *функцией роста* (growth function) множества алгоритмов A [11, 215]. Максимум берётся по всевозможным выборкам $\mathbb{X} \subset \mathscr{X}$ длины L из некоторого (как правило, бесконечного) множества допустимых объектов $\mathscr{X}$. Функция роста является мерой сложности множества алгоритмов A. В отличие от локального коэффициента разнообразия, она не зависит ни от задачи (выборки $\mathbb{X}$ и восстанавливаемой зависимости $y(x)$), ни от метода обучения μ. Поэтому $\Delta^A(L)$ может оказаться существенно больше $\Delta_L^\ell(\mu, \mathbb{X})$. Справедлива верхняя оценка $\Delta^A(L) \leqslant 2^L$.

$A_m = \{a \in A \colon n(a, \mathbb{X}) = m\}$ — множество алгоритмов из A с m ошибками на генеральной выборке $\mathbb{X}$. Будем называть подмножества A_m *слоями* и говорить, что A *расслаивается по уровням ошибок*. Очевидно, $A = A_0 \sqcup \cdots \sqcup A_L$.

$\Delta_m \equiv \Delta_m(\mu, \mathbb{X}) = \Delta\big((A_L^\ell)_m, \mathbb{X}\big)$ — локальный коэффициент разнообразия m-го слоя множества алгоритмов $A_L^\ell(\mu, \mathbb{X})$. Совокупность величин $(\Delta_m)_{m=0}^L$ будем называть *профилем расслоения*[3]. Очевидно, $\Delta_L^\ell = \Delta_0 + \cdots + \Delta_L$.

1.5.4 Принцип равномерной сходимости и VC-оценка

Чтобы получить верхние оценки вероятности переобучения, справедливые для любого метода μ, в VC-теории и многочисленных последующих работах (см. обзоры [217, 108, 24]) применяется *принцип равномерной сходимости*. Функционал Q_ε заменяется его верхней оценкой $\widetilde{Q}_\varepsilon$ — вероятностью большого равномерного отклонения частот в двух подвыборках:

$$Q_\varepsilon \leqslant \widetilde{Q}_\varepsilon = \mathsf{P}\Big[\max_{a \in A_L^\ell} \delta(a, X, \bar{X}) \geqslant \varepsilon\Big]. \tag{1.49}$$

Когда говорят о «сходимости», имеют в виду, что $\widetilde{Q}_\varepsilon \to 0$ при $\ell, k \to \infty$, а «равномерность» означает, что величина $\delta(a, X, \bar{X})$ сходится к нулю одновременно для всех алгоритмов a из A_L^ℓ.

Следующая теорема играет центральную роль в VC-теории. Ниже приводится её доказательство, существенно более краткое, чем в [10, 11, 8].

Теорема 1.12. *Для любых μ, $\mathbb{X}$ и $\varepsilon \in [0, 1]$ справедлива оценка*

$$Q_\varepsilon \leqslant \Delta_L^\ell(\mu, \mathbb{X}) \max_{m=1,\ldots,L} H_L^{\ell,m}\left(s_m^-(\varepsilon)\right), \quad s_m^-(\varepsilon) = \left\lfloor \tfrac{\ell}{L}(m - \varepsilon k) \right\rfloor. \tag{1.50}$$

Доказательство. Заменим функционал Q_ε, согласно принципу равномерной сходимости, функционалом $\widetilde{Q}_\varepsilon$ и заметим, что максимум в (1.49) достаточно взять не по

[3] В статье [219] она называлась *профилем разнообразия* множества алгоритмов A на выборке $\mathbb{X}$.

всему множеству A_L^ℓ, а только по алгоритмам, неразличимым на выборке $\mathbb{X}$, т. е. по множеству векторов ошибок $\vec{A}_L^\ell$.

$$Q_\varepsilon \leqslant \widetilde{Q}_\varepsilon = \mathsf{P}\Big[\max_{\vec{a}\in\vec{A}_L^\ell} \delta(a, X, \bar{X}) \geqslant \varepsilon\Big] = \mathsf{P}\max_{\vec{a}\in\vec{A}_L^\ell}\big[\delta(a, X, \bar{X}) \geqslant \varepsilon\big]. \tag{1.51}$$

Оценим максимум бинарных величин $\big[\delta(a, X) \geqslant \varepsilon\big]$ их суммой (или можно говорить об оценке вероятности объединения событий суммой их вероятностей, называемой также union bound или *неравенством Буля*):

$$\widetilde{Q}_\varepsilon \leqslant \widetilde{\widetilde{Q}}_\varepsilon = \mathsf{P}\sum_{\vec{a}\in\vec{A}_L^\ell}\big[\delta(a, X, \bar{X}) \geqslant \varepsilon\big]. \tag{1.52}$$

Воспользуемся расслоением по уровням ошибок $A_L^\ell = \big(A_L^\ell\big)_0 \sqcup \cdots \sqcup \big(A_L^\ell\big)_L$ и заметим, что вероятность большого отклонения частот для отдельного алгоритма a уже известна, согласно Теореме 1.11:

$$\widetilde{\widetilde{Q}}_\varepsilon = \sum_{m=0}^{L}\sum_{\vec{a}\in(\vec{A}_L^\ell)_m} \mathsf{P}\big[\delta(a, X, \bar{X}) \geqslant \varepsilon\big] = \sum_{m=0}^{L} \Delta_m H_L^{\ell,m}\left(s_m^-(\varepsilon)\right). \tag{1.53}$$

Оценим сверху $H_L^{\ell,m}\left(s_m^-(\varepsilon)\right)$ и вынесем его за знак суммирования:

$$Q_\varepsilon \leqslant \widetilde{Q}_\varepsilon \leqslant \widetilde{\widetilde{Q}}_\varepsilon \leqslant \Delta_L^\ell \max_m H_L^{\ell,m}\left(s_m^-(\varepsilon)\right). \tag{1.54}$$

Теорема доказана. ■

Следствие 1.12.1. *Аналогичная оценка верна и для функционала R_ε: для любых μ, $\mathbb{X}$ и $\varepsilon \in [0, 1]$*

$$R_\varepsilon \leqslant \Delta_L^\ell(\mu, \mathbb{X}) \max_{m=1,\ldots,L} H_L^{\ell,m}\left(\lfloor m - \varepsilon k\rfloor\right).$$

VC-Оценка (1.50) имеет следующую простую интерпретацию. Вероятность переобучения не превышает вероятности большого отклонения частот для наихудшего алгоритма (максимум $H_L^{\ell,m}\left(s_m^-(\varepsilon)\right)$ достигается при $m \approx L/2$), умноженной на число алгоритмов с различными векторами ошибок.

В исходных работах [10, 11] вместо (1.49) используется ещё более грубая оценка — максимум берётся по всем алгоритмам исходного семейства алгоритмов A, включая и те, которые никогда не выбираются методом обучения. При этом функционал $\widetilde{Q}_\varepsilon$ фактически принимается за определение обобщающей способности, а понятия метода обучения и вероятности переобучения вообще не вводятся. Таким образом, завышенность оценок закладывается на уровне аксиоматики VC-теории.

Легко понять, что замена A_L^ℓ на A в функционале $\widetilde{Q}_\varepsilon$ приводит к замене в (1.50) локального коэффициента разнообразия на функцию роста:

$$\Delta_L^\ell(\mu, \mathbb{X}) \leqslant \Delta^A(L). \tag{1.55}$$

Наконец, функцию гипергеометрического распределения заменяют экспоненциальной верхней оценкой, имеющей особо простой вид при $\ell = k$ [8]:

$$\max_m H_L^{\ell,m}\left(s_m^-(\varepsilon)\right) \leqslant \tfrac{3}{2}e^{-\varepsilon^2\ell}, \quad \ell = k. \tag{1.56}$$

В итоге получается наиболее известная из VC-оценок [215]:

Следствие 1.12.2. *Для любых μ, $\mathbb{X}$, $\varepsilon \in [0,1)$ при $\ell = k$ справедлива оценка:*

$$Q_\varepsilon \leqslant \Delta^A(2\ell) \cdot \tfrac{3}{2}e^{-\varepsilon^2\ell}. \tag{1.57}$$

Понятие ёмкости (VC-размерности) семейства алгоритмов. Минимальное число h, при котором $\Delta^A(h) < 2^h$, называется *ёмкостью* или *размерностью Вапника-Червоненкиса* (VC-dimension) семейства алгоритмов A.

Если такого числа h не существует, то говорят, что ёмкость A бесконечна. Тогда можно лишь утверждать, что $\Delta^A(L) \leqslant 2^L$, что при подстановке в правую часть (1.57) даёт оценку, всегда бóльшую единицы. Следовательно, равномерной сходимости нет, и гарантировать хорошее качество обучения нельзя.

Если же A имеет конечную ёмкость h, то его функция роста ограничена сверху функцией, полиномиально растущей по L:

$$\Delta^A(L) \leqslant C_L^0 + C_L^1 + \cdots + C_L^h \leqslant \frac{3}{2}\frac{L^h}{h!}. \tag{1.58}$$

В этом случае имеет место равномерная сходимость, и можно говорить, что семейство A является обучаемым. Таким образом, согласно VC-теории, для оценивания качества обучения достаточно знать только длину выборки L и ёмкость h.

Оценивание ёмкости конкретных семейств часто оказывается сложной комбинаторной задачей. Практически сразу было доказано, что в задачах классификации на два класса при $\mathbb{X} = \mathbb{R}^n$ ёмкость семейства линейных разделяющих правил

$$a(x) = \left[\alpha_1 x^1 + \cdots + \alpha_n x^n > 0\right], \quad x = (x^1, \ldots, x^n) \in \mathbb{X},$$

равна числу параметров α_j, то есть размерности n линейного пространства, в котором строится разделяющая гиперплоскость. Оценки ёмкости получены для нейронных сетей [96, 92, 149, 181], решающих деревьев [42], корректных алгебраических замыканий АВО [66], комитетных решающих правил [178], и других семейств.

Ёмкость — нетривиальное понятие, и далеко не всегда она связана с числом параметров алгоритма. Известны примеры многопараметрических семейств ёмкости 1 и однопараметрических семейств бесконечной ёмкости [215].

Ёмкость семейств, основанных на явном хранении всей обучающей выборки, как правило, бесконечна (например, у алгоритма ближайших соседей). Ёмкость семейств, гарантирующих корректность (отсутствие ошибок) на обучающей выборке, также, как правило, бесконечна. Хотя, есть и исключения: в работах В. Л. Матросова построена композиция алгоритмов вычисления оценок, имеющая конечную ёмкость, и одновременно гарантирующая корректность [63, 66, 67, 68].

1.5.5 Степень некорректности и её влияние на переобучение

Алгоритм a называется *корректным на выборке* X, если $n(a, X) = 0$.

Метод μ называется *корректным на выборке* X, если $n(\mu X, X) = 0$.

В VC-теории отдельно рассматривается частный случай, когда метод μ корректен на любой обучающей выборке X. Этот случай называется *детерминистской постановкой задачи обучения*[4]. В то же время общая постановка задачи предполагает полное отсутствие какой-либо априорной информации о величине $n(\mu X, X)$. Чтобы исследовать промежуточные ситуации, введём понятие степени некорректности [22].

Определение 1.6. *Степенью некорректности метода обучения μ на выборке $\mathbb{X}$ будем называть максимальную частоту ошибок на всевозможных обучающих подвыборках длины ℓ:*

$$\sigma(\mu, \mathbb{X}) = \max_{X \in [\mathbb{X}]^\ell} \nu(\mu X, X).$$

Метод μ называется *корректным на генеральной выборке* $\mathbb{X}$, если $\sigma(\mu, \mathbb{X}) = 0$.

Для случая, когда степень некорректности ограничена сверху, следующая теорема уточняет оценку (1.50) Теоремы 1.12.

Теорема 1.13. *Для любых μ, $\mathbb{X}$ с ограниченной некорректностью $\sigma(\mu, \mathbb{X}) \leqslant \sigma$ и любого $\varepsilon \in [0, 1]$ справедлива оценка*

$$Q_\varepsilon \leqslant \Delta_L^\ell \max_{m \in M(\varepsilon,\sigma)} H_L^{\ell,m}\left(s_m^-(\varepsilon, \sigma)\right), \tag{1.59}$$

где $M(\varepsilon, \sigma) = \{m\colon \varepsilon k \leqslant m \leqslant k + \sigma\ell\}$, $s_m^-(\varepsilon, \sigma) = \min\{s_m^-(\varepsilon), \sigma\ell\}$.

Доказательство. Модифицируем функционал вероятности переобучения, добавив в него условие ограниченной некорректности, что не повлияет на его значение:

$$Q_\varepsilon = \mathsf{P}\big[\delta_\mu(X, \bar{X}) \geqslant \varepsilon\big] = \mathsf{P}\big[\delta_\mu(X, \bar{X}) \geqslant \varepsilon\big]\big[\nu(\mu X, X) \leqslant \sigma\big].$$

К этому модифицированному функционалу применим принцип равномерной сходимости аналогично тому, как это было сделано в доказательстве Теоремы 1.12:

$$Q_\varepsilon \leqslant \widetilde{Q}_\varepsilon = \mathsf{P} \max_{\vec{a} \in \vec{A}_L^\ell} \big[\delta(a, X, \bar{X}) \geqslant \varepsilon\big]\big[\nu(a, X) \leqslant \sigma\big].$$

Применим неравенство Буля (заменим максимум бинарных величин их суммой) и переставим местами знаки суммирования:

$$\widetilde{Q}_\varepsilon \leqslant \widetilde{\widetilde{Q}}_\varepsilon = \sum_{m=0}^{L} \sum_{\vec{a} \in (\vec{A}_L^\ell)_m} \mathsf{P}\big[\delta(a, X, \bar{X}) \geqslant \varepsilon\big]\big[\nu(a, X) \leqslant \sigma\big].$$

---

[4] В зарубежной литературе сложилась другая терминология. Детерминистскую постановку задачи называют реализуемым обучением (realizable learning), имея в виду, что с помощью семейства алгоритмов A возможно реализовать истинную зависимость $y(x)$. Общую постановку задачи называют нереализуемым или агностическим обучением (agnostic learning), подчёркивая принципиальную невозможность знать, находится ли истинная зависимость в семействе A, или нет.

Вставим во внутреннюю сумму тождество $\sum\limits_{s=0}^{\ell}\big[n(a,X)=s\big]=1$, затем воспользуемся расслоением по уровням ошибок $\vec{A}^\ell_L=\big(\vec{A}^\ell_L\big)_0\sqcup\cdots\sqcup\big(\vec{A}^\ell_L\big)_L$ и определением профиля расслоения $\Delta_m=\big|\big(\vec{A}^\ell_L\big)_m\big|$:

$$\widetilde{\widetilde{Q}}_\varepsilon=\sum_{m=0}^{L}\Delta_m\sum_{s=0}^{\ell}\left[\frac{m-s}{k}-\frac{s}{\ell}\geqslant\varepsilon\right]\left[\frac{s}{\ell}\leqslant\sigma\right]\mathsf{P}\big[n(a,X)=s\big].$$

Согласно Лемме 1.2 вероятность $\mathsf{P}\big[n(a,X)=s\big]$ равна гипергеометрической функции $h^{\ell,m}_L(s)$. В силу ограничения $\frac{s}{\ell}\leqslant\sigma$ все члены суммы по m при $m>\sigma\ell+k$ равны нулю, а в силу ограничения $0\leqslant s\leqslant\frac{\ell}{L}(m-\varepsilon k)$ все её члены при $m<\varepsilon k$ также равны нулю. Поэтому суммировать достаточно по $m\in M(\varepsilon,\sigma)$. Итак,

$$\widetilde{\widetilde{Q}}_\varepsilon=\sum_{m\in M(\varepsilon,\sigma)}\Delta_m\sum_{s=0}^{\ell}\big[s\leqslant\tfrac{\ell}{L}(m-\varepsilon k)\big]\big[s\leqslant\sigma\ell\big]\,h^{\ell,m}_L(s)\,. \tag{1.60}$$

Сумма по s представляет собой «левый хвост» гипергеометрического распределения $H^{\ell,m}_L\big(s^-_m(\varepsilon,\sigma)\big)$. Оценим её сверху и вынесем за знак суммирования по m, а сумму компонент профиля расслоения $\Delta_m(\mu,\mathbb{X})$ оценим сверху локальным коэффициентом разнообразия $\Delta^\ell_L(\mu,\mathbb{X})$:

$$Q_\varepsilon\leqslant\widetilde{Q}_\varepsilon\leqslant\widetilde{\widetilde{Q}}_\varepsilon\leqslant\Delta^\ell_L\max_{m\in M(\varepsilon,\sigma)}H^{\ell,m}_L\left(s^-_m(\varepsilon,\sigma)\right).$$

Теорема доказана. ■

Следствие 1.13.1. *Если метод μ корректен на генеральной выборке $\mathbb{X}$, то для любого $\varepsilon\in[0,1]$ справедлива цепочка оценок*

$$Q_\varepsilon\leqslant\sum_{m\in M(\varepsilon,0)}\Delta_m\frac{C^\ell_{L-m}}{C^\ell_L}\leqslant\Delta^A(L)\frac{C^\ell_{L-\lceil\varepsilon k\rceil}}{C^\ell_L}\leqslant\Delta^A(L)\left(\frac{k}{L}\right)^{\varepsilon k}. \tag{1.61}$$

Доказательство.

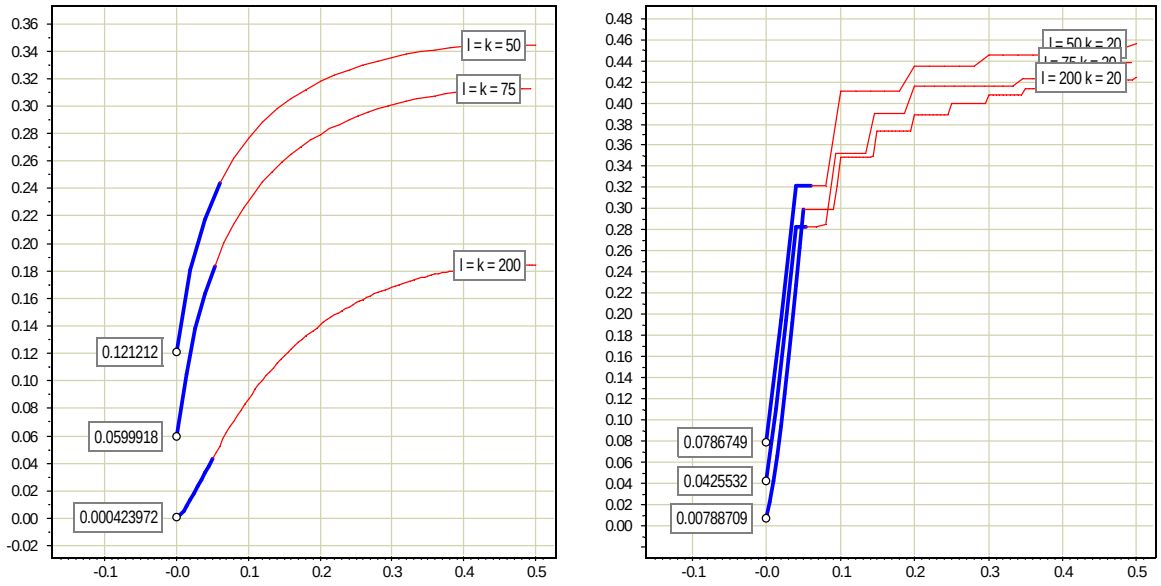
Первое неравенство следует из условия корректности $\sigma=0$, равенства (1.60) и $h^{\ell,m}_L(0)=\frac{C^\ell_{L-m}}{C^\ell_L}$.

Второе неравенство следует из того, что C^ℓ_{L-m} максимально, когда m минимально, а минимальное значение m в $M(\varepsilon,0)$ как раз и есть $\lceil\varepsilon k\rceil$.

Третье неравенство получается из $\frac{C^\ell_{L-m}}{C^\ell_L}=\prod\limits_{j=0}^{m-1}\frac{k-j}{L-j}\leqslant\big(\frac{k}{L}\big)^m$. ■

Следствие 1.13.2. *Если $\ell=k$ и метод μ корректен на генеральной выборке $\mathbb{X}$, то для любого $\varepsilon\in[0,1]$ справедлива оценка*

$$Q_\varepsilon\leqslant\Delta^A(2\ell)\,2^{-\varepsilon\ell}. \tag{1.62}$$



(а) при $\ell = k = 50, 75, 200$.

(б) при $\ell = 50, 75, 200; k = 20$.

Рис. 1.12. График зависимости $\Gamma_L^\ell(\varepsilon, \sigma)$ от степени некорректности σ при $\varepsilon = 0.05$. Жирной линией выделен интервал $0 \leqslant \sigma \leqslant \varepsilon$.

В случае $\sigma = 1$, когда априорной информации о корректности метода нет, оценка вероятности переобучения (1.59) переходит в (1.50).

В случае $\sigma = 0$, когда метод μ корректен, оценка вероятности переобучения Q_ε становится существенно более точной и принимает более простой вид.

Оценка (1.59) меньше оценки (1.50), главным образом, благодаря замене квантили гипергеометрического распределения $s_m^-(\varepsilon)$ на меньшее значение $s_m^-(\varepsilon, \sigma)$. Наибольший выигрыш достигается при $\sigma = 0$.

По мере увеличения σ гипергеометрический сомножитель в (1.59)

$$\Gamma_L^\ell(\varepsilon, \sigma) = \max_{m \in M(\varepsilon,\sigma)} H_L^{\ell,m}\left(s_m^-(\varepsilon, \sigma)\right)$$

возрастает очень быстро и достигает значительной величины, сравнимой с $\Gamma_L^\ell(\varepsilon, 1)$, уже при σ порядка ε, см. рис. 1.12.

Отсюда следует вывод о важности требования корректности. Очевидно, что для обеспечения корректности необходимо усложнять конструкцию алгоритмов. Согласно VC-теории это приводит к значительному увеличению функции роста $\Delta^A(L)$, на фоне которого эффект уменьшения комбинаторного множителя остаётся незаметным. Поэтому в VC-теории принято считать, что не следует добиваться безошибочной работы алгоритма на обучающем материале. С точки зрения комбинаторных оценок, усложнение конструкции алгоритма не обязательно приводит к существенному увеличению локальной функции роста $\Delta_L^\ell(\mu, \mathbb{X})$. С другой стороны, корректность резко уменьшает комбинаторный множитель.

Отметим, что требование корректности всегда было основополагающим в *алгебраическом подходе* к построению алгоритмических композиций, развиваемом научной школой академика РАН Ю. И. Журавлёва [48, 49].

1.5.6 Проблема завышенности VC-оценок

Основная проблема VC-оценок в том, что они чрезвычайно завышены — настолько, что их применение практически теряет смысл. Чтобы в этом убедиться, достаточно выполнить численный расчёт требуемой длины обучающей выборки ℓ как функции от ёмкости h, точности ε и уровня значимости (надёжности) Q_ε.

Результаты приведены в таблицах 1.1, 1.2, 1.3. Все данные вычислены для случая $\ell = k$. Первые две таблицы построены для недетерминистского случая ($\sigma = 1$), третья — для детерминистского ($\sigma = 0$). Первая таблица использует завышенные аппроксимации функции роста (1.58) и гипергеометрического сомножителя (1.56).

Основные выводы следующие.

1. В детерминистском случае требуемая длина обучения заметно меньше, но всё же она на несколько порядков превышает те характерные длины выборок, с которыми обычно приходится иметь дело на практике. Опыт решения прикладных задач показывает, что хорошая обучаемость возможна и по выборкам существенно меньшей длины.

2. Оценки в таблице 1.1 существенно хуже. Поэтому применение завышенных аппроксимаций функции роста (1.58) и гипергеометрического сомножителя (1.56) едва ли оправдано в компьютерных вычислениях.

3. Правые половины таблиц соответствуют значению $\eta = 1$ и показывают границу применимости VC-оценок. При меньших ℓ верхняя оценка вероятности вырождается — становится больше 1.

4. Сопоставление правой и левой половин таблиц позволяет сделать вывод о высокой чувствительности достаточной длины обучения к ёмкости h и точности ε, но относительно слабой чувствительности к выбору уровня значимости η.

5. Первая строка таблицы соответствует семейству из одного алгоритма, тогда $h = 0$. При этом достигается наилучшая возможная оценка. Однако этот случай не интересен с точки зрения статистического обучения.

6. Учёт априорной информации о степени некорректности уточняет VC-оценки, но не устраняет ни одного из основных факторов завышенности VC-оценок.

1.5.7 Причины завышенности VC-оценок

Причины завышенности видны из доказательства Теоремы 1.12, в котором сделаны три оценки сверху, а при получении Следствия 1.12.2 — ещё две. Те же факторы завышенности остаются и при учёте степени некорректности в Теореме 1.13.

Итого, имеется пять причин завышенности. Рассмотрим их подробнее.

	$\eta = 0.01$				$\eta = 1$			
h	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
0	60106	2404	601	150	14054	562	140	35
2	314692	9813	2149	460	265220	7786	1634	328
5	715120	21605	4631	961	665470	19565	4111	827
10	1386763	41427	8808	1806	1337061	39382	8287	1671
20	2733709	81218	17200	3504	2683987	79171	16677	3369
50	6780774	200844	42438	8616	6731042	198797	41916	8481
100	13530370	400406	84550	17149	13480635	398359	84027	17014

Таблица 1.1. Зависимость достаточной длины обучения ℓ от ёмкости h, точности ε и надёжности η для случая $\sigma = 1$ и оценки $Q_\varepsilon \leqslant \eta = \frac{3}{2}\frac{L^h}{h!} \cdot \frac{3}{2}\exp(-\varepsilon^2\ell)$. Это наименее точная оценка Теоремы 1.12, использующая аппроксимацию функции роста (1.58) и аппроксимацию гипергеометрического сомножителя (1.56).

	$\eta = 0.01$				$\eta = 1$			
h	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
0	35900	1440	360	91	506	20	10	5
2	259300	7619	1600	316	210035	5579	1089	186
5	632633	18260	3770	741	582841	16219	3250	610
10	1262928	36396	7521	1470	1213200	34320	6989	1335
20	2531001	72918	15069	2936	2481120	70820	14549	2805
50	6348132	182980	37821	7381	6298001	180900	37290	7250
100	7373100	295440	73821	14811	7373100	295440	73821	14671

Таблица 1.2. Зависимость достаточной длины обучения ℓ от ёмкости h, точности ε и надёжности η для случая $\sigma = 1$ и оценки $Q_\varepsilon \leqslant \eta = (C_L^0 + \cdots + C_L^h)\Gamma_L^\ell(\varepsilon, 1)$. Это также оценка Теоремы 1.12, но функция роста и гипергеометрический сомножитель вычисляются по точным формулам, без применения аппроксимаций.

	$\eta = 0.01$				$\eta = 1$			
h	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
0	800	140	70	35	100	20	10	5
2	2800	440	200	85	2000	300	120	45
5	6200	960	410	165	5400	800	330	125
10	11900	1820	770	310	11200	1660	690	270
20	23500	3560	1500	600	22700	3400	1420	555
50	58100	8780	3700	1465	57400	8620	3620	1425
100	107000	17480	7370	2915	107000	17320	7290	2875

Таблица 1.3. Зависимость достаточной длины обучения от ёмкости h, точности ε и надёжности η для случая σ=0 и оценки $Q_\varepsilon \leqslant \eta = (C_L^0 + \cdots + C_L^h)C_{L-\lceil\varepsilon k\rceil}^\ell / C_L^\ell$. Это оценка из Теоремы 1.13, в которой не используются аппроксимации.

1. Применение принципа равномерной сходимости. Следующая теорема показывает, что неравенство (1.49) является точным, когда множество алгоритмов A_L^ℓ не расслаивается по уровням ошибок.

Теорема 1.14. *Если метод μ минимизирует эмпирический риск, и все векторы $a \in A_L^\ell$ имеют одинаковый уровень ошибок $m = n(a, \mathbb{X})$, то верхняя оценка* (1.49) *обращается в точное равенство: $Q_\varepsilon = \widetilde{Q}_\varepsilon$.*

Доказательство. Минимизация эмпирического риска $\nu(a, X)$ при фиксированном m эквивалентна максимизации переобученности, поскольку

$$\delta(a, X, \bar{X}) = \frac{m - \ell\nu(a, X)}{k} - \nu(a, X) = \frac{m}{k} - \frac{L}{k}\nu(a, X).$$

Теорема доказана. ■

Если же множество A_L^ℓ расслаивается по уровням ошибок, то (1.49) может оказаться как точной, так и сильно завышенной верхней оценкой. В главе 4 будут рассмотрены оба примера: для интервала булева куба она остаётся точной (стр. 145), а для монотонной цепочки — сильно завышена (стр. 153). Возможная завышенность (1.49) связана с тем, что максимум переобученности достигается на алгоритмах a, у которых не только мало́ $s = n(a, X)$, но и велико $m = n(a, \mathbb{X})$.

В общем случае требование равномерной сходимости является чрезмерно сильным и даёт лишь достаточное условие обучаемости.

2. Применение неравенства Буля в (1.52) приводит к сильной завышенности, особенно, когда среди векторов ошибок имеется много похожих. На практике часто применяются *связные семейства* алгоритмов, в которых для каждого алгоритма $a \in A$ найдутся другие алгоритмы $a' \in A$ такие, что векторы ошибок алгоритмов a и a' отличаются только на одном объекте [205]. Связные семейства порождаются методами классификации с непрерывной по параметрам разделяющей поверхностью. Это линейные классификаторы, машины опорных векторов с непрерывными ядрами, нейронные сети с непрерывными функциями активации, решающие деревья с пороговыми условиями ветвления, и многие другие. Таким образом, сильная завышенность *неравенства Буля* проявляется как раз в тех случаях, которые наиболее распространены и интересны с практической точки зрения.

3. Пренебрежение профилем расслоения при переходе от (1.53) к (1.54). Гипергеометрическое распределение $H_m = H_L^{\ell,m}(s_m^-(\varepsilon))$ оценивается сверху максимумом по m, что и позволяет заменить векторную характеристику сложности — профиль расслоения Δ_m — одной скалярной характеристикой — локальным коэффициентом разнообразия $\Delta_L^\ell = \sum_m \Delta_m$. Данный фактор завышенности не очень существенный, поскольку и гипергеометрическое распределение H_m, и профиль расслоения Δ_m достигают максимальных значений в окрестности $m \approx L/2$. Большие погрешности оценивания $H_m < \max H_m$ при малых m компенсируются относительно малым весом Δ_m этих слагаемых в суммарной оценке $Q_\varepsilon = \sum_m \Delta_m H_m$.

Заметим, что, хотя оценка (1.53) и зависит от профиля расслоения, она уже не учитывает эффект расслоения в полной мере, поскольку доказательство опирается на принцип равномерной сходимости.

4. Пренебрежение эффектом локализации. Применение функции роста в (1.55) приводит к тому, что оценка перестаёт зависеть от конкретной выборки $\mathbb{X}$, восстанавливаемой зависимости $y(x)$ и метода обучения μ. Такая оценка ориентирована на худший случай (worst case bound), который, скорее всего, никогда не возникнет на практике.

В каждой конкретной задаче $\langle \mathbb{X}, y, \mu \rangle$ в результате обучения выбираются только те алгоритмы семейства A, которые в определённом смысле подходят для данной задачи. Остальные алгоритмы остаются незадействованными. Таким образом, задача индуцирует локальное подмножество $A_L^\ell(\mu, \mathbb{X}) \subset A$, которое может оказаться намного меньше A. Этот эффект будем называть *локализацией семейства алгоритмов*.

При решении прикладных задач эффект локализации возникает практически всегда. Это связано с универсальностью применяемых семейств алгоритмов A. Лишь малая доля алгоритмов семейства имеет низкий уровень ошибок для каждой конкретной задачи. Подавляющее большинство алгоритмов аппроксимируют совсем другие зависимости $y(x)$, и в данной задаче допускают около 50% ошибок. Эксперименты [167, 164] подтверждают, что профиль расслоения $\Delta(A_m, \mathbb{X})$, как правило, имеет форму узкого пика, сконцентрированного в средних слоях $m \approx L/2$. В то же время, метод обучения гораздо чаще выбирает алгоритмы из нижних слоёв.

Всё это означает, что для получения точных оценок вероятности переобучения необходимо каким-то образом оценивать $\mathsf{P}[\mu X = a]$ — вероятности получения отдельных алгоритмов $a \in A$ в результате обучения заданным методом μ. Хорошая обучаемость возможна даже в рамках очень широких семейств A, при условии, что эти вероятности быстро убывают с ростом номера слоя $m = n(a, \mathbb{X})$.

5. Применение экспоненциальной оценки гипергеометрического распределения в (1.56). Эта причина завышенности наименее существенна, так как использовать экспоненциальную оценку, вообще говоря, не обязательно, хотя она и даёт представление результата в удобной аналитической форме.

В следующих главах будут сделаны экспериментальные оценки основных факторов завышенности; будет показано, что *получение оценок приемлемой точности невозможно без совместного учёта эффектов расслоения и сходства алгоритмов*; будут предложены новые комбинаторные методы, дающие точные оценки вероятности переобучения.

1.6 Основные выводы

1. Вводится *слабая вероятностная аксиоматика*, основанная на единственном вероятностном предположении, что все разбиения конечной генеральной выборки на наблюдаемую и скрытую подвыборки равновероятны. При столь слабых

вероятностных допущениях уже удаётся доказывать многие фундаментальные факты теории вероятностей и математической статистики. В их числе закон больших чисел, сходимость эмпирических распределений, многие непараметрические статистические критерии и доверительные оценки.

2. Преимущество слабой аксиоматики в том, что она имеет дело только с финитарными вероятностями, не требует привлечения теории меры, позволяет получать не завышенные, не асимптотические оценки вероятностей, пользуясь лишь *точными комбинаторными методами*.
3. В теории статистического обучения (SLT) слабая аксиоматика естественным образом приводит к задаче оценивания *вероятности переобучения*. В то же время, в классической VC-теории принято оценивать вероятность равномерного уклонения частоты ошибок от их вероятности. Этот функционал является завышенной оценкой вероятности переобучения. Таким образом, при классическом подходе завышенность оценок закладывается в саму аксиоматику теории.
4. В отличие от классического подхода, предлагается существенным образом учитывать свойства конкретного *метода обучения* как отображения, которое произвольной обучающей выборке ставит в соответствие некоторый алгоритм.
5. Приводятся результаты численного расчёта, показывающие, что VC-оценки завышены чрезвычайно, даже при самых благоприятных дополнительных предположениях. В рамках слабой аксиоматики выводятся аналогичные оценки для вероятности переобучения, зависящие от *локального профиля разнообразия* и *степени некорректности* метода обучения. Это новые для SLT понятия. В отличие от стандартных характеристик сложности семейства алгоритмов, они характеризуют степень соответствия метода обучения и решаемой задачи. Тем не менее, радикального улучшения точности оценок они не дают. В следующих главах проводится более глубокий анализ причин завышенности.

Глава 2

Теория статистического обучения

В данной главе представлен обзор оценок обобщающей способности, начиная с теории Вапника-Червоненкиса (параграф 2.1), и заканчивая работами последних лет. Все эти оценки исходно были получены в рамках стандартной (колмогоровской) вероятностной аксиоматики, и именно так они излагаются в данной главе.

Особое внимание уделяется оценкам, показывающим, что обобщающая способность может не ухудшаться с ростом сложности семейства (параграф 2.2), а также оценкам, учитывающим эффекты расслоения и сходства в семействах алгоритмов.

Комбинаторный подход, развиваемый в данной работе, основан на оценивании вероятности переобучения $Q_\varepsilon(\mu, \mathbb{X})$. В следующей главе, сопоставив теоретические и эмпирические оценки данного функционала, мы придём к заключению, что без совместного учёта эффектов *расслоения* и *сходства* в семействах алгоритмов получение достаточно точных его оценок едва ли возможно. Поэтому в данном обзоре особое внимание уделяется оценкам, учитывающим эффекты расслоения (параграф 2.3) и сходства (параграф 2.4) по отдельности. Автору не известны работы, в которых они учитывались бы совместно, а также работы, в которых ставилась бы задача получения точных (не верхних, не асимптотических) оценок вероятности переобучения.

В параграфе 2.5 затрагиваются проблемы обоснования скользящего контроля и адекватного выбора функционала обобщающей способности.

Данная глава является целиком обзорной и не содержит новых результатов. Тем не менее, многие концепции излагаются на русском языке впервые.

2.1 Теория Вапника-Червоненкиса

Статистическая теория восстановления зависимостей по эмпирическим данным (VC-теория) предложена В. Н. Вапником и А. Я. Червоненкисом в конце 60-х — начале 70-х годов [9, 10, 11, 8]. В середине 80-х она получила широкую мировую известность [213, 214, 215] и вместе с работами Валианта [212] на многие годы определила генеральное направление развития теории статистического обучения.

Рассмотрим основные предположения и результаты VC-теории, используя некоторые обозначения, введённые в параграфе 1.5 предыдущей главы.

2.1.1 Основные предположения VC-теории

Пусть множество объектов $\mathscr{X}$ является вероятностным пространством с неизвестной вероятностной мерой P. Задано множество (семейство) алгоритмов A и индикатор ошибок $I\colon A \times \mathscr{X} \to \{0,1\}$. Качество произвольного алгоритма $a \in A$ характеризуется *вероятностью ошибки* $P(a) = \mathsf{E}I(a,x)$.

Метод минимизации эмпирического риска. В идеале хотелось бы найти алгоритм a с минимальной вероятностью ошибки $P(a)$. Однако на практике это невозможно, поскольку вероятностная мера P не известна. Вместо минимизации $P(a)$ применяется *минимизация эмпирического риска* — ищется алгоритм a^* с минимальной частотой ошибок на заданной обучающей выборке $X = (x_1, \ldots, x_\ell)$:

$$a^* = \arg\min_{a \in A} \nu(a, X), \qquad \nu(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} I(a, x_i).$$

Предполагается, что X является *простой выборкой*, то есть что это конечное множество объектов, выбранных из множества $\mathscr{X}$ случайно и независимо согласно вероятностной мере P.

Если в семействе существует много алгоритмов, минимизирующих эмпирический риск $\nu(a, X)$, то в качестве решения может быть выбран любой из них. Метод обучения как отображение $\mu\colon \mathscr{X}^\ell \to A$ в VC-теории не рассматривается.

Требование равномерной сходимости. Чтобы вероятность ошибки $P(a^*)$ была близка к минимальной независимо от того, какой именно алгоритм a^* будет выбран, вводится требование *равномерной ограниченности* отклонения частоты ошибок от их вероятности: $\bigl|P(a) - \nu(a, X)\bigr| \leqslant \varepsilon$ для всех алгоритмов $a \in A$. Поскольку выборка X случайная, это условие приходится смягчать, требуя, чтобы оно выполнялось не всегда, а лишь с вероятностью, близкой к единице.

Итак, получаем постановку задачи: найти условия, при которых для достаточно малых значений *точности* ε и *надёжности* η справедлива оценка

$$\bar{P}_\varepsilon(A) = \mathsf{P}\Bigl\{\sup_{a \in A} \bigl|P(a) - \nu(a, X)\bigr| > \varepsilon\Bigr\} \leqslant \eta. \tag{2.1}$$

Если $\bar{P}_\varepsilon(A) \to 0$ при $\ell \to \infty$, то говорят о *равномерной сходимости частоты ошибок к их вероятности*.

Можно также характеризовать качество алгоритма a^* частотой ошибок $\nu(a^*, \bar{X})$ на случайной независимой контрольной выборке $\bar{X}$ длины k, выбранной из того же неизвестного распределения P на множестве $\mathscr{X}$. Тогда оценки получаются более точными в силу «основной леммы», доказанной в [8, стр. 219] при $\ell = k$:

$$\mathsf{P}\Bigl\{\sup_{a \in A} \bigl|\,\mathsf{P}(a) - \nu(a, X)\bigr| > \varepsilon\Bigr\} \leqslant 2\,\mathsf{P}\Bigl\{\max_{a \in A} \bigl|\nu(a, \bar{X}) - \nu(a, X)\bigr| > \tfrac{1}{2}\varepsilon\Bigr\}. \tag{2.2}$$

В [215] это неравенство было уточнено: в правой части $\frac{1}{2}\varepsilon$ заменено на $\varepsilon - \frac{1}{\ell}$.

Если правая часть стремится к нулю при $\ell, k \to \infty$, то говорят о *равномерной сходимости частот в двух выборках*.

Заметим, что в (2.1) и (2.2) можно убирать модули и рассматривать только односторонние оценки, поскольку отклонения $P(a)$ ниже $\nu(a,X)$ являются благоприятными. При этом точность оценки повышается вдвое. Таким образом, вместо (2.1) приходим к требованию *равномерной ограниченности сверху* величины *переобученности* $\delta(a,X,\bar{X}) = \nu(a,\bar{X}) - \nu(a,X)$:

$$P_\varepsilon(A) = \mathsf{P}\Big\{\max_{a\in A} \delta(a,X,\bar{X}) > \varepsilon\Big\} \leqslant \eta. \tag{2.3}$$

О симметризации и комбинаторном подходе. Неравенство (2.2) применяется не только в VC-теории, но и в большинстве современных подходов к оцениванию обобщающей способности. Обычно его рассматривают как вспомогательный технический приём, называемый *симметризацией* (symmetrization). Введение дополнительной *призрачной выборки* (ghost sample) $\bar{X}$ считается «неизбежным злом», так как оно ухудшает оценку, однако более удобного способа оценивать отклонение частоты от вероятности пока не найдено. В комбинаторном подходе, развиваемом в данной работе, призрачная выборка $\bar{X}$ получает естественную интерпретацию — это контрольная выборка, скрытая в момент обучения. Именно на ней и проверяется обобщающая способность алгоритма $a = \mu X$. Удивительно, что данная интерпретация не встречается в работах по теории статистического обучения. Остаётся незамеченной также и очевидная связь симметризации с полным скользящим контролем.

В слабой вероятностной аксиоматике симметризация не нужна, так как оцениваемые функционалы изначально формулируются в терминах конечных выборок. Кроме того, снимается искусственное требование равенства длин обучающей и скрытой выборки, $\ell = k$.

2.1.2 Основные результаты VC-теории

Одним из основных результатов VC-теории является *оценка скорости равномерной сходимости* [10, 11]. При $\ell = k$ для любого распределения вероятностей на $\mathscr{X}$ и любой фиксированной функции потерь $I(a,x)$ справедливо неравенство

$$P_\varepsilon(A) \leqslant \Delta^A(2\ell) \cdot \tfrac{3}{2}\, e^{-\varepsilon^2\ell}, \tag{2.4}$$

где $\Delta^A(L)$ — *функция роста* семейства алгоритмов A. Эта оценка совпадает с комбинаторной оценкой (1.57) из предыдущей главы.

Функция роста не зависит ни от выборки, ни от метода обучения, и является мерой сложности множества алгоритмов A. Справедлива верхняя оценка $\Delta^A(L) \leqslant 2^L$. Однако если семейство A имеет конечную *ёмкость* h, то оценка становится не экспоненциальной, а полиномиальной по длине выборки, $\Delta^A(L) \leqslant \frac{3}{2}\frac{L^h}{h!}$, и правая часть (2.4) стремится к нулю при $\ell \to \infty$. В таком случае говорят, что семейство A обладает свойством *обучаемости* (learnability).

Метод структурной минимизации риска. Оценив зависимость $\eta(\varepsilon)$, легко выразить из неё ε как функцию от ёмкости h, длины обучения ℓ и надёжности η. При $\ell = k$

для любого распределения на множестве $\mathbb{X}$ с вероятностью не менее $1-\eta$ одновременно для всех алгоритмов $a \in A$ справедливо неравенство

$$\nu(a, \bar{X}) < \nu(a, X) + \sqrt{\frac{h}{\ell} \ln\left(\frac{2e\ell}{h}\right) + \frac{4}{9\ell} \ln\frac{1}{\eta}}. \tag{2.5}$$

Первое слагаемое в этой оценке — эмпирический риск, убывающий с ростом ёмкости h. Второе слагаемое возрастает с ростом ёмкости, и его можно рассматривать как *штраф за сложность* (complexity penalty). Сумма в общем случае достигает минимума при некотором h.

Для определения оптимальной сложности модели в VC-теории предлагается метод *структурной минимизации риска*. В семействе A заранее задаётся *структура* вложенных подсемейств возрастающей ёмкости $A_1 \subset A_2 \subset \cdots \subset A_h = A$. Задача обучения решается в каждом из этих подсемейств, всего h раз. Выбирается подсемейство оптимальной ёмкости, для которого достигается минимум правой части (2.5). Тем самым гарантируется заданное качество обучения.

Основной проблемой VC-теории является сильная завышенность сложностных оценок вида (2.4) и (2.5). Как мы видели в параграфе 1.5.6, степень завышенности может достигать нескольких порядков. Более того, завышенность возрастает с ростом ёмкости h. В методе структурной минимизации риска это приводит к выбору подсемейства заниженной ёмкости, то есть к переупрощению алгоритмов, что подтверждается и в экспериментах на модельных данных [154].

Распространённой ошибкой в интерпретации VC-теории является вывод о необходимости ограничивать сложность семейства алгоритмов. Такой вывод был бы справедлив, если бы VC-оценки были достаточно точными.

О скользящем контроле. При практическом применении структурной минимизации риска обычно рекомендуется вместо завышенной теоретической оценки (2.5) применять оценку скользящего контроля $\hat{\mathsf{E}}\nu(a, \bar{X})$. Однако такая замена ставит под сомнение ценность теоретических результатов, поскольку скользящий контроль не опирается на VC-теорию.

С практической точки зрения скользящий контроль имеет ряд существенных недостатков. Во-первых, это ресурсоёмкая процедура. Во-вторых, оценка скользящего контроля имеет большую дисперсию, что может приводить к ошибкам при выборе оптимальной сложности подсемейства h. В-третьих, скользящий контроль удобен для эмпирического оценивания качества метода обучения, но не удобен для конструирования новых оптимальных методов обучения. По этим причинам задача получения точных теоретических оценок не теряет актуальности.

Об оценках равномерного относительного отклонения частот. Чтобы оценить частоту ошибок на контроле $\nu(a, \bar{X})$ по частоте ошибок на обучении $\nu(a, X)$, достаточно потребовать равномерной сходимости не по всему семейству, а только в области низких частот. Построить эту область в явном виде достаточно трудно. Поэтому в VC-теории предлагается другой способ ослабить требование равномерной сходимости. Вводится функционал вероятности большого равномерного относительного

отклонения частот в двух подвыборках. Для него в [8] получена оценка:

$$\mathsf{P}\left\{\sup_{a\in A}\frac{\nu(a,\bar{X})-\nu(a,X)}{\sqrt{\nu(a,\mathbb{X})}}>\varepsilon\right\}\leqslant\Delta^A(2\ell)\max_{m\geqslant(\varepsilon k)^2/L}H_L^{\ell,m}\left(\tfrac{l}{L}\big(m-\varepsilon k\sqrt{m/L}\big)\right)\leqslant$$
$$\leqslant\Delta^A(2\ell)\cdot 4e^{-\frac{1}{4}\varepsilon^2\ell}.$$

Это оценка заметно точнее оценки (2.4) в области малых частот. Она выводится аналогично доказательству Теорем 1.11 и 1.12, если в ходе доказательства сделать замену переменной $\varepsilon\sqrt{\frac{m}{L}}\to\varepsilon$. По сути дела, относительная оценка отличается от абсолютной лишь тем, что по-другому оценивает сверху комбинаторный множитель. Однако она не описывает эффект локализации семейства алгоритмов и не решает проблему завышенности.

В слабой аксиоматике нет необходимости рассматривать относительные оценки, поскольку они вводятся, чтобы компенсировать неточность верхней оценки комбинаторного множителя в области низких частот. Если же комбинаторный множитель не оценивается сверху, то и компенсация не нужна.

О необходимых и достаточных условиях равномерной сходимости. В VC-теории много внимания уделяется *необходимым и достаточным условиям равномерной сходимости*. Однако равномерная сходимость, как следует из неравенства (1.49), является лишь достаточным условием обучаемости. Вообще, любые завышенные оценки дают лишь достаточные условия и могут оказаться не способными объяснить обучаемость тех или иных методов.

Если ёмкость бесконечна и равномерной сходимости нет, то рано делать вывод, что нет обучаемости. В середине 90-х годов на примере алгоритмов *бустинга* было показано, сначала экспериментально, а затем и теоретически, что обобщающая способность алгоритмических композиций может улучшаться при практически неограниченном росте их сложности [133].

Таким образом, получение необходимых условий равномерной сходимости можно считать чисто теоретическим упражнением, имеющим лишь косвенное отношение к проблеме обучаемости.

2.1.3 Бритва Оккама

Философский принцип, приписываемый английскому философу Уильяму Оккаму, «не плодить сущности без необходимости» имеет прямое отношение к основному результату VC-теории — не пользоваться слишком сложными семействами алгоритмов. В теории статистического обучения *бритвой Оккама* (Occam Razor) принято называть следующее обобщение VC-оценки [164, 165].

Метод обучения μ индуцирует на множестве алгоритмов A распределение вероятностей $q(a)=\mathsf{P}(\mu X{=}a)$. Найти его в явном виде весьма затруднительно, но можно попытаться приблизить его некоторым другим распределением $p(a)$. Оказывается,

для любых A и $p(a)$ с вероятностью $1-\eta$ одновременно для всех $a \in A$

$$P(a) \leqslant \nu(a,X) + \sqrt{\frac{1}{\ell}\ln\frac{1}{p(a)} + \frac{1}{\ell}\ln\frac{1}{\eta}}. \tag{2.6}$$

Эта оценка переходит в (2.5), если взять равномерное распределение $p(a) = \frac{1}{\Delta^A(L)}$. Роль *штрафа за сложность* в данном случае играет величина $\frac{1}{p(a)}$.

Чем точнее $p(a)$ приближает $q(a)$, тем меньше правая часть оценки (2.6). Можно доказать, что математическое ожидание правой части (2.6) по выборке X минимально, когда $p(a) = q(a)$, то есть когда мы точно «угадали» распределение (здесь можно говорить о математическом ожидании, поскольку $a = \mu X$ — функция выборки, следовательно $p(a)$ является случайной величиной).

На самом деле (2.6) является ослабленным следствием более точной оценки, которая, собственно, и называется «бритвой Оккама».

Теорема 2.1 (Бритва Оккама). *Для любых A и $p(a)$ с вероятностью не менее $1-\eta$ одновременно для всех $a \in A$*

$$P(a) \leqslant \operatorname{Bin}_\ell^{-1}\big(\nu(a,X), \eta p(a)\big), \tag{2.7}$$

где $\operatorname{Bin}_\ell^{-1}\big(\frac{s}{\ell},\eta\big) = \max\limits_{p\in[0,1]}\big\{p\colon \operatorname{Bin}_\ell(s,p) = \eta\big\}$ *— функция, обратная по параметру p к функции биномиального распределения*

$$\operatorname{Bin}_\ell(s,p) = \sum_{t=0}^{s} \operatorname{bin}_\ell(t,p), \quad \text{где} \quad \operatorname{bin}_\ell(t,p) = C_\ell^t p^t (1-p)^{\ell-t}.$$

Доказательство. Возьмём произвольное $\delta \in (0,1)$. Заметим, что, согласно определению функции $\operatorname{Bin}_\ell^{-1}$, величина $\bar{P}(a,\delta) = \operatorname{Bin}_\ell^{-1}\big(\nu(a,X),\delta\big)$ является верхней границей доверительного полуинтервала для неизвестного значения вероятности ошибки $P(a)$, оцениваемого по частоте ошибок $\nu(a,X)$:

$$\mathsf{P}\big\{P(a) > \bar{P}(a,\delta)\big\} \leqslant \delta.$$

Положим $\delta = \eta p(a)$ и воспользуемся *неравенством Буля* (union bound):

$$\mathsf{P}\big\{\forall a \in A \;\; P(a) > \bar{P}(a,\eta p(a))\big\} \leqslant \sum_{a\in A} \mathsf{P}\big\{P(a) > \bar{P}(a,\eta p(a))\big\} \leqslant \sum_{a\in A} \eta p(a) \leqslant \eta,$$

что и требовалось доказать. ■

Оценка (2.7) устраняет ту часть завышенности (2.6), которая возникает из-за применения неравенства Хёфдинга — экспоненциальной аппроксимации функции биномиального распределения.

Оценка (2.7) учитывает также расслоение семейства алгоритмов, но косвенным образом — путём априорного «угадывания» функции $q(a)$. Чем точнее приближение $p(a) \approx q(a)$, тем менее завышенной будет оценка.

Второй причиной завышенности является использование неравенства Буля при доказательстве теоремы. Чем более неравномерно (сконцентрировано) распределение $q(a)$, тем менее завышенной будет оценка. Однако сконцентрированное распределение труднее «угадать» априори, и в этом заключается принципиальная сложность применения оккамовского подхода.

Ещё одна сложность заключается в том, что функцию $p(a)$ необходимо задавать до того, как станет известна выборка X, поскольку в доказательстве предполагается, что $p(a)$ не зависит от X. Таким образом, при «угадывании» истинного распределения $q(a)$ мы не имеем права использовать выборку, и вынуждены полагаться лишь на априорную информацию о выборке X и методе μ.

2.2 Оценки, зависящие от задачи

Появление VC-теории вызвало большое количество исследований, направленных на уточнение оценок. Однако проблема получения численно точных оценок, непосредственно применимых на практике, оказалась вызывающе трудной, и до сих пор остаётся открытой.

Все известные оценки опираются на те или иные неравенства *концентрации вероятностной меры* (measure concentration). В первых работах Вапника и Червоненкиса использовались классические неравенства Хёфдинга и Бернштейна. Более точные результаты удаётся получать с помощью неравенств Чернова [119], метода ограниченных разностей МакДиармида [180] и изопериметрических неравенств Талагранда [209, 210]. Вводное изложение этих математических техник можно найти в обзорах [90, 172].

Здесь приведём только *неравенство ограниченных разностей* (bounded differences inequality) МакДиармида, которое является обобщением классической теоремы о сходимости среднего к математическому ожиданию.

Теорема 2.2. *Пусть $g\colon \mathscr{X}^\ell \to \mathbb{R}$ — функция ℓ переменных такая, что существуют неотрицательные константы $c_1, \ldots, c_\ell$, для которых*

$$\sup_{x_1,\ldots,x_\ell,x'\in\mathscr{X}} \bigl|g(x_1,\ldots,x_\ell) - g(x_1,\ldots,x_{i-1},x',x_{i+1},\ldots,x_\ell)\bigr| \leqslant c_i, \quad i=1,\ldots,\ell.$$

Тогда для любого распределения на $\mathscr{X}$ и любой случайной независимой выборки $X=\{x_1,\ldots,x_\ell\}$ из этого распределения случайная переменная $g(X)=g(x_1,\ldots,x_\ell)$ удовлетворяет неравенству

$$\mathsf{P}\bigl\{|g(X)-\mathsf{E}g|>\varepsilon\bigr\} \leqslant 2\exp\bigl(-\tfrac{2}{C}\varepsilon^2\bigr), \qquad C=\sum_{i=1}^{\ell} c_i^2. \tag{2.8}$$

В частности, если положить $g(X)=\nu(a,X)$ при некотором фиксированном a, то $c_i=\frac{1}{\ell}$, $\mathsf{E}g(X)=P(a)$, и из теоремы следует неравенство Бернштейна, описывающее скорость сходимости частоты ошибок к их вероятности:

$$\mathsf{P}\bigl\{|\nu(a,X)-P(a)|>\varepsilon\bigr\} \leqslant 2\exp\bigl(-2\ell\varepsilon^2\bigr).$$

2.2.1 Локальные меры сложности

При решении конкретной задачи далеко не каждый алгоритм из выбранного семейства имеет шансы быть полученным в результате обучения. Как правило, реально задействуется не всё семейство, а лишь небольшая его часть.

Этот факт был замечен ещё В. Н. Вапником, предложившим понятие эффективной ёмкости вместе с алгоритмом её практического измерения [216, 107], см. также стр. 115. *Эффективная ёмкость* не превосходит полной ёмкости семейства и зависит от выборки. Она учитывает особенности исходного распределения объектов, но не принимает во внимание особенностей целевой зависимости и метода обучения. Никакого метода для теоретического оценивания эффективной ёмкости так и не было предложено, и про неё довольно быстро забыли.

В дальнейшем концепция *оценок, зависящих от задачи* (data dependent bounds), получила развитие во многих работах, см., например, [199, 223, 109, 110, 93]. Однако надо принимать во внимание, что когда говорят «оценка зависит от данных», подразумевается, что учтены *некоторые* свойства данных, но не обязательно все, и не обязательно наилучшим образом.

Статья [217] содержит обзор ранних VC-оценок, зависящих от задачи. Отмечается, что наилучшая оценка, справедливая при достаточно общих предположениях, получена М. Талаграндом [209], и на её основе выводится более точная оценка, справедливая при некотором «разумном» ограничении класса вероятностных распределений на множестве объектов $\mathscr{X}$.

Существует много подходов, учитывающих те или иные особенности выборки или метода обучения путём введения локальных мер сложности. Многие из этих мер оцениваются друг через друга, см. обзоры [208, 185, 108]. Это позволяет брать ту меру сложности, которая более удобна для оценивания в конкретном случае. Локальные меры сложности дают более точные оценки вероятности ошибки, чем VC-теория, но и они в большинстве случаев завышены на порядки.

Вещественные функции потерь вида $\mathscr{L}\colon A \times \mathscr{X} \to \mathbb{R}$ используются как в методах восстановления регрессии, так и во многих методах классификации. VC-теория несложно обобщается на этот случай [8]. Вместо вероятности ошибки $P(a)$ вводится *ожидаемая потеря*:

$$\tilde{P}(a) = \mathsf{E}\mathscr{L}(a, x);$$

а вместо частоты ошибок $\nu(a, X)$ — *средняя потеря на выборке* X:

$$\tilde{\nu}(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathscr{L}(a, x_i).$$

Понятия *коэффициента разнообразия*, *функции роста* и *ёмкости* распространяются на случай вещественной функции потерь путём введения индикатора ошибки $I_\delta(a, x) = \big[\mathscr{L}(a, x) \geqslant \delta\big]$ с дополнительным скалярным параметром δ. Тогда в качестве коэффициента разнообразия можно взять мощность множества векторов ошибок, индуцируемых всевозможными $a \in A$ при всевозможных значениях параметра δ.

Дальнейшее обобщение VC-оценок на задачи восстановления регрессии и интерпретации результатов косвенных экспериментов подробно рассматривается в [8].

Fat-разнообразие и fat-размерность — это ещё одно обобщение коэффициентов разнообразия и VC-размерности на вещественный случай, предложенное М. Кернсом и Р. Шапиром в [156].

Пусть множество алгоритмов $A = \{f_w \colon \mathscr{X} \to \mathbb{R}\}$ — это параметрическое семейство вещественных функций с параметром w.

Для заданного вектора $s = (s_1, \ldots, s_\ell) \in \mathbb{R}^\ell$ обозначим через $\delta_\theta(A, X, s)$ число бинарных векторов $b \in \{0,1\}^\ell$ таких, что найдётся функция $f_w \in A$, для которой

$$f_w(x_i) \begin{cases} \geqslant s_i + \theta, & b_i = 1, \\ \leqslant s_i - \theta, & b_i = 0, \end{cases} \qquad i = 1, \ldots, \ell.$$

Геометрически $\delta_\theta(A, X, s)$ — это число вершин ℓ-мерного куба с центром в s и длиной ребра 2θ, «заметаемых» всевозможными алгоритмами из A. Максимальное (по всевозможным положениям центра s) число таких бинарных векторов называется *fat-разнообразием* (fat-shattering) множества A по выборке X:

$$\Delta_\theta(A, X) = \max_{s \in \mathbb{R}^\ell} \delta_\theta(A, X, s).$$

Очевидно, fat-разнообразие не возрастает по θ и не превосходит 2^ℓ. Максимальная длина выборки, при которой оно равно в точности 2^ℓ, называется *fat-размерностью* (fat-shattering dimension) множества A:

$$\mathrm{fat}_\theta(A) = \max\bigl\{\ell \colon \exists X \in \mathscr{X} \colon |X| = \ell, \;\; \Delta_\theta(A, X) = 2^\ell\bigr\}.$$

Геометрически fat-размерность — это максимальная размерность куба с ребром 2θ, целиком «заметаемого» всевозможными алгоритмами из A.

Известно большое количество верхних оценок вероятности ошибки, в которых *штраф за сложность* выражается через fat-размерность. Они отличаются константами, условиями применимости и прочими техническими деталями [102, 97, 200, 199, 201, 202, 91, 93]. Обычно fat-размерность входит в эти оценки аналогично VC-размерности, но с дополнительным множителем порядка $\ln^2 \ell$. В параграфе 2.2.2 будет приведена одна из типичных оценок.

Мощность покрытия и мощность упаковки. Пусть $\rho(a, a')$ — произвольная метрика на множестве A.

Множество $a_1, \ldots, a_\mathcal{N}$ минимальной мощности такое, что для любого $a \in A$ найдётся a_n на расстоянии $\rho(a, a_n) \leqslant \varepsilon$, называется *$\varepsilon$-сетью* ($\varepsilon$-covering). Мощность ε-сети $\mathcal{N}(\varepsilon, A)$ будем называть также *мощностью ε-покрытия* (ε-covering number).

Множество $a_1, \ldots, a_\mathcal{M}$ максимальной мощности такое, что любые два его элемента $a_m, a_{m'} \in A$ находятся друг от друга на расстоянии $\rho(a_m, a_{m'}) \geqslant \varepsilon$, называется *$\varepsilon$-упаковкой* ($\varepsilon$-packing). Мощность этого множества $\mathcal{M}(\varepsilon, A)$ называется *мощностью ε-упаковки* (ε-packing number).

Мощности ε-покрытия и ε-упаковки связаны двусторонними неравенствами [185]:

$$\mathcal{N}(\varepsilon, A) \leqslant \mathcal{M}(\varepsilon, A) \leqslant \mathcal{N}(\varepsilon/2, A).$$

Если метрика ρ порождается выборкой X и функцией потерь $\mathscr{L}$,

$$\rho^p(a, a') = \sum_{i=1}^{\ell} \big|\mathscr{L}(a, x_i) - \mathscr{L}(a', x_i)\big|^p,$$

то говорят о *локальных* мощностях покрытия и упаковки, и обозначают их, соответственно, $\mathcal{N}(\varepsilon, A, X)$ и $\mathcal{M}(\varepsilon, A, X)$. В оценках, не зависящих от задачи, используются их глобальные (говорят также *равномерные*) аналоги:

$$\mathcal{N}(\varepsilon, A, \ell) = \max_{X \in \mathscr{X}^\ell} \mathcal{N}(\varepsilon, A, X), \qquad \mathcal{M}(\varepsilon, A, \ell) = \max_{X \in \mathscr{X}^\ell} \mathcal{M}(\varepsilon, A, X).$$

Известно большое количество верхних оценок вероятности ошибки, в которых *штраф за сложность* выражается через мощность покрытия. Как правило, она входит в эти оценки аналогично коэффициенту разнообразия [208, 185]. В параграфе 2.2.2 будет приведена одна из таких оценок.

Радемахеровская сложность. Понятие *радемахеровской сложности* (Rademacher complexity) было введено в теорию статистического обучения В. Кольчинским [159, 158, 160]. Оно даёт наиболее точные оценки обобщающей способности, отличается математическом изяществом и удобством использования в приложениях [98, 103, 99].

Дискретная случайная величина σ, принимающая два значения $-1, +1$, каждое с вероятностью $\frac{1}{2}$, называется *радемахеровской*.

Локальной радемахеровской сложностью (local Rademacher complexity) семейства A на выборке X называется величина

$$\mathcal{R}(A, X) = \mathsf{E}_\sigma \sup_{a \in A} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} \sigma_i \mathscr{L}(a, x_i) \right|, \tag{2.9}$$

где $\sigma = (\sigma_1, \ldots, \sigma_\ell)$ — независимые радемахеровские случайные величины.

Смысл этой величины легко понять, если заметить, что $r(a, \sigma) = \frac{1}{\ell} \sum_{i=1}^{\ell} \sigma_i \mathscr{L}(a, x_i)$ есть выборочная ковариация величины потерь и случайного (радемахеровского) шума. Супремум $\sup_{a \in A} \big|r(a, \sigma)\big|$ выделяет в семействе A алгоритм, поведение которого на выборке X наиболее близко к шуму. Математическое ожидание E_σ по вектору шума необходимо для того, чтобы величина сложности $\mathcal{R}$ не зависела от выбора конкретного вектора шума. Степень близости вектора потерь и вектора шума, выраженная с помощью ковариации, вполне соответствует интуитивным представлениям о сложности: чем сложнее семейство, тем выше шансы найти в нём алгоритм, «похожий» на произвольный шум. Нетрудно также понять, почему в качестве меры близости двух векторов берётся именно ковариация, а не функция расстояния —

благодаря линейности ковариации рademахеровская сложность обладает удобными алгебраическими свойствами, некоторые из которых будут приведены ниже.

Математическое ожидание $\mathcal{R}^A(\ell) = \mathsf{E}_X \mathcal{R}(A, X)$ по выборке X фиксированной длины ℓ называется (глобальной) *рademахеровской сложностью* семейства A.

Локальная рademахеровская сложность $\mathcal{R}(A, X)$ является аналогом *коэффициента разнообразия* $\Delta(A, X)$, а глобальная $\mathcal{R}^A(\ell)$ — аналогом *функции роста* $\Delta^A(\ell)$. Более того, эти величины связаны неравенствами [108]

$$\mathcal{R}(A, X) \leqslant \sqrt{\frac{2 \ln \Delta(A, X)}{\ell}}, \qquad \mathcal{R}^A(\ell) \leqslant \sqrt{\frac{2 \ln \Delta^A(\ell)}{\ell}}. \tag{2.10}$$

Переходя к рассмотрению свойств рademахеровской сложности, заметим, что она может определяться (при фиксированной функции потерь $\mathscr{L}$) как для семейства A, так и непосредственно для множества векторов $\vec{A} = \big\{\vec{a} = \big(\mathscr{L}(a, x_i)\big)_{i=1}^{\ell} \colon a \in A\big\}$. Очевидно, что $\mathcal{R}(A, X) = \mathcal{R}(\vec{A}, X)$, хотя в общем случае $|\vec{A}| \leqslant |A|$, и даже может оказаться так, что множество A бесконечно, а множество $\vec{A}$ — конечно (как в случае с коэффициентами разнообразия при бинарной функции потерь).

1. Пусть $\vec{A}$, $\vec{B}$ — ограниченные подмножества в $\mathbb{R}^\ell$, и c — константа. Тогда

$$\begin{aligned}
&\mathcal{R}(A \cup B, X) \leqslant \mathcal{R}(A, X) + \mathcal{R}(B, X);\\
&\mathcal{R}(c \cdot A, X) = |c| \cdot \mathcal{R}(A, X), \qquad\qquad c \cdot A = \{c\vec{a} \colon a \in A\};\\
&\mathcal{R}(A \oplus B, X) \leqslant \mathcal{R}(A, X) + \mathcal{R}(B, X), \qquad A \oplus B = \{\vec{a} + \vec{b} \colon a \in A,\ b \in B\}.
\end{aligned}$$

2. Пусть A — конечное множество, $\|a\|^2 = \sum\limits_{i=1}^{\ell} \mathscr{L}^2(a, x_i)$ — евклидова норма вектора потерь. Тогда

$$\mathcal{R}(A, X) = \frac{\sqrt{2 \ln |A|}}{\ell} \max_{a \in A} \|a\|.$$

3. Пусть $\operatorname{conv} A = \Big\{ \sum\limits_{a \in A} c_a \vec{a} \colon \sum\limits_{a \in A} |c_a| \leqslant 1,\ a \in A \Big\}$ — выпуклая оболочка множества векторов A. Тогда

$$\mathcal{R}(\operatorname{conv} A, X) = \mathcal{R}(A, X).$$

4. Пусть функция $\varphi \colon \mathbb{R} \to \mathbb{R}$ липшицева с константой L_φ, и $\varphi(0) = 0$. Обозначим через $\varphi \circ A$ множество векторов вида $\big(\varphi(a_1), \ldots, \varphi(a_\ell)\big)$. Тогда φ является сжимающим отображением в следующем смысле:

$$\mathcal{R}(\varphi \circ A, X) \leqslant L_\varphi \mathcal{R}(A, X).$$

Более полный обзор свойств рademахеровской сложности можно найти в [108, 99].

В верхних оценках ожидаемых потерь величина $\mathcal{R}(A, X)$ непосредственно играет роль *штрафа за сложность* [160, 108]:

Теорема 2.3. *Пусть* $\mathscr{L}\colon A \times \mathscr{X} \to [-1,1]$ *— ограниченная функция потерь. Тогда с вероятностью не менее* $1-\eta$ *одновременно для всех* $a \in A$

$$\tilde{P}(a) \leqslant \tilde{\nu}(a,X) + 2\mathcal{R}(A) + \sqrt{\frac{2}{\ell}\ln\frac{1}{\eta}}\,.$$

Справедлива также локальная оценка

$$\tilde{P}(a) \leqslant \tilde{\nu}(a,X) + 2\mathcal{R}(A,X) + \sqrt{\frac{2}{\ell}\ln\frac{2}{\eta}}\,. \tag{2.11}$$

Чтобы прояснить суть радемахеровской сложности, рассмотрим основные шаги доказательства этой теоремы.

1. Вводится случайная величина $g(X) = \sup\limits_{a\in A}\bigl|\tilde{P}(a) - \tilde{\nu}(a,X)\bigr|$. К ней применяется неравенство МакДиармида (2.8), которое записывается в виде оценки сверху, справедливой с вероятностью не менее $1-\eta$:

$$\sup_{a\in A}\bigl|\tilde{P}(a) - \tilde{\nu}(a,X)\bigr| \leqslant \mathsf{E}\sup_{a\in A}\bigl|\tilde{P}(a) - \tilde{\nu}(a,X)\bigr| + \sqrt{\frac{2}{\ell}\ln\frac{1}{\eta}}.$$

2. Вводится *призрачная выборка* (ghost sample) $X' = (x'_1,\ldots,x'_\ell)$ — случайная, независимая, из того же распределения и той же длины, что X. С помощью неравенства Йенсена выводится *симметризованная* оценка для математических ожиданий, аналогичная «основной лемме» Вапника:

$$\mathsf{E}\sup_{a\in A}\bigl|\tilde{P}(a) - \tilde{\nu}(a,X)\bigr| \leqslant \mathsf{E}\sup_{a\in A}\bigl|\tilde{\nu}(a,X') - \tilde{\nu}(a,X)\bigr|.$$

3. Вводятся независимые радемахеровские случайные величины $\sigma_1,\ldots,\sigma_\ell$, с помощью которых, собственно и определяется величина $\mathcal{R}(A,X)$:

$$\begin{aligned}
\mathsf{E}\sup_{a\in A}\bigl|\tilde{\nu}(a,X') - \tilde{\nu}(a,X)\bigr| &= \mathsf{E}\sup_{a\in A}\frac{1}{\ell}\left|\sum_{i=1}^{\ell}\mathscr{L}(a,x_i) - \mathscr{L}(a,x'_i)\right| = \\
&= \mathsf{E}\sup_{a\in A}\frac{1}{\ell}\left|\sum_{i=1}^{\ell}\sigma_i\bigl(\mathscr{L}(a,x_i) - \mathscr{L}(a,x'_i)\bigr)\right| \leqslant \\
&\leqslant 2\,\mathsf{E}\sup_{a\in A}\frac{1}{\ell}\left|\sum_{i=1}^{\ell}\sigma_i\mathscr{L}(a,x_i)\right| = 2\mathcal{R}(A,X).
\end{aligned}$$

4. Для получения локальной оценки (2.11) неравенство МакДиармида применяется к случайной переменной $\mathcal{R}(A,X)$. ■

В этом доказательстве три раза сделаны оценки сверху, однако факторы их завышенности совсем невелики, по сравнению с VC-оценками. Здесь вообще не используется сильно завышенное неравенство Буля. Поэтому $\mathcal{R}(A,X)$ является наиболее точной мерой сложности из всех рассмотренных выше.

Подставляя в (2.11) верхнюю оценку радемахеровской сложности через коэффициент разнообразия (2.10), можно получить аналог сильно завышенной VC-оценки (2.5).

Вторая оценка в Теореме 2.3 несколько хуже, но она использует *локальную* радемахеровскую сложность, которую можно вычислять по обучающей выборке X. Вычисление супремума в (2.9) — это оптимизационная задача, мало чем отличающаяся от задачи минимизации эмпирического риска. Математическое ожидание E_σ можно приближённо вычислить, усредняя по некоторому множеству реализаций случайного вектора $(\sigma_1, \ldots, \sigma_\ell)$, то есть фактически *методом Монте-Карло*. Эта вычислительная процедура мало чем отличается от обычного скользящего контроля. Однако она даёт завышенную оценку $P(a)$, тогда как скользящий контроль — несмещённую. Таким образом, эмпирическое оценивание радемахеровской сложности практически лишено смысла. Следует искать пути её аналитического оценивания.

Выводы и интерпретации.

1. Локальные меры сложности зависят от выборки X и позволяют учитывать индивидуальные особенности конкретной задачи. В этом, несомненно, заключается их достоинство перед классической функцией роста в VC-теории. Недостатки же связаны с тем, что локальные меры сложности слабо учитывают восстанавливаемую зависимость $y(x)$ и метод обучения μ.

2. Наиболее точной является локальная радемахеровская сложность $\mathcal{R}$, но и она является оценкой худшего случая, поскольку определяется по алгоритмам, максимально неудачным для данной задачи, тогда как интерес представляют алгоритмы, выбираемые методом обучения.

3. Меры сложности VC, fat, $\mathcal{N}$ и $\mathcal{M}$, включая их локальные варианты, основаны на подсчёте числа алгоритмов в семействе A. Это означает, что неявно за ними стоит неравенство Буля, которое сильно завышено, когда в семействе A есть много похожих алгоритмов. Меры сложности $\mathcal{N}$ и $\mathcal{M}$ позволяют учитывать сходство путём увеличения параметра ε, но при этом снижается точность аппроксимации произвольного алгоритма a ближайшим элементом ε-покрытия или ε-упаковки, так что увеличение параметра ε улучшает точность оценки лишь до некоторого предела.

4. Все рассмотренные локальные меры сложности основаны на принципе *равномерной сходимости*. Значит, оценки вероятности переобучения, получаемые с их помощью, будут завышены, как минимум, настолько же, насколько завышена оценка

$$\mathsf{P}\Big\{P(\mu X) - \nu(\mu X, X) \geqslant \varepsilon\Big\} \leqslant \mathsf{P}\Big\{\sup_{a \in A}\big(P(a) - \nu(a, X)\big) \geqslant \varepsilon\Big\}.$$

Итак, известные сложностные оценки, даже если они опираются на *локальные* меры сложности, не учитывают в полной мере эффекты расслоения и сходства в семействе алгоритмов, а также особенности конкретного метода обучения μ.

2.2.2 Оценки, учитывающие отступы объектов

Рассмотрим задачу классификации на два класса $\mathbb{Y} = \{-1, +1\}$ с обучающей выборкой $X = (x_i, y_i)_{i=1}^{\ell}$, где $y_i = y(x_i)$. Пусть *алгоритм классификации* имеет вид

$a(x,w) = \operatorname{sign} f(x,w)$, где w — вектор параметров, f — фиксированная функция, называемая *дискриминантной*. В частности, в *линейном классификаторе* $f(x,w) = \langle x,w\rangle$ есть скалярное произведение векторов x и w из $\mathbb{R}^n$. Координаты векторов x и w соответствуют n признакам.

Множество $\{x\colon f(x,w)=0\}$ называется *разделяющей поверхностью*.

Величина $M_i(w) = y_i f(x_i,w)$ называется *отступом* (margin) объекта x_i.

Отступ характеризует расположение объекта x_i относительно разделяющей поверхности. Если алгоритм неверно классифицирует объект, то отступ отрицателен.

Наряду с частотой ошибок $\nu(a,X)$ будем рассматривать частоту *ошибок с отступом* (margin error), зависящую от неотрицательного параметра θ:

$$\nu_\theta(a,X) = \frac{1}{\ell}\sum_{i=1}^{\ell}\bigl[M_i(w) \leqslant \theta\bigr].$$

Это частота ошибок при более строгой бинарной функции потерь, принимающей значение 1 не только в случае обычной ошибки, но и когда объект приближается к разделяющей поверхности ближе, чем на θ. Очевидно, $\nu(a,X) \leqslant \nu_\theta(a,X)$.

Оценка VC-размерности. Достаточно давно была выдвинута гипотеза, что чем больше отступы объектов обучающей выборки, тем лучше (увереннее) разделяются классы, и тем меньше должна быть вероятность ошибки.

Первое математическое обоснование этим соображениям дали Вапник и Червоненкис [11, 8], показав, что ёмкость линейного классификатора ограничивается сверху не только размерностью пространства n, но и величиной $\left(\frac{D}{r}\right)^2 + 1$, где $D = \max\|x_i\|$ — половинный диаметр выборки, r — минимальное расстояние между выпуклыми оболочками векторов выборки, отнесённых к разным классам. Использование этой оценки в методе структурной минимизации риска послужило обоснованием для метода обобщённого портрета [11, 8], позже получившему широкую известность как *машина опорных векторов* (support vector machine, SVM).

Непосредственное использование этой оценки весьма затруднительно, так как она является *ненаблюдаемой* — здесь под «выборкой» понимается полная выборка $\mathbb{X}$, включающая обучение и контроль. Позже были получены наблюдаемые оценки достаточно общего вида, основанные на fat-размерности, мощности покрытия или радемахеровской сложности. В случае линейного классификатора все они оценивают *штраф за сложность* через геометрическую величину вида $\left(\frac{D}{r}\right)^2$, возможно, с несколько иной интерпретацией параметров D и r.

Оценка fat-размерности приводится здесь согласно [100, 208]. Для любых A и $\theta > 0$ с вероятностью $1-\eta$ одновременно для всех $a \in A$

$$P(a) \leqslant \nu_\theta(a,X) + C\sqrt{\frac{1}{\ell}\operatorname{fat}_\theta(A)\,\ln^2\frac{\ell}{\theta} + \frac{1}{\ell}\ln\frac{1}{\eta}}, \tag{2.12}$$

где C — некоторая константа. Для линейного классификатора $\operatorname{fat}_\theta(A) \leqslant \left(\frac{D}{\theta}\right)^2$, где D — половинный диаметр выборки.

В этой оценке первое слагаемое не убывает по θ, второе — не возрастает. Величина эмпирического риска искусственно завышается, зато локальная (зависящая от задачи) сложность семейства становится существенно меньше, если её оценивать только по объектам, далеко отстоящим от разделяющей поверхности. Чтобы получить максимально точную оценку, необходимо выбрать значение θ, при котором правая часть (2.12) минимальна.

Оценка мощности покрытия приводится здесь согласно [201, 97, 208]. Для любых A и $\theta > 0$ с вероятностью $1 - \eta$ одновременно для всех $a \in A$

$$P(a) \leqslant \nu_\theta(a, X) + \sqrt{\frac{8}{\ell} \ln \mathcal{N}(\theta/2, A, 2\ell) + \frac{8}{\ell} \ln \frac{2}{\eta}}.$$

В этой оценке, как и в предыдущей, первое слагаемое не убывает по θ, второе не возрастает, и возможно распорядиться свободой выбора параметра θ, чтобы минимизировать правую часть неравенства.

Оценка рademахеровской сложности оказывается более плодотворной и обосновывает не только введение штрафа за сложность, но и применение разнообразных *вещественных функций потерь* в задачах классификации.

Рассмотрим в качестве функции $\mathscr{L}(a, x)$ непрерывную или гладкую верхнюю оценку пороговой функции потерь:

$$I(a, x_i) = \big[a(x_i, w) \neq y_i\big] = \big[M_i(w) < 0\big] \;\leqslant\; \mathscr{L}(a, x_i) \equiv \lambda\big(M_i(w)\big),$$

где $\lambda(M)$ — невозрастающая функция отступа (чем больше отступ, тем меньше потеря). В теории классификации известно много методов обучения, основанных на минимизации средней потери $\tilde{\nu}(a, X) \to \min\limits_{a \in A}$ при том или ином выборе функции $\lambda(M)$. С вычислительной точки зрения это очень выгодный шаг, поскольку NP-полная задача поиска максимальной совместной подсистемы [55] в системе неравенств $M_i(w) \geqslant 0$, $i = 1, \ldots, \ell$ заменяется задачей минимизации непрерывного или даже гладкого функционала, для решения которой существуют эффективные численные методы. При этом исходный функционал частоты ошибок подменяется другим, и возникает опасение, что качество классификации от этого ухудшится. Однако практика показывает, что минимизация средней потери (или максимизация отступов), наоборот, улучшает обобщающую способность. Этот парадокс требует как должного теоретического обоснования, так и ответа на важный практический вопрос — при какой функции потерь $\lambda(M)$ достигается наилучшая обобщающая способность.

Перечислим некоторые функции потерь, соответствующие широко известным методам классификации, см. также рис. 2.1.

$\lambda(M) = \log_2(1 + e^{-M})$ — логарифмическая функция потерь, используемая в *логистической регрессии*; эта функция является следствием принципа максимума правдоподобия и формулы байесовского оптимального классификатора при предположении, что плотности распределения классов являются экспоненциальными, с равными значениями параметров разброса [1, 145];

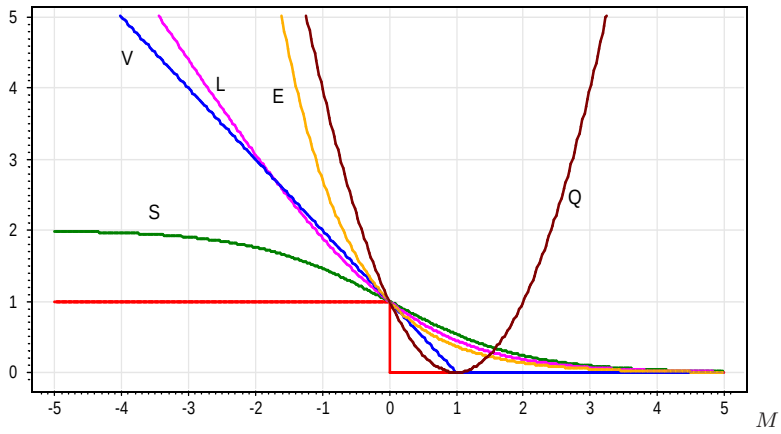


Рис. 2.1. Некоторые непрерывные аппроксимации пороговой функции потерь: Q — квадратичная, V — кусочно-линейная (SVM); S — сигмоидная; L — логарифмическая; E — экспоненциальная.

$\lambda(M) = (1-M)_+$ — кусочно-линейная функция потерь, используемая в *методе опорных векторов*; эта функция появляется вследствие обобщения принципа максимума ширины зазора между классами на случай, когда классы линейно неразделимы [123, 118, 208];

$\lambda(M) = (-M)_+$ — кусочно-линейная функция потерь, используемая в *правиле Хэбба* для обучения однослойного персептрона Розенблатта; эта функция имеет некоторые обоснования в нейрофизиологи и использовалась в ранних работах по нейронным сетям [141, 192, 69];

$\lambda(M) = e^{-M}$ — экспоненциальная функция потерь, используемая в алгоритме *бустинга* AdaBoost [133];

$\lambda(M) = (1-M)^2$ — квадратичная функция потерь; соответствует «наивному» сведению задачи классификации к задаче регрессии и применению метода наименьших квадратов; в случае предварительного центрирования объектов (перед решением задачи наименьших квадратов из каждого вектора x_i вычитается вектор среднего арифметического всех векторов класса y_i) соответствует *линейному дискриминанту Фишера* [40];

$\lambda(M) = \frac{2}{1+e^M}$ — сигмоидная функция потерь; часто применяется при настройке нейронных сетей как наиболее близкий гладкий аналог пороговой функции потерь [70, 34, 84];

$\lambda(M) = \left(\frac{2}{1+e^M}\right)^2$ — квадрат сигмоидной функции потерь; также применяется при настройке нейронных сетей; возникает при использовании сигмоидной функции активации выходного нейрона совместно с квадратичной функцией потерь;

$\lambda(M) = \Phi(-M)$ — гауссовская функция потерь, где $\Phi(z)$ — функция нормального распределения; эта функция находит обоснование в рамках *PAC-байесовского подхода* [164, 165], который будет рассмотрен ниже, см. стр. 97;

$\lambda(M) = [M \leqslant \theta],\ \theta > 0$ — пороговая функция *ошибки с отступом* (margin error), штрафующая не только за ошибки, но и за приближение к разделяющей поверхности ближе чем на θ;

$\lambda(M) = \max\{(1-M/\theta)_+, 1\}$ — непрерывный кусочно-линейный аналог пороговой функции ошибки с отступом.

Следующая теорема утверждает, что для широкого класса функций потерь λ вероятность ошибки может быть ограничена сверху суммой средней потери $\tilde{\nu}(a, X)$ и радемахеровской сложности [108]:

Теорема 2.4. *Пусть A — семейство линейных классификаторов, функция $\lambda(M)$ липшицева с константой L_λ и ограничена: $[M<0] \leqslant \lambda(M) \leqslant C$ для всех $M \in \mathbb{R}$. Тогда с вероятностью $1-\eta$ одновременно для всех $a \in A$*

$$P(a) \leqslant \tilde{\nu}(a, X) + 2L_\lambda \mathcal{R}(A, X) + C\sqrt{\frac{2}{\ell}\ln\frac{1}{\eta}}\,. \tag{2.13}$$

Эта теорема служит обоснованием сразу для двух приёмов, широко используемых для обучения алгоритмов классификации.

Первое слагаемое в правой части (2.13) служит обоснованием для *принципа минимизации средних потерь* $\tilde{\nu}(a, X) \to \min\limits_{a \in A}$, согласно которому пороговая функция потерь (индикатор ошибки) заменяется некоторой непрерывной оценкой сверху.

Второе слагаемое служит обоснованием для *принципа регуляризации*, согласно которому к минимизируемому функционалу добавляется *штраф за сложность*. Использование таких оценок в методе *структурной минимизации риска* приводит к тому, что из всех алгоритмов $a \in A$, доставляющих функционалу средних потерь $\tilde{\nu}(a, X)$ значение, близкое к минимальному, будет выбираться алгоритм наименьшей сложности.

Теорема 2.4 является достаточно общей и даёт верхние оценки вероятности ошибки для линейных методов классификации (включая SVM), нейронных сетей [101], алгоритмических композиций типа взвешенного голосования, методов *явной максимизации отступов* (direct optimization of margin) [176].

Метод опорных векторов в спрямляющем пространстве. Для оценивания радемахеровских сложностей $\mathcal{R}(A, X)$ в конкретных случаях применяются их алгебраические свойства, рассмотренные выше. Здесь мы приведём один из наиболее известных примеров, второй пример будет показан в следующем параграфе.

В *методе опорных векторов* (SVM) строится линейный классификатор вида

$$a(x, w) = \operatorname{sign}\left(\sum_{i=1}^{\ell} w_i \langle x_i, x\rangle - w_0\right),$$

где коэффициенты w_i ненулевые только для некоторых объектов выборки x_i, называемых *опорными векторами*.

Чтобы обобщить линейный классификатор и строить нелинейные разделяющие поверхности, скалярное произведение $\langle x_i, x\rangle$ заменяют *ядром* $K(x_i, x)$ — некоторой функцией вида $K\colon \mathscr{X} \times \mathscr{X} \to \mathbb{R}$, удовлетворяющей требованиям симметричности и неотрицательной определённости. Эти требования гарантируют, что функция K

является скалярным произведением в некотором гипотетическом гильбертовом пространстве, называемом также *спрямляющим пространством*. Тогда

$$a(x,w) = \operatorname{sign}\left(\sum_{i=1}^{\ell} w_i K(x_i,x) - w_0\right),$$

а отступы объектов x_i определяются функцией

$$M_i(w) = \left(\sum_{j=1}^{\ell} w_j K(x_i,x_j) - w_0\right) y_i.$$

Применяя алгебраические свойства рademахеровской сложности, рассмотренные в предыдущем параграфе, нетрудно получить из (2.13) следующий результат [108]:

Теорема 2.5. *Пусть ядро ограничено в смысле* $\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{\ell} w_i w_j K(x_i,x_j) \leqslant \lambda^2$. *Тогда для любого* $\theta > 0$ *с вероятностью* $1-\eta$ *одновременно для всех* $a \in A$

$$P(a) \leqslant \nu_\theta(a,X) + \frac{2\lambda}{\theta\ell}\sqrt{\sum_{i=1}^{\ell} K(x_i,x_i)} + \sqrt{\frac{2}{\ell}\ln\frac{2}{\eta}}\,. \tag{2.14}$$

С ростом параметра θ первое слагаемое возрастает, второе уменьшается. Поскольку метод обучения SVM основан на принципе максимизации ширины зазора между классами, правая часть неравенства (2.14) может быть минимизирована путём выбора достаточно большого значения θ.

Оценка (2.14) по своей структуре похожа на fat-оценку (2.12). Второе слагаемое представляет собой *штраф за сложность*. Под радикалом находится отношение среднего квадрата нормы обучающих векторов x_i в спрямляющем пространстве, $D^2 = \frac{1}{\ell}\sum\limits_{i=1}^{\ell} K(x_i,x_i)$, к квадрату зазора между классами θ. При этом рademахеровская оценка точнее, поскольку в ней нет логарифмического множителя, имеющего порядок $\ln^2\ell$.

Основные выводы. Теория *оценок обобщающей способности, зависящих от отступов* (margin-based generalization bounds) даёт критерии выбора непрерывных функций потерь и регуляризаторов, и тем самым позволяет улучшать конструкции методов обучения. Рademахеровская сложность позволяет оценивать вероятность ошибки наиболее точно. Как и другие понятия сложности, она основана на принципе равномерной сходимости, следовательно даёт лишь завышенные оценки худшего случая. Как мы увидим далее, завышенность рademахеровских оценок может приводить к неверным прогнозам переобучения.

2.2.3 Композиции алгоритмов

Композиция алгоритмов — это объединение нескольких базовых алгоритмов, имеющих по отдельности относительно невысокое качество, в один алгоритм более

высокого качества. Повышение качества достигается за счёт того, что ошибки одних алгоритмов компенсируются другими алгоритмами.

Наиболее общая теория алгоритмических композиций разработана в *алгебраическом подходе к построению корректных алгоритмов*, развиваемом научной школой академика Ю. И. Журавлёва [43, 44, 45, 46, 47, 48, 49, 77, 17, 50]. Здесь мы приведём лишь несколько основных определений, довольно сильно их упростив.

Рассматривается задача классификации на два класса, $\mathbb{Y} = \{-1, +1\}$.

Алгоритмической композицией называется суперпозиция вида

$$a(x) = \operatorname{sign} F\big(b_1(x), \ldots, b_T(x)\big), \tag{2.15}$$

где отображения $b_t\colon \mathscr{X} \to \mathbb{R},\ t = 1, \ldots, T$ называются *алгоритмическими операторами*[1], отображение $F\colon \mathbb{R}^T \to \mathbb{R}$ — *корректирующей операцией*.

Наиболее известный вид корректирующей операции — *взвешенное голосование*:

$$a(x) = \operatorname{sign} \sum_{t=1}^{T} \alpha_t b_t(x).$$

Обычно накладывают дополнительные ограничения $\alpha_t \geqslant 0,\ \sum\limits_{t=1}^{T} \alpha_t \leqslant 1$, и тогда говорят о *выпуклой комбинации классификаторов*.

Применяются также упрощённые варианты взвешенного голосования: *простое голосование* ($\alpha_t = \pm 1$) и *голосование с логикой старшинства* [184, 190, 174]. Они не имеют весов, которые можно было бы настраивать по обучающей выборке, следовательно, являются менее гибкими и могут требовать большего числа базовых алгоритмов.

Многие методы классификации, даже считающиеся «простыми», явно или не явно также используют принцип взвешенного голосования, в частности, метод k ближайших соседей, метод парзеновского окна, метод потенциальных функций [2], линейные классификаторы (в частности, метод опорных векторов), нейронные сети [84], алгоритмы вычисления оценок [48], и другие. Поэтому оценки обобщающей способности, получаемые для взвешенного голосования, могут иметь гораздо более широкое применение.

Обобщениями взвешенного голосования являются: *полиномиальные корректирующие операции*, в которых F является полиномом [45, 46, 47, 68]; *смеси экспертов* (mixture of experts), в которых вместо весов базовых алгоритмов α_t используются *функции компетентности* $\alpha_t(x)$ [75, 148]; *монотонные корректирующие операции* [76, 17], в которых F — неубывающая нелинейная функция своих аргументов.

---

[1] В зарубежных публикациях функции b_t называют *базовыми классификаторами* (base classifier) или *слабыми классификаторами* (weak classifier). Если b_t выдают не элементы множества $\mathbb{Y}$, а вещественные оценки степени принадлежности классу $+1$, то их называют также *вещественнозначными классификаторами* (real-valued classifier), хотя, строго говоря, классификаторами в данном случае являются функции $a_t(x) = \operatorname{sign} b_t(x)$.

Все эти конструкции порождают семейства $A = \{a(x)\}$ настолько высокой сложности, что VC-оценки непосредственно к ним не применимы. Ёмкость таких семейств, как правило, бесконечна. Тем не менее, эксперименты показывают, что при «аккуратной» настройке параметров базовых алгоритмов и корректирующей операции композиции часто обладают лучшей обобщающей способностью, чем составляющие их базовые алгоритмы по отдельности.

Одно из первых обоснований этому феномену предложил В. Л. Матросов в начале 80-х годов [63, 64, 65, 66, 67, 68], показав, что для линейных и полиномиальных композиций *алгоритмов вычисления оценок* (ABO) [48] возможно обеспечить корректное распознавание любой заданной выборки, пользуясь подмножеством алгоритмов ограниченной ёмкости. Нетривиальные оценки вероятности ошибки были получены благодаря эксплуатации специфических свойств конкретного семейства базовых алгоритмов, конкретного вида корректирующей операции и конкретного метода обучения. К сожалению, эти оценки сильно завышены, поскольку опираются на VC-теорию — другие подходы в то время были ещё не известны.

Бустинг. В 1988 М. Кернс и Л. Валиант сформулировали следующую проблему [153]. Будем понимать под *слабой обучаемостью* (weak learnability) возможность построить за полиномиальное время алгоритм, вероятность ошибок которого лишь немного меньше 50%, а под *сильной обучаемостью* (strong learnability) — возможность построить за полиномиальное время алгоритм, вероятность ошибок которого сколь угодно мало отличается от нуля при $\ell \to \infty$. Была выдвинута гипотеза, что понятия сильной и слабой обучаемости эквивалентны, то есть что любую слабую модель обучения можно *усилить* (to boost).

Эта гипотеза была теоретически подтверждена в работе Р. Шапира [195], который предложил первый алгоритм *бустинга* (boosting). Годом позже И. Фройнд предложил свой алгоритм [130]. Оба алгоритма основывались на взвешенном голосовании, но были неудобны для практического применения. Лишь пятью годами позже им совместно удалось разработать алгоритм *адаптивного бустинга* AdaBoost, получивший впоследствии широкую известность благодаря простоте, эффективности и высокой обобщающей способности [132, 194]. Аналогичные методы, но с несколько иной стратегией оптимизации базовых классификаторов, были разработаны несколькими годами раньше Д. Уолпертом [224].

Заметим также, что проблема Кернса–Валианта была поставлена и положительно решена гораздо раньше в *алгебраическом подходе к проблеме распознавания* Ю. И. Журавлёва [45, 46, 47]. Однако в алгебраическом подходе проблема обобщающей способности оставалась непроработанной до появления работ В. Л. Матросова, а получение точных или слабо завышенных оценок, пригодных для практического использования, остаётся открытой до сих пор. Тогда как алгоритм AdaBoost был изначально «снабжён» оценками обобщающей способности.

В основе AdaBoost лежат две эвристики. Первая эвристика заключается в том, что базовые классификаторы строятся последовательно. На каждой итерации добавляется очередной базовый алгоритм b_t и определяется коэффициент α_t, при этом

предыдущие элементы композиции $\alpha_1 b_1, \ldots, \alpha_{t-1} b_{t-1}$ полагаются фиксированными. Вторая эвристика заключается в применении экспоненциальной функции потерь $\mathscr{L}(a, x_i) = \exp\bigl(-M_i(\alpha_t, b_t)\bigr)$. Из этих двух эвристик автоматически получаются все остальные результаты: формула расчёта весов α_t, критерий минимума взвешенной частоты ошибок для обучения базового алгоритма b_t, формула для пересчёта весов объектов, теорема о сходимости (безошибочной классификации обучающей выборки) за конечное число шагов. Перед настройкой каждого базового алгоритма, начиная со второго, веса обучающих объектов пересчитываются так, чтобы наибольший вес получили объекты с наименьшими отступами. Это те объекты, на которых чаще ошибались предыдущие алгоритмы. Таким образом, каждый последующий алгоритм стремится компенсировать совокупные ошибки предыдущих. После настройки очередного базового алгоритма b_t вычисляется его вес в композиции α_t, причём чем лучше алгоритм, тем больший вес он получает.

Довольно неожиданным свойством бустинга оказалась его высокая обобщающая способность. В первых экспериментах наблюдалось практически неограниченное улучшение качества обучения при увеличении числа T алгоритмов в композиции [133]. Более того, качество на тестовой выборке, как правило, продолжало улучшаться даже после достижения безошибочного распознавания обучающей выборки. В качестве базовых алгоритмов использовались стандартные решающие деревья. Бустинг над решающими деревьями до сих пор считается одним из наиболее эффективных методов классификации с точки зрения обобщающей способности.

Эти наблюдения противоречат непосредственным выводам VC-теории, основанным только на анализе сложности. В конце 90-х годов одним из самых актуальных направлений в SLT стало исследование феноменов бустинга и поиск нетривиальных оценок его обобщающей способности.

Бэггинг. Похожий на бустинг метод *бэггинга* (bagging, bootstrap aggregation) был предложен практически одновременно Л. Брейманом [114, 115, 116], исходя из несколько иных соображений. Базовые алгоритмы должны существенно различаться, чтобы их ошибки взаимно компенсировались при голосовании. Для повышения *различности* (diversity) предлагается обучать базовые алгоритмы либо на различных подвыборках данных, либо на различных частях признакового описания объектов. Выделение подмножеств объектов и/или признаков можно производить случайным образом — этого вполне достаточно для обеспечения различности. Композиция строится с помощью простого или взвешенного голосования.

Имеется много работ по сравнительному анализу бустинга и бэггинга. Бэггинг направлен исключительно на уменьшение *вариации* (variance), в то время как бустинг способствует уменьшению и вариации, и *смещения* (bias) модели классификации [134]. Эмпирические исследования [207] показывают, что бустинг работает лучше на больших обучающих выборках, бэггинг — на малых. При увеличении длины выборки бустинг повышает различность алгоритмов активнее, чем бэггинг. Бустинг лучше воспроизводит границы классов сложной формы. С другой стороны, бэггинг

гораздо проще распараллеливается, чем бустинг, благодаря тому, что базовые алгоритмы обучаются по отдельности.

Верхние оценки вероятности ошибки для линейных композиций. В работе П. Бартлетта [101] впервые было показано, что эффективная сложность выпуклой комбинации классификаторов равна не суммарной, и даже не максимальной, как ранее предполагалось, а средней взвешенной сложности отдельных классификаторов, взятых с теми же весами α_t, с которыми они входят в комбинацию. Иными словами, взвешенное голосование не увеличивает эффективную сложность семейства алгоритмов, а лишь сглаживает ответы базовых классификаторов. Вытекающие отсюда оценки обобщающей способности существенно точнее классических сложностных VC-оценок, хотя и они всё ещё сильно завышены (требуемая длина обучения имеет порядок 10^4–10^5). Этот результат обосновывает ряд эвристических приёмов, направленных на уменьшение весов синаптических связей при обучении нейронных сетей, таких как *сокращение весов* (weight decay) и *ранний останов* (early stopping).

Результаты, первоначально полученные для линейных комбинаций, оказались применимы и к более широкому классу алгоритмов. В частности, бинарные решающие деревья и дизъюнктивные нормальные формы допускают представление в виде выпуклой комбинации булевых функций с пороговым решающим правилом [139]. Получены оценки обобщающей способности и для более сложных алгоритмических композиций, представимых в виде пороговых выпуклых комбинаций над пороговыми выпуклыми комбинациями. Примерами таких конструкций являются сигмоидальные нейросети с одним скрытым уровнем и взвешенное голосование решающих деревьев [177]. Для всех этих случаев оценки обобщающей способности выражаются через долю обучающих объектов с малым отступом.

Верхние оценки вероятности ошибки для линейных выпуклых композиций классификаторов, в том числе для бустинга и бэггинга, наиболее изящным способом выводятся с помощью рademахеровской сложности [160]. Применяя свойство 3 (стр. 81) и оценку (2.10), находим, что

$$\mathcal{R}(A, X) \leqslant \sqrt{\frac{2h}{\ell} \ln \frac{\ell e}{h}},$$

где h — ёмкость семейства базовых классификаторов B.

Подставляя в (2.13), получаем верхнюю оценку вероятности ошибки:

$$P(a) \leqslant \tilde{\nu}(a, X) + 2L_\lambda \sqrt{\frac{2h}{\ell} \ln \frac{\ell e}{h}} + C\sqrt{\frac{2}{\ell} \ln \frac{1}{\eta}}. \tag{2.16}$$

Отсюда следует основной вывод: обобщающая способность взвешенного голосования классификаторов зависит только от сложности базового семейства B, но не зависит от числа базовых алгоритмов T. Таким образом, количество базовых алгоритмов можно наращивать практически неограниченно.

С другой стороны, бустинг активно максимизирует отступы обучающих объектов по мере увеличения числа базовых алгоритмов. Тем самым уменьшается первое слагаемое $\tilde{\nu}(a, X)$ в правой части оценки. В экспериментах на реальных задачах

классификации строились графики эмпирических распределений отступов объектов на разных итерациях бустинга [133, 196], что позволило заметить важное свойство бустинга — зазор между классами продолжает увеличиваться даже после достижения безошибочной классификации обучающей выборки.

С третьей стороны, в бустинге легко минимизировать вклад второго слагаемого в оценку (2.16), применяя базовые семейства с предельно низкой ёмкостью $h = 1$. Примером являются одномерные пороговые правила (data stumps), неплохо зарекомендовавшие себя на практике

$$B = \Big\{ b(x) = \big[f_j(x) \leqslant d \big],\ j = 1, \ldots, n,\ d \in \mathbb{R} \Big\},$$

где $f_1(x), \ldots, f_n(x)$ — признаки объекта x. Порог d легко оптимизировать по критерию взвешенной частоты ошибок. Столь слабого базового семейства оказывается вполне достаточно для достижения конкурентоспособного качества классификации на широком классе реальных задач.

Ещё одно объяснение феномена бустинга состоит в том, что он строит выпуклую комбинацию вещественнозначных классификаторов, которая проявляет свойство устойчивости [129].

Работы Матросова, Бартлетта, Фройнда, Шапира и др. решительным образом изменили представления о соотношении качества и сложности. Если ранее считалось, что для надёжного восстановления зависимости необходимо ограничивать сложность семейства алгоритмов, то теперь стало очевидно, что семейство A может быть сколь угодно сложным, а первостепенную роль играет *метод обучения*, который по обучающей выборке выбирает алгоритм из A. По всей видимости, некоторые разновидности взвешенного голосования, такие как бустинг, являются «удачными» методами, способными эффективно сужать изначально широкое семейство алгоритмов, подстраиваясь под конкретную задачу.

О переобучении бустинга. Более основательные эксперименты показали, что иногда бустинг всё же переобучается [189, 188]. Если продолжать итерации и наращивать число T базовых алгоритмов, то средняя ошибка на контрольных данных $\nu(a, \bar{X})$ всё же может пройти через точку минимума T^* и далее начать увеличиваться. Значение T^* зависит от конкретной задачи и может достигать нескольких тысяч. Оценка (2.16), как и многие другие завышенные оценки, не способна объяснить этот феномен. Кривая зависимости $\nu(a, \bar{X})$ от T с её характерным минимумом проходит существенно ниже верхней оценки $P(a)$, невозрастающей по T.

В этом заключается «коварство» завышенных оценок вероятности ошибки. Иногда они дают правильные качественные выводы, но если оценка не является количественно точной, то тонкие эффекты могут ускользнуть от внимания исследователей.

На сегодняшний день не существует оценок обобщающей способности, объясняющих, в каких случаях возможно переобучение бустинга. Более того, не существует и математического аппарата, который позволял бы получать такие оценки. По мнению автора статьи [140] причины эффективности бустинга до конца ещё не поняты и его дальнейшее улучшение остаётся открытой проблемой.

О нелинейных композициях алгоритмов классификации. Идея последовательной компенсации ошибок предыдущих алгоритмов реализована также в *оптимизационных (проблемно-ориентированных) методах* алгебраического подхода [76, 16, 17]. В отличие от бустинга, здесь используется не выпуклая комбинация, а более сложная нелинейная *монотонная корректирующая операция*. Монотонность можно рассматривать как обобщение выпуклости: любая выпуклая корректирующая операция является монотонной, обратное в общем случае неверно. При выпуклой коррекции вес каждого базового алгоритма остается постоянным на всём пространстве объектов, что представляется не вполне обоснованной эвристикой. Монотонная коррекция обладает существенно более богатыми возможностями для настройки. С другой стороны, для монотонных корректирующих операций, как для более широкого семейства функций, существенно выше опасность переобучения.

Поэтому значительный теоретический интерес представляет получение оценок обобщающей способности для семейства монотонных отображений. В параграфе 5.3 главы 5 будут рассмотрены комбинаторные оценки функционала полного скользящего контроля для данного семейства.

2.2.4 Стохастические методы обучения: байесовский подход

Байесовский подход к вероятно приближённо корректному обучению[2] (PAC-bayesian approach) [179, 164, 165] является одним из наиболее плодотворных и активно развиваемых современных подходов к оцениванию обобщающей способности. Он позволяет получать достаточно точные оценки [165, 193, 95, 89, 138].

Идея заключается в том, чтобы улучшить оценки «бритвы Оккама» (2.6), (2.7), взяв в качестве *штрафа за сложность* не завышенную величину $\frac{1}{p(a)}$, а меру расхождения неизвестного истинного распределения $q(a) = \mathsf{P}(\mu X{=}a)$ и его априорной оценки $p(a)$. При этом оценки обобщающей способности должны оставаться справедливыми при любом выборе функции p, достигая наибольшей точности при идеальном «угадывании» истинного распределения, $q(a) = p(a)$.

Реализация этой идеи потребовала обобщить концепцию обучения и рассматривать *стохастический метод обучения* $\tilde{\mu}\colon X \mapsto Q$, который по обучающей выборке X

[2]Понятие *вероятно приближённо корректного обучения* (probably approximately correct learning, PAC-learning) введено в работах Валианта [212]. По сути, это другое название той же постановки задачи, что и в VC-теории — получить верхнюю оценку вида $\mathsf{P}\big(\sup_a\big(P(a)-\nu(a,X)\big) \geqslant \varepsilon\big) \leqslant \eta$, где «*вероятно*» означает «с вероятностью $1-\eta$», «*приближённо*» означает «с точностью ε, «*корректное*» означает, что «$\nu(a,X)$ близко к нулю». В исходном варианте теории Валианта рассматривался только детерминистский случай $\nu(a,X) = 0$, позже были сделаны соответствующие обобщения. Теория Валианта отличается от теории Вапника–Червоненкиса, главным образом, привлечением соображений вычислительной сложности. Вводятся понятия полиномиального и неполиномиального (экспоненциального) обучения, чтобы характеризовать зависимость времени обучения от длины выборки и от параметра желаемой точности ε. Подход Валианта представляет, главным образом, теоретический интерес, поскольку на практике дилемма «полиномиальной–экспоненциальной» сложности даже не возникает, а в большинстве современных приложений практическая применимость методов обучения определяется дилеммой «квадратичной–субквадратичной» сложности.

выдаёт не единственный алгоритм $a \in A$, а вероятностное распределение $Q\colon A \to \mathbb{R}$. Если распределение Q сконцентрировано в одной точке, то стохастический метод обучения эквивалентен обычному детерминированному. С другой стороны, если полагать, что всегда выдаётся одно и то же распределение, и затем из этого распределения случайно выбирается алгоритм $a \sim Q$, то такая схема обучения будет эквивалентна предположению о существовании истинного распределения $q(a) \equiv Q(a)$. В общем случае, когда распределение Q зависит от обучающей выборки, стохастический метод обучения является обобщением детерминированного.

Вместо вероятности ошибки $P(a)$ и частоты ошибок $\nu(a, X)$ алгоритма $a = \mu X$, обученного по выборке X, вводятся соответственно их обобщения:

$\nu_Q(X) = \mathsf{E}_{a\sim Q}\nu(a, X)$ — *ожидаемая частота ошибок* (expected empirical error);

$P_Q = \mathsf{E}_{a\sim Q}P(a)$ — *ожидаемая вероятность ошибки* (expected true error).

Впервые верхняя оценка ожидаемой вероятности ошибки была получена МакАллестером в [179]. В качестве меры расхождения распределения Q и его априорной оценки p в этой и практически во всех последующих работах используется *дивергенция Кульбака–Лейблера* (Kullback–Leibler divergence):

$$\mathrm{KL}(Q\|p) = \mathsf{E}_{a\sim Q} \ln \frac{Q(a)}{p(a)}.$$

KL — это несимметричная мера различия вероятностных распределений. С вероятностью 1 справедливо $\mathrm{KL}(Q\|p) = 0 \;\Leftrightarrow\; Q(a) \equiv p(a)$. Другие свойства KL, её верхние и нижние оценки, а также способы применения в оценках обобщающей способности рассматриваются в [164].

Теорема 2.6 (МакАллестер, 1999). *Для любого априорного распределения p и любого $\eta \in (0, 1)$ с вероятностью $1 - \eta$ одновременно для всех распределений Q*

$$P_Q \leqslant \nu_Q(X) + \sqrt{\frac{\mathrm{KL}(Q\|p) + \ln\frac{\ell}{\eta} + 2}{2\ell - 1}}.$$

Эта оценка похожа на сложностные оценки, только теперь роль штрафа за сложность играет KL-дивергенция двух распределений: априорного p и апостериорного Q, в общем случае зависящего от обучающей выборки X.

В работах [168, 198] оценка МакАллестера была уточнена, также были найдены более простые доказательства. Приведём окончательный вариант согласно [165]:

Теорема 2.7 (Лангфорд, 2002). *Для любого априорного распределения p и любого $\eta \in (0, 1)$ с вероятностью $1 - \eta$ одновременно для всех распределений Q*

$$\mathrm{KL}\big(\nu_Q(X)\|P_Q\big) \leqslant \frac{1}{\ell}\mathrm{KL}(Q\|p) + \frac{1}{\ell}\ln\frac{\ell+1}{\eta}. \tag{2.17}$$

В левой части KL-дивергенцию между двумя числовыми величинами из интервала $(0, 1)$ следует понимать в смысле $\mathrm{KL}(\nu\|P) = \nu\ln\frac{\nu}{P} + (1-\nu)\ln\frac{1-\nu}{1-P}$. Чтобы выразить верхнюю оценку P через ν, на практике применяют численные методы.

Если стохастический метод обучения выдаёт распределение Q, сконцентрированное в точке a, то есть работает как обычный детерминированный метод обучения, то $\mathrm{KL}(Q\|p) = \ln \frac{1}{p(a)}$, и оценка (2.17) переходит в оценку Оккама (2.6) с точностью до «лишнего» слагаемого $\frac{1}{\ell}\ln(\ell+1)$. Пока остаётся не ясно, возможно ли устранить этот добавочный член, или это естественная плата за то, что оценка (2.17) верна для любого распределения Q.

Стохастические методы обучения непосредственно не применяются на практике. Однако полученные для них оценки в ряде случаев удаётся приспособить для обоснования и усовершенствования нестохастических методов обучения.

Чтобы применить теорему 2.7, необходимо распорядиться свободой выбора как априорного распределения p, так и апостериорного Q. Этот выбор граничит с искусством, поскольку к функциям p и Q предъявляется нетривиальная совокупность требований. Во-первых, величина $\mathrm{KL}(Q\|p)$ должна удобно оцениваться аналитически. Во-вторых, распределения p и Q должны действительно оказаться близкими, иначе оценка будет неточной. В-третьих, полученная оценка должна быть полезной для конструктивного улучшения метода обучения.

Известно много применений PAC-байесовского подхода к различным методам обучения: к взвешенному голосованию классификаторов [95, 170, 137], к простому голосованию логических закономерностей [193], к машинам покрывающих множеств [171], и др. Мы рассмотрим в качестве «канонического» примера его применение к линейным алгоритмам классификации.

Линейные классификаторы. Рассмотрим задачу классификации на два класса $\mathbb{Y} = \{-1, +1\}$ и *линейный классификатор* $a(x, w) = \mathrm{sign}\langle x, w\rangle,\ \ x \in \mathbb{R}^n$ с направляющим вектором разделяющей гиперплоскости $w \in \mathbb{R}^n$.

Нормированным отступом (normalized margin) объекта x_i называется величина

$$M_i(w) = \frac{y_i\langle x_i, w\rangle}{\|x_i\| \cdot \|w\|}.$$

В остальном будем придерживаться обозначений, введённых выше на стр. 83.

Следующая конкретизация распределений p и Q предложена в [169].

Положим $p(w) = \mathcal{N}(w; 0, I_n)$ — n-мерное нормальное распределение с нулевым вектором математического ожидания и единичной ковариационной матрицей.

Положим $Q(w) = \mathcal{N}(w; w_0, I_n)$ — n-мерное нормальное распределение с вектором математического ожидания w_0 и единичной ковариационной матрицей. Вектор w_0 является параметром и характеризует наиболее вероятное положение направляющего вектора разделяющей гиперплоскости. Потом w_0 окажется тем самым направляющим вектором, который будет строиться нестохастическим методом обучения по выборке X. Обозначим $\beta = \|w_0\|$.

При сделанном выборе функций p и Q доказываются следующие факты.

Во-первых, $\mathrm{KL}(Q\|p) = \frac{1}{2}\beta^2$, зависит только от нормы вектора w_0.

Во-вторых, функция $\nu_Q(X)$ есть *средняя потеря на выборке* X при *вещественной функции потерь* $\mathscr{L}(a, x_i) = \Phi\big(-\beta M_i(w)\big)$ и при $w = w_0$:

$$\nu_Q(X) = \frac{1}{\ell}\sum_{i=1}^{\ell}\Phi\big(-\beta M_i(w_0)\big),$$

где $\Phi(z) = \int\limits_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\exp\big(-\frac{1}{2}t^2\big)\,dt$ — функция одномерного нормального распределения.

В-третьих, справедлива следующая конкретизация теоремы 2.7:

Теорема 2.8 (Лангфорд и Сиджер, 2002). *Для любого распределения на множестве объектов* $\mathscr{X}$ *и любого* $\eta \in (0,1)$ *с вероятностью* $1-\eta$ *одновременно для всех* w_0

$$\mathrm{KL}\big(\nu_Q(X)\|P_Q\big) \leqslant \frac{\beta^2}{2\ell} + \frac{1}{\ell}\ln\frac{\ell+1}{\eta}. \tag{2.18}$$

Чтобы минимизировать ожидаемую вероятность ошибки P_Q, необходимо минимизировать функционал средних потерь $\nu_Q(X)$ и одновременно норму направляющего вектора разделяющей гиперплоскости $\beta = \|w_0\|$. Таким образом, (2.18) даёт оптимизационный критерий для поиска параметра w_0 по обучающей выборке X, похожий на *принцип максимума зазора между классами*, лежащий в основе *метода опорных векторов* (SVM). Соответствие с SVM, однако, не полное.

В SVM оптимизационная задача обучения вектора w_0 по выборке X ставится следующим образом: $\|w_0\|^2 + C\sum\limits_{i=1}^{\ell}\big(1-M_i(w_0)\big)_+ \to \min\limits_{w_0}$, где C — некоторая константа.

Первое отличие: в SVM функция потерь кусочно-линейная (выпуклая, неограниченная), тогда как в нашем случае это нормальная функция распределения (гладкая, невыпуклая, ограниченная). Из-за существенных различий в свойствах функций потерь для решения оптимизационной задачи должны использоваться принципиально различные методы.

Второе отличие: из-за нелинейности функции KL минимизация ожидаемой вероятности ошибки P_Q непосредственно не сводится к минимизации *суммы* двух функционалов $\|w_0\|^2$ и $\nu_Q(X)$. Вместо этого приходится численно решать уравнение (2.18) относительно верхней оценки $\bar{P}_Q$ при различных значениях параметра β.

Третье отличие: вместо подбора параметра C приходится подбирать параметр β. Однако этот выбор не требует трудоёмкого скользящего контроля, и с помощью (2.18) может быть выполнен очень эффективно.

Остался последний вопрос: оценка выводилась для стохастического метода обучения, так почему же она применяется для построения обычного нестохастического метода обучения? Ответ даёт теорема, утверждающая, что если определить *усреднённый классификатор* (averaging classifier)

$$\bar{a}(x) = \operatorname{sign}\int_A a(x)\,dQ(a) \equiv \operatorname{sign}\mathsf{E}_{a\sim Q}a(x),$$

то оказывается, что $\bar{a}(x) = \text{sign}\langle w_0, x\rangle$. Таким образом, строить единственный вектор w_0 — это всё равно, что строить распределение $Q(w) = \mathcal{N}(w; w_0, I_n)$ и затем классификацию произвольного объекта x производить путём усреднения (интегрирования) по всему семейству линейных классификаторов $w \sim Q$.

В более поздних работах [95, 89, 138] рассматриваются другие варианты конкретизации p и Q и показывается, что они приводят к ещё более точным оценкам для линейных классификаторов.

Выводы. PAC-байесовский подход в настоящее время активно развивается и позволяет получать одни из самых точных оценок. Однако точность существенно зависит от того, насколько удачен априорный выбор распределений p и Q. Кроме того, приходится доказывать, что получаемый в итоге нестохастический метод обучения строит усреднённый классификатор для исходного стохастического метода обучения.

Заметим, что радемахеровский подход даёт сопоставимые по точности оценки, и в то же время лишён указанных недостатков.

2.3 Оценки, учитывающие расслоение алгоритмов

Эффект *расслоения семейства алгоритмов* связан с тем, что восстановление фиксированной зависимости $y(x)$ с помощью конкретного метода обучения μ по случайной выборке X индуцирует на множестве алгоритмов A существенно неравномерное распределение вероятностей $P_a = \mathsf{P}(a{=}\mu X)$.

Весьма существенная часть семейства может получить ничтожно малую вероятность, и тогда можно говорить об эффекте *локализации семейства* как о частном случае расслоения.

Пренебрежение эффектом расслоения — это очень существенный фактор завышенности многих сложностных оценок обобщающей способности. Если понятие сложности определяется через подсчёт числа различных алгоритмов в семействе, то неявно это означает, что всем алгоритмам приписываются равные вероятности.

В параграфе 2.3.1 рассматриваются подходы, в которых учитывается только локализация, то есть выделяется подмножество алгоритмов с достаточно высокими вероятностями P_a, но в рамках этого подмножества расслоение не учитывается. Некоторые универсальные обобщения этих подходов рассматривается в параграфе 2.3.2. В параграфе 2.3.3 рассматриваются оценки расслоения, учитывающие неравномерность распределения P_a.

Все эти подходы естественным образом сочетаются со структурной минимизации риска, причём, в отличие от классической VC-теории, структура вложенных подсемейств может зависеть от обучающей выборки.

2.3.1 Самооценивающие методы обучения

При использовании оценок, зависящих от задачи, метод структурной минимизации риска трансформируется и приводит к построению *самооценивающих методов*

обучения (self bounding learning algorithms) [131]. От исходного метода структурной минимизации риска они отличаются тем, что структура вложенных подсемейств не задаётся заранее, а формируется в процессе обучения. Результатом обучения является не только сам алгоритм, но и оценка его обобщающей способности, которая также вычисляется в процессе обучения.

Первоначальной мотивацией для введения принципа самооценивания послужили следующие соображения о методах обучения *решающих деревьев*, таких как ID3 [186], C4.5 [187], CART [117], LISTBB [39] и другие [71]. На каждом шаге построения дерева одна из внутренних вершин разветвляется на две, при этом условие ветвления подбирается по некоторому критерию информативности. Если предположить, что на каждом шаге выбранный оптимум критерия информативности оказывается существенно лучше всех остальных альтернативных условий ветвления, то, скорее всего, другая обучающая выборка при той же *целевой зависимости* приведёт к построению ровно такого же решающего дерева. Тогда сложность используемого семейства алгоритмов будет равна единице, а не мощности множества всех различных решающих деревьев. Вообще, основная проблема состоит в том, чтобы решить, какие именно альтернативы являются заведомо плохими и не должны учитываться при оценивании локальной сложности семейства, зависящей от задачи.

Аналогичные соображения переносятся на широкий класс методов обучения, основанных на *статистических запросах* (statistical query) [150, 152, 147, 146]. Это достаточно общая парадигма обучения, охватывающая не только методы индукции решающих деревьев, но и локальные методы синтеза логических закономерностей, градиентные итерационные методы, в частности, *метод обратного распространения ошибок* (error back propagation, BackProp) — один из стандартных методов обучения нейронных сетей.

Статистический запрос — это предикат $\varphi\colon \mathscr{X} \times \mathbb{Y} \to \{0,1\}$ из заданного семейства предикатов Φ. Ответом на запрос является оценка вероятности события $\varphi(x,y) = 1$, вычисленная по обучающей выборке X:

$$\hat{p}(\varphi, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \varphi(x_i, y_i).$$

Предполагается, что обучение — это итерационный процесс, на каждом шаге которого выбирается некоторый статистический запрос $\varphi_t \in \Phi$ и вычисляется ответ на него $\hat{p}_t = \hat{p}(\varphi_t, X)$. Эта информация используется для того, чтобы ещё немного конкретизировать конструкцию алгоритма. Тем самым сужается подмножество A_t, в котором потенциально может находиться результат обучения — алгоритм $a = \mu X$. Критерий прекращения итераций в общем случае не фиксируется.

Отклонение каждого ответа от ожидаемого значения $p_t = \mathsf{E}\varphi(x,y)$ оценивается с помощью стандартных доверительных интервалов. Для некоторых алгоритмов классификации, таких как решающие деревья, знания значений $p_1, \ldots, p_t$ достаточно, чтобы оценить вероятность ошибки. Таким образом, оценка обобщающей способности постепенно уточняется в ходе итераций [131].

Алгоритм 2.3.1. Обучение на основе статистических запросов.

1: **для всех** $t = 1, \ldots, T$, пока не выполнится условие прекращения итераций
2: выбрать запрос $\varphi_t \in \Phi$ на основе ответов предыдущих итераций $\hat{p}_1, \ldots, \hat{p}_{t-1}$;
3: вычислить ответ $\hat{p}_t = \hat{p}(\varphi_t, X)$;
4: выдать результат обучения $a \in A$ на основе накопленной информации.

Чтобы применить метод *структурной минимизации риска*, не нужно заранее знать сами множества A_t, достаточно знать только их мощности или иные характеристики сложности. Таким образом, снимается требование классической VC-теории, чтобы структура вложенных подсемейств задавалась до того, как станет известна обучающая выборка X.

Эти идеи были обобщены в работах Дж. Лангфорда [166, 164] введением понятия *микровыборов* (microchoice). Для применения структурной минимизации риска, вообще говоря, не важно, выполняет ли метод обучения на каждой итерации t статистический запрос, или каким-либо другим способом производит выбор из множества альтернатив, сужая тем самым подмножество A_t. Выбор, производимый на каждой итерации, называется *микровыбором*. Если известно количество альтернатив на каждой итерации, то можно оценить сверху сложность всего множества алгоритмов, которые могут быть получены в результате всевозможных последовательностей микровыборов. В таком виде подсчёт микровыборов не зависит от задачи и является не более чем ещё одним способом оценить сверху функцию роста. Однако на практике, когда выборка и целевая зависимость фиксированы, альтернативы на каждой итерации становятся существенно неравноценными.

Как правило, имеется небольшое число подходящих альтернатив и большое число неподходящих. Метод *адаптивных микровыборов* (adaptive microchoice) [166, 164] позволяет отбросить неподходящие альтернативы и оценить сложность эффективно используемого подсемейства алгоритмов. При этом удаётся ограничить возможные потери от отбрасывания неподходящих альтернатив.

Хотя рассмотренные подходы учитывают локализацию семейства алгоритмов, оценки обобщающей способности всё же остаются завышенными. Проблема в том, что все они основаны на подсчёте числа различных алгоритмов, то есть, фактически, опираются на неравенство Буля и не учитывают эффектов сходства алгоритмов.

В экспериментах на реальных данных [164] оценки метода адаптивных микровыборов несколько уступали по точности оценкам расслоения, которые рассматриваются ниже в параграфе 2.3.3.

2.3.2 Функции удачности и подсемейства, зависящие от выборки

В работе Дж. Шо-Тейлора и П. Бартлетта [199] вводится понятие *функции удачности* (luckiness function), с помощью которой все алгоритмы семейства ранжируются по предпочтительности относительно заданной выборки. Это ранжирование,

фактически, и задаёт систему вложенных подсемейств для последующего применения структурной минимизации риска. В роли функции удачности могут выступать любые характеристики сложности семейства, но не только. Подойдёт любая функция, удовлетворяющая специальному требованию ω-малости (ω-smallness). В частности, для метода SVM функция удачности определяется через отступы объектов обучающей выборки. В общем случае не вполне понятно, как строить функции удачности, так как они не несут собственного содержательного смысла, а являются лишь формальным обобщением нескольких ранее известных конструкций.

Развитием этой концепции стало понятие *функции удачности метода обучения* (algorithmic luckiness function), введённое Р. Хербричем и Р. Уильямсоном [142]. От обычной функции удачности она отличается тем, что учитываются лишь алгоритмы локального подмножества $A_L^\ell(\mu, \mathbb{X}) = \{\mu X \colon X \in [\mathbb{X}]^\ell\}$, которое определяется так же, как и в данной работе, см. стр. 59. Это одна из немногих работ в теории статистического обучения, непосредственно эксплуатирующая понятие *метода обучения*. К сожалению, усилия, направленные на максимально полный учёт эффекта расслоения, сводятся на нет применением завышенных сложностных оценок, основанных на мощности покрытия и не учитывающих эффект сходства алгоритмов.

Последующим обобщением стало понятие *подсемейства, зависящего от выборки* (random subclass) [185], подразумевающее, что теперь в функционале равномерной сходимости (2.1) супремум берётся не по всему семейству A, а по некоторому подсемейству $A(X) \subseteq A$, зависящему от обучающей выборки X. В этом подходе ставится задача получения верхних оценок вида

$$\mathsf{P}\Big\{ \sup_{a \in A(X)} \big(P(a) - \nu(a, X)\big) > \varepsilon \Big\} \leqslant \eta(\varepsilon),$$

которая путём *симметризации* сводится к задаче получения верхних оценок вида

$$\mathsf{P}\Big\{ \sup_{a \in A(X)} \big(\nu(a, \bar{X}) - \nu(a, X)\big) > \varepsilon \Big\} \leqslant \eta(\varepsilon).$$

В предельном случае, если взять одноэлементное множество $A(X) = \{\mu X\}$, данный функционал переходит в функционал *вероятности переобучения*. Задача его оценивания технически сложна и требует привлечения слишком детальных знаний о методе обучения μ. Можно даже констатировать как факт, что теория статистического обучения старательно избегает решения данной задачи на протяжении всего периода своего существования.

В другом предельном случае, когда семейство не зависит от выборки, $A(X) \equiv A$, получаем стандартный принцип равномерной сходимости по Вапнику-Червоненкису.

В промежуточных случаях можно рассчитывать на некий компромисс: подсемейство $A(X)$ должно быть достаточно широким, чтобы упростить анализ и гарантировать выполнение условия $\mu X \in A(X)$, но и достаточно узким, чтобы учесть эффект локализации и получить оценки, существенно более точные, чем VC-оценки.

В работе [185] показывается, что многие предложенные ранее подходы естественным образом переформулируются в терминах подсемейств, зависящих от выборки.

В их числе: локальные сложностные оценки, в том числе рademахеровские, самооценивание, статистические запросы, функции удачности, и другие.

Выводы. В данном параграфе вкратце упомянуто несколько подходов, претендующих на роль удобного универсального языка для получения широкого класса оценок, зависящих от задачи. Радикального улучшения точности оценок эти подходы пока не дали, но они способствуют лучшему пониманию взаимосвязей между различными направлениями внутри теории статистического обучения.

2.3.3 Оценка расслоения по Лангфорду

Оценка расслоения по Лангфорду (shell bound) при совместном использовании с методом *структурной минимизации риска* показала наиболее точные результаты в экспериментах на реальных задачах классификации [164, 167]. Поэтому мы рассмотрим этот подход подробнее. Он приводит к довольно громоздким оценкам, которые не выражаются в виде явных формул, а вычисляются в процессе обучения.

Вводится функция $P(\varepsilon, s)$ — верхняя оценка того, что в конечном семействе A найдётся переобученный алгоритм a с s ошибками на обучающей выборке, то есть такой алгоритм, для которого одновременно $\nu(a, X) = \frac{s}{\ell}$ и $P(a) > \frac{s}{\ell} + \varepsilon$:

$$P(\varepsilon, s) = \sum_{a \in A} \big[P(a) > \tfrac{s}{\ell} + \varepsilon\big] \operatorname{bin}_\ell\big(\tfrac{s}{\ell}, P(a)\big).$$

Эта верхняя оценка выводится с помощью неравенства Буля.

Функция $P(\varepsilon, s)$, вообще говоря, неизвестна, так как в ней используется неизвестная вероятность ошибки $P(a)$. Но если бы она была известна, то мы имели бы *ненаблюдаемую оценку*, называемую в оригинальной работе «оценкой полного знания» (full knowledge bound): с вероятностью $1 - \eta$ одновременно для всех $a \in A$

$$P(a) \leqslant \nu(a, X) + \varepsilon\big(\eta, \nu(a, X)\big), \tag{2.19}$$

где $\varepsilon\big(\eta, \frac{s}{\ell}\big) = \min\big\{\varepsilon\colon P(\varepsilon, s) \leqslant \frac{\eta}{\ell}\big\}$ — функция, обратная к $P(\varepsilon, s)$ по аргументу ε. Заметим, что здесь уровень значимости η делится поровну между всеми возможными значениями числа ошибок на наблюдаемой выборке. Это довольно грубая оценка, за которой, опять-таки, стоит неравенство Буля.

Следующий шаг заключается в том, чтобы оценить функцию $P(\varepsilon, s)$ сверху по наблюдаемой выборке X. Для этого доказывается, что с вероятностью $1 - \eta$

$$P(\varepsilon, s) \leqslant \hat{P}(\varepsilon, s, \eta) = \sum_{i=1}^{\ell} \Delta_i \operatorname{bin}_\ell\Big(\tfrac{s}{\ell}, \max\big\{\tfrac{s}{\ell} + \varepsilon, \underline{e}\big(\tfrac{i}{\ell}, \eta\big)\big\}\Big),$$

где $\Delta_i = \#\big\{a \in A\colon n(a, X) = i\big\}$ — *наблюдаемый профиль расслоения*, вычисляемый по обучающей выборке,

$$\underline{e}\big(\tfrac{i}{\ell}, \eta\big) = \min\big\{p\colon \operatorname{Bin}_\ell\big(\tfrac{i}{\ell}, p\big) \leqslant 1 - \eta\big\}$$

— нижняя граница доверительного полуинтервала вероятности ошибки, оценённая для выборки с частотой ошибок $\frac{i}{\ell}$. Для получения оценки $\hat{P}$ приходится в третий раз использовать неравенство Буля.

Имея функцию $\hat{P}(\varepsilon, s, \eta)$, которая, в отличие от $P(\varepsilon, s)$, определяется только по наблюдаемой выборке X, можно вычислить т. н. *наблюдаемую оценку расслоения* (observable shell bound): с вероятностью $1-\eta$ одновременно для всех $a \in A$

$$P(a) \leqslant \nu(a, X) + \hat{\varepsilon}\big(\eta, \nu(a, X)\big),$$

где $\hat{\varepsilon}\big(\eta, \frac{s}{\ell}\big) = \min\big\{\varepsilon\colon \hat{P}\big(\varepsilon, s, \frac{1}{2}\eta\big) \leqslant \frac{\eta}{2\ell}\big\}$ — функция, обратная к $\hat{P}\big(\varepsilon, s, \frac{1}{2}\eta\big)$ по ε.

Осталось решить последнюю проблему: как оценить наблюдаемый профиль расслоения $(\Delta_i)_{i=0}^{\ell}$, если в большинстве практических случаев перебрать всё множество A не представляется возможным. В [164] предлагается оценивать Δ_i *методом Монте-Карло*, выбирая относительно небольшое случайное подмножество алгоритмов $A' \subset A$: $\Delta'_i = \#\big\{a \in A'\colon n(a, X) = i\big\}$. Тогда верхняя граница для Δ_i оценивается с помощью той же техники обращения функции биномиального распределения. А именно, с вероятностью $1-\eta$

$$\Delta_i \leqslant \bar{\Delta}_i\big(|A'|, \Delta'_i, \eta, |A|\big) = |A| \cdot \max_p \Big\{p\colon \operatorname{Bin}_{|A'|}\big(\tfrac{\Delta'_i}{|A'|}, p\big) = \eta\Big\}.$$

Окончательная *оценка расслоения по Лангфорду* формулируется следующим образом: с вероятностью $1-\eta$ одновременно для всех $a \in A$

$$P(a) \leqslant \nu(a, X) + \tilde{\varepsilon}\big(\eta, \nu(a, X)\big),$$

где

$$\tilde{\varepsilon}\big(\eta, \tfrac{s}{\ell}\big) = \min\big\{\varepsilon\colon \tilde{P}\big(\varepsilon, s, \tfrac{1}{2}\eta\big) \leqslant \tfrac{\eta}{2\ell}\big\};$$

$$\tilde{P}\big(\varepsilon, s, \eta\big) = \sum_{i=1}^{\ell} \bar{\Delta}_i\big(|A'|, \Delta'_i, \tfrac{\eta}{2\ell}, |A|\big) \cdot \operatorname{bin}_\ell\Big(\tfrac{s}{\ell}, \max\big\{\tfrac{s}{\ell} + \varepsilon, \underline{e}\big(\tfrac{i}{\ell}, \tfrac{1}{2}\eta\big)\big\}\Big),$$

функции $\bar{\Delta}_i$ и $\underline{e}$ определены выше.

Полученная оценка расслоения точнее VC-оценок по нескольким причинам.

1. Хотя неравенство Буля применяется трижды, но каждый раз ко множеству, существенно меньшему, чем всё семейство алгоритмов.

2. VC-оценки (1.53) и (1.60) также выражаются через ненаблюдаемый профиль расслоения $\Delta_m(\mu, \mathbb{X})$. Ничто не мешает оценить его по наблюдаемому профилю Δ_i, аналогично тому, как это было проделано выше. Однако такие оценки будут существенно менее точны, чем оценка Лангфорда, которая явно учитывает, что метод обучения — это *минимизация эмпирического риска*, и оценивает вероятность получения алгоритма из i-го слоя в результате обучения. В оценке Лангфорда вклады слоёв (коэффициенты при $\bar{\Delta}_i$) довольно быстро убывают с ростом i. В то же время, в VC-оценках вклады слоёв (коэффициенты при Δ_m) имеют пологий максимум в окрестности $m = L/2$, см. рис. 1.6 на стр. 39. Это означает, что в VC-оценках наиболее мощные средние слои учитываются с наибольшим весом.

Мы достаточно подробно описали общий ход построения оценки расслоения Лангфорда (с техническими деталями и доказательствами это заняло бы 30–40 страниц), преследуя несколько целей.

1. Показать, как осуществляется переход от ненаблюдаемых оценок к наблюдаемым. Этот переход достаточно громоздкий и требует численного обращения функций распределения. Аналогичная техника может быть применена для перехода от ненаблюдаемых оценок к наблюдаемым в рамках слабой аксиоматики. Основное отличие в том, что вместо биномиального распределения придётся обращать гипергеометрическое. В последующих главах мы опускаем такого рода построения[3].

2. Хотя оценка расслоения Лангфорда позиционируется как *численно точная* (quantitatively tight sample complexity bound), она всё же довольно сильно завышена. Она учитывает эффект расслоения, но опирается на неравенство Буля, следовательно, не учитывает эффект сходства алгоритмов. Численные эксперименты, проведённые Д. Кочедыковым, показывают, что оценки расслоения Лангфорда в некоторых практически интересных случаях несущественно лучше VC-оценок. Они начинают сильно выигрывать только при оценивании подсемейств малой ёмкости, например, в методе структурной минимизации риска.

2.4 Оценки, учитывающие сходство алгоритмов

Важный фактор завышенности VC-оценок — пренебрежение сходством алгоритмов в результате применения неравенства Буля. Это неравенство тем сильнее завышено, чем более схожи векторы ошибок алгоритмов. Влияние сходства алгоритмов на вероятность переобучения изучалось относительно мало.

Мы рассмотрим три идеи учёта сходства: кластеризацию или покрытие семейства ε-сетью [105], представление семейства в виде связного графа [205] и устойчивость метода обучения относительно малых вариаций обучающей выборки [112, 113, 163]. Во всех трёх случаях радикального улучшения VC-оценок добиться не удаётся, хотя третья идея оказалась более плодотворной и позволила получить нетривиальные оценки для некоторых семейств бесконечной ёмкости.

2.4.1 Кластеризация семейства алгоритмов

В работе Э. Бакса [105] оценивается вероятность большого равномерного отклонения частот в двух выборках. Фактически, задача ставится в рамках слабой аксиоматики — оценить сверху функционал

$$\widetilde{Q}_\varepsilon(A) = \mathsf{P}\Big[\max_{a\in A}\big(\nu(a,\bar{X}) - \nu(a,X)\big) \geqslant \varepsilon\Big].$$

[3]Перенос теории Лангфорда в слабую аксиоматику с несколькими исправлениями и уточнениями был недавно выполнен Д. Кочедыковым (готовится публикация).

Вводится *расстояние Хэмминга* между векторами ошибок алгоритмов

$$\rho(a,a') = \sum_{i=1}^{L} \big|I(a,x_i) - I(a',x_i)\big|, \quad \forall a,a' \in A,$$

и предполагается, что на множество векторов ошибок можно выделить r-сеть мощности не более $\mathcal{N}(r,A)$. Иначе говоря, множество алгоритмов A разбивается на $\mathcal{N}(r,A)$ кластеров радиуса не более r каждый.

Теорема 2.9 (Бакс). *Для произвольного семейства A, любого $\varepsilon \in (0,1)$ и любого значения $r \in \big[0, \varepsilon \min\{\ell,k\}\big]$*

$$\mathsf{P}\Big[\max_{a\in A}\big(\nu(a,\bar{X}) - \nu(a,X)\big) > \varepsilon\Big] \leqslant \mathcal{N}(r,A) \max_m H_L^{\ell,m}\Big(s_1\big(\varepsilon - \tfrac{r}{\min\{\ell,k\}}\big)\Big).$$

При $r=0$ данная оценка переходит в VC-оценку, поскольку $\mathcal{N}(0,A) = \Delta^A(L)$. При увеличении r мощность покрытия $\mathcal{N}(r,A)$ уменьшается, но при этом увеличивается гипергеометрический сомножитель. Поэтому оценка Бакса не даёт существенного улучшения точности по сравнению с VC-оценками. Она остаётся сильно завышенной даже после минимизации правой части по r.

Для частного случая, когда семейство A линейно по параметрам, в [105] доказывается, что $\mathcal{N}(r,A) \leqslant \frac{1}{2r+1}\Delta^A(L)$. По всей видимости, этот результат может быть улучшен, так как размер кластеров должен расти не линейно по радиусу кластера r, а пропорционально r^d, где d — размерность пространства.

2.4.2 Связные семейства алгоритмов

В работах Ж. Силла [205, 203] рассматриваются параметрические семейства алгоритмов классификации $A = \big\{a(x,w) = \operatorname{sign} f(x,w) \colon w \in \mathbb{R}^m\big\}$ с непрерывной по вектору параметров w *дискриминантной функцией* $f(x,w)$. Доказывается *теорема связности*, утверждающая, что если плотность распределения на множестве $\mathscr{X}$ непрерывна, то множество векторов ошибок $\vec{A} = \big\{\vec{a} = (I(a,x_i))_{i=1}^L \colon a \in A\big\}$ с вероятностью 1 образует связный граф, рёбра которого соответствуют парам векторов, отличающихся только на одном объекте. Иными словами, при непрерывном изменении любой из координат вектора параметров w каждое изменение вектора ошибок алгоритма $a(x,w)$ с вероятностью 1 происходит только на одном объекте, а одновременное изменение нескольких координат имеет нулевую вероятность.

Свойством *связности* обладают многие алгоритмы классификации с непрерывной по параметрам разделяющей поверхностью: линейные классификаторы, машины опорных векторов с непрерывными ядрами, нейронные сети с непрерывными функциями активации, решающие деревья с пороговыми условиями над непрерывными признаками, и многие другие.

Теорема 2.10 (Силл). *Для произвольного связного семейства A*

$$\mathsf{P}\Big[\max_{a\in A}\big(\nu(a,\bar{X}) - \nu(a,X)\big) > \varepsilon\Big] \leqslant \frac{6\Delta^A(L)}{\sqrt{\pi L}} \exp\big(-\varepsilon^2 L\big).$$

Данный результат отличается от VC-оценки (1.50) множителем $\sqrt{\pi L}$, что существенно меньше степени завышенности VC-оценки. Таким образом, подход Силла не даёт радикального улучшения точности. Причина, по всей видимости, в том, что при получении оценки используется лишь тот факт, что любые два вектора ошибок можно соединить путём на графе, но не учитывается количество связей.

В параграфе 4.3 будет введён параметр связности q, выражающий среднее число вершин, смежных с произвольной вершиной, а также будет показано, что VC-оценка вероятности переобучения может быть улучшена на множитель порядка 2^{-q}.

2.4.3 Устойчивость метода обучения

Третья идея анализа сходства алгоритмов связана с понятием *устойчивости* (stability) метода обучения [112, 113, 163]. В отличие от предыдущих подходов, здесь делается попытка учесть не столько структуру семейства алгоритмов A, сколько свойства метода обучения μ.

Метод обучения называется устойчивым, если небольшие вариации обучающей выборки, такие как вставка, удаление или замена одного обучающего объекта, приводят к незначительным изменениям получаемого алгоритма. Существуют различные способы формализации этого общего представления об устойчивости. В работе [163] вводится 12 различных определений и устанавливаются взаимосвязи между ними.

Как правило, оценки качества устойчивых методов обучения не зависят от сложностных характеристик семейства. Первые оценки обобщающей способности через устойчивость были получены в конце 70-х годов для локальных методов классификации типа ближайших соседей и потенциальных функций [191, 124, 125]. Эти методы порождают семейства алгоритмов бесконечной ёмкости. Позже была доказана устойчивость бустинга [162] и других методов взвешенного голосования [129], машин опорных векторов, методов минимизации эмпирического риска с регуляризирующей штрафной функцией, рандомизированных методов обучения [128], и других.

Приведём в качестве примера оценку, основанную на понятии *равномерной устойчивости* (uniform stability). Рассмотрим задачу классификации с двумя классами $\mathbb{Y} = \{-1, +1\}$, семейство алгоритмов классификации вида $a(x, w) = \operatorname{sign} f(x, w)$ и функцию потерь вида $\mathscr{L}(a, x_i) = \lambda\big(y_i f(x_i, w)\big)$, ограниченную константой C. Обозначим через $f_X(x, w)$ дискриминантную функцию, полученную методом обучения μ по обучающей выборке $X = (x_i, y_i)_{i=1}^{\ell}$. Обозначим через $X|i$ выборку X, в которой обучающий прецедент (x_i, y_i) заменён прецедентом (x'_i, y'_i). Будем говорить, что метод μ *удовлетворяет условию равномерной устойчивости*, если существует такая функция $\beta(\ell)$, что для любой выборки X и любой замены $(x_i, y_i) \to (x'_i, y'_i)$

$$\big|\lambda\big(y f_X(x, w)\big) - \lambda\big(y f_{X|i}(x, w)\big)\big| \leqslant \beta(\ell), \quad \text{для всех } (x, y) \in \mathscr{X} \times \mathbb{Y}.$$

Из условия равномерной устойчивости нетрудно получить, что

$$\mathsf{E}_X\big(\tilde{P}(\mu X) - \tilde{\nu}(\mu X, X)\big) \leqslant \beta(\ell).$$

Отсюда и из неравенства ограниченных разностей МакДиармида следует, что с вероятностью не менее $1-\eta$

$$\tilde{P}(\mu X) \leqslant \tilde{\nu}(\mu X, X) + \beta(\ell) + \big(2\ell\beta(\ell) + C\big)\sqrt{\frac{1}{2\ell}\ln\frac{1}{\eta}}.$$

Чтобы воспользоваться данной оценкой, необходимо показать, что для конкретного метода обучения μ функция $\beta(\ell)$ является невозрастающей, более того, $\sqrt{\ell}\beta(\ell) \to 0$ при $\ell \to \infty$. Такие оценки получены к настоящему времени для многих методов обучения.

К сожалению, известные оценки стабильности столь же сильно завышены, как сложностные, и дают только качественные обоснования соответствующих методов обучения.

2.5 Скользящий контроль

Ещё одно важное направление исследований в SLT связано с использованием *скользящего контроля* (cross-validation), см. стр. 30 и [127, 157].

Несмотря на известную громоздкость, в отдельных случаях техника скользящего контроля непосредственно приводит к простым изящным результатам. В частности, для машин опорных векторов доказано, что значение функционала LOO (leave-one-out cross-validation) не превосходит доли опорных векторов во всей выборке. В силу несмещённости LOO отсюда немедленно вытекает, что вероятность ошибки не превосходит математического ожидания доли опорных векторов [123]. На практике частота ошибок на контроле часто оказывается ещё в несколько раз меньше.

При исследовании локальных методов обучения использование функционала LOO становится естественной «вынужденной мерой» из-за очевидной смещённости эмпирического риска. В частности, для метода одного ближайшего соседа частота ошибок на обучении всегда равна нулю, поскольку каждый объект является ближайшим соседом самого себя. Если каждый обучающий объект исключать из его собственной окрестности, то стандартный функционал эмпирического риска (частоты ошибок на обучении) трансформируется в LOO. Начиная с работ Девроя, Роджерса, Вагнера [191, 124, 125] функционал LOO остаётся одним из основных инструментов исследования устойчивости методов обучения.

На практике скользящий контроль чаще всего применяется либо для выбора одной модели алгоритмов из нескольких (model selection) [151], либо для оптимизации небольшого числа параметров, определяющих структуру алгоритма, таких, как число информативных признаков, степень аппроксимирующего полинома, параметр регуляризации, количество нейронов на скрытом уровне нейронной сети, и т. д. Считается, что настройка значительной доли параметров по скользящему контролю лишена смысла. Когда контрольные данные существенно вовлекаются в процесс обучения, скользящий контроль даёт заниженную оценку обобщающей способности.

Причиной является всё то же переобучение, которое приводит к заниженности эмпирического риска [183]. Известно, что скользящий контроль даёт несмещённую оценку вероятности ошибки в том случае, когда он используется для проверки качества по окончании обучения. Однако до сих пор нет исчерпывающих исследований, показывающих, в какой степени скользящий контроль может использоваться как критерий оптимизации на стадии обучения.

Интуиция подсказывает, что скользящий контроль должен характеризовать обобщающую способность алгоритма лучше, чем частота ошибок на обучении. Тем не менее, этот факт долгое время не удавалось доказать. Попытки предпринимались неоднократно [151, 155, 144], но были получены лишь *«разумные» верхние границы* (sanity-check bounds) для отклонения оценки скользящего контроля от вероятности ошибок алгоритма. Эти границы даже несколько хуже, чем классические VC-оценки для эмпирического риска.

Причина этих неудач анализируется в [106], где вводятся и сравниваются два альтернативных способа формализации понятия обобщающей способности. При первом способе, близком к подходу Вапника и Червоненкиса, оценивается качество *отдельного алгоритма* $a = \mu X$, получаемого в результате обучения. Это приводит к громоздким и сильно завышенным оценкам, зависящим от ёмкости семейства и требующим дополнительных предположений об устойчивости метода обучения [155]. При втором способе оценивается качество *метода обучения* μ. Оказывается, что в этом случае оценка отклонения скользящего контроля от вероятности ошибки алгоритма, обученного на случайной выборке, не зависит от ёмкости семейства, а только от длины обучения и контроля. Данный результат проясняет природу скользящего контроля и показывает, что завышенность предыдущих оценок связана с неудачным выбором самой постановки задачи и функционала качества.

Отсюда вытекают два важных вывода.

Во-первых, адекватность теоретических оценок обобщающей способности существенно зависит от исходной аксиоматики, в частности, от способа формализации понятий обобщающей способности и переобучения.

Во-вторых, скользящий контроль характеризует обобщающую способность метода ничуть не хуже, чем вероятность ошибки. Нет особой необходимости вводить избыточную величину «*вероятность ошибки*», которую, к тому же, невозможно точно измерить в эксперименте.

Эти идеи нашли непосредственно выражение в комбинаторном подходе к обоснованию обучаемых алгоритмов, развиваемом в данной работе.

2.6 Основные выводы

1. Основной задачей теории статистического обучения (SLT) является получение оценок обобщающей способности и создание на их основе статистически обоснованных методов обучения по прецедентам. Развитие SLT на протяжении последних 40 лет было мотивировано, главным образом, стремлением уточнить оценки

теории Вапника-Червоненкиса (VC-теории). Тем не менее, *завышенность оценок* до сих пор остаётся открытой проблемой.

2. *Причины завышенности* VC-оценок к настоящему времени поняты хорошо. Основная причина в том, что это оценки «худшего случая», справедливые для любой задачи (в понятие «задача» входят: вероятностное распределение в пространстве объектов, восстанавливаемая зависимость, конкретная обучающая выборка), и любого метода обучения (в понятие «метод» входят: параметрическое семейство алгоритмов, оптимизируемый функционал качества, численный метод оптимизации). Получить существенно более точные оценки возможно только на пути учёта специфических свойств конкретных задач и методов обучения.
3. Большое разнообразие подходов в SLT связано с неоднозначностью ответов на вопросы: какие именно характеристики задачи и метода обучения наиболее существенны, и в то же время достаточно удобны для практического оценивания и управления качеством алгоритма в процессе его обучения.
4. В классе оценок, зависящих от задачи, наиболее точные результаты даёт подход, основанный на оценивании *радемахеровской сложности*, а также *PAC-байесовский подход*. В задачах классификации оба они приводят к введению различных гладких функций потерь и различных видов регуляризации. Однако некоторые тонкие эффекты, в частности, переобучение бустинга, не находят объяснения даже в рамках этих, наиболее точных, подходов.
5. В настоящее время в SLT не известно ни одного подхода (за исключением предлагаемого в данной работе), который устранял бы все известные причины завышенности VC-оценок и давал бы *точные оценки вероятности переобучения*.

Глава 3

Эмпирический анализ факторов завышенности VC-оценок

В 1.5.6 были выделены пять причин завышенности VC-оценки (1.50). В данной главе описана методика экспериментального количественного измерения факторов завышенности. Цель таких измерений — выяснить, какие из факторов наиболее значимы. Следующим за этим шагом будет формулировка новых постановок задач, направленных на устранение наиболее значимых факторов.

В рамках классической VC-теории произвести необходимые измерения практически невозможно, поскольку в *функционале равномерной сходимости*

$$P_\varepsilon = \mathsf{P}\Big\{ \sup_{a\in A}(P(a) - \nu(a,X)) \geqslant \varepsilon \Big\}$$

вероятности $P(a)$ неизвестны, а супремум берётся по чрезвычайно широкому множеству A. Возникает парадоксальная ситуация: теоретически вероятность любого события может быть оценена в эксперименте по конечному числу наблюдений. Однако в данном случае вероятность P_ε не удаётся оценить как частоту события, поскольку само наступление этого события трудно идентифицировать.

В рамках слабой аксиоматики оценивается *вероятность переобучения*

$$Q_\varepsilon = \mathsf{P}\big\{\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon\big\},$$

для которой справедливы те же *VC-оценки* (стр. 60), но которую легко измерить по небольшому подмножеству разбиений $(X, \bar{X})$. В частности, в *методе Монте-Карло* разбиения выбираются случайным образом (стр. 24).

С методологической точки зрения интересно отметить, что именно стремление выделить и измерить факторы завышенности VC-оценок и привело к отказу от принципа равномерной сходимости и введению слабой вероятностной аксиоматики.

В слабой аксиоматике становится очевидной связь VC-оценок со скользящим контролем. Классические VC-оценки — это верхние оценки вероятности переобучения, получаемые аналитическим путём. Под *скользящим контролем* обычно понимают среднюю частоту ошибок на контроле $\nu(\mu X, \bar{X})$, вычисленную по подмножеству

разбиений. Ничто не мешает вычислить заодно её эмпирическое распределение, а также эмпирическое распределение *переобученности* $\delta_\mu(X, \bar{X}) = \nu(\mu X, \bar{X}) - \nu(\mu X, X)$, которое и является эмпирической оценкой вероятности переобучения $\hat{Q}_\varepsilon$. Таким образом, VC-оценки и скользящий контроль — это два разных способа оценить одну и ту же величину — вероятность переобучения. В данной работе *принцип скользящего контроля* понимается в широком смысле как вычисление числа ошибок на контроле $n(\mu X, \bar{X})$ для некоторого представительного подмножества разбиений $(X, \bar{X})$ и затем некоторое «разумное» агрегирование этих величин.

Зная эмпирическую оценку $\hat{Q}_\varepsilon$, можно задаться вопросом: какое значение должна была бы иметь функция роста Δ, чтобы VC-оценка $Q_\varepsilon \leqslant \Delta\eta(\varepsilon)$ не была завышенной и обращалась в точное равенство? Здесь $\eta(\varepsilon)$ — вероятность большого отклонения частот в двух выборках для одного отдельного алгоритма. Гипотетическое точное значение функции роста $\Delta = Q_\varepsilon/\eta(\varepsilon)$ предлагается называть *эффективным локальным коэффициентом разнообразия* (ЭЛКР). Возможна ещё одна интерпретация ЭЛКР — он показывает, во сколько раз снижается надёжность эмпирических предсказаний в результате обучения (оптимизации параметров) алгоритма по наблюдаемой выборке по сравнению с оценкой надёжности некоторого фиксированного алгоритма, которую даёт *закон больших чисел*.

В параграфе 3.1 понятие ЭЛКР определяется более корректно и предлагается методика его эмпирического измерения.

В параграфе 3.2 описываются эксперименты на реальных задачах классификации и практических методах обучения. Количественное измерение факторов завышенности VC-оценок показывает, что наиболее значимы два фактора — расслоение и связность семейства алгоритмов. На практике ЭЛКР принимает значения порядка 10^0–10^2, тогда как функция роста обычно имеет порядок 10^5–10^{10} и выше.

В параграфе 3.3 описываются эксперименты на модельных семействах алгоритмов. Рассматривается монотонная цепочка алгоритмов — простейшее семейство, обладающее свойствами расслоения и связности. Строятся его естественные аналоги, не обладающие либо свойством расслоения, либо свойством связности. Оказывается, что в обоих случаях вероятность переобучения становится значительной уже при нескольких десятках алгоритмов в семействе. Отсюда следуют два принципиально важных вывода. Во-первых, все реальные семейства с необходимостью расслоены и связны (а если и не связны, то обладают какой-либо иной структурой сходства алгоритмов). Во-вторых, только при совместном учёте обоих свойств возможно получение точных оценок вероятности переобучения.

Поскольку матрица ошибок монотонной цепочки алгоритмов обладает вполне определённой структурой, для неё нетрудно получить точную комбинаторную оценку вероятности переобучения. На этом этапе исследования точные оценки были получены ещё для нескольких «искусственных» семейств простой структуры — единичной окрестности, пары алгоритмов, и некоторых других. Затем эти оценки были обобщены и обнаружен общий механизм их вывода, основанный на порождающих и разрушающих множествах объектов. Эти техники будут описаны в следующей главе.

3.1 Эффективный локальный коэффициент разнообразия

3.1.1 Определение и эмпирическое измерение ЭЛКР

Введём функционал $Q_{\varepsilon,m}(\mu, \mathbb{X})$, выражающий долю разбиений $\mathbb{X} = X \sqcup \bar{X}$, при которых наблюдается переобучение и число ошибок алгоритма на генеральной выборке $\mathbb{X}$ в точности равно m:

$$Q_{\varepsilon,m} = \mathsf{P}\big[\delta_\mu(X, \bar{X}) \geqslant \varepsilon\big]\big[n(\mu X, \mathbb{X}) = m\big].$$

Лемма 3.1. *Вероятность переобучения представима в виде суммы $L+1$ слагаемых:*

$$Q_\varepsilon = Q_{\varepsilon,0} + Q_{\varepsilon,1} + \ldots + Q_{\varepsilon,L}.$$

Доказательство.

$$\sum_{m=0}^{L} Q_{\varepsilon,m} = \mathsf{P}\big[\delta_\mu(X, \bar{X}) \geqslant \varepsilon\big] \underbrace{\sum_{m=0}^{L} \big[n(\mu X, \mathbb{X}) = m\big]}_{1} = Q_\varepsilon.$$

■

Лемма 3.2. *Для любых $\varepsilon \in [0, 1)$ и $m = 0, \ldots, L$ справедлива оценка*

$$Q_{\varepsilon,m} \leqslant \Delta_m(\mu, \mathbb{X}) \cdot H_L^{\ell,m}\left(s_m^-(\varepsilon)\right), \tag{3.1}$$

где $\Delta_m(\mu, \mathbb{X})$ — локальный коэффициент разнообразия m-го слоя, $s_m^-(\varepsilon) = \left\lfloor \frac{\ell}{L}(m - \varepsilon k) \right\rfloor$.

Доказательство мы опускаем, так как оно аналогично Теореме 1.12, стр. 60. В ходе доказательства делается две оценки сверху: первая — переход к функционалу равномерной сходимости, вторая — применение неравенства Буля. В доказательстве Теоремы 1.12 есть и третья оценка, однако в данном случае она обращается в равенство, так как рассматривается только один m-й слой.

Значение $Q_{\varepsilon,m}$ легко может быть измерено эмпирически. Для этого среднее по всем разбиениям $\mathsf{P} \equiv \frac{1}{C_L^\ell} \sum\limits_{(X,\bar{X})}$ заменяется средним $\hat{\mathsf{P}} \equiv \frac{1}{|N|} \sum\limits_{(X,\bar{X})\in N}$ по некоторому подмножеству разбиений N. В частности, в *методе Монте-Карло* предполагается брать случайное множество разбиений.

Введём обозначения:

$$\hat{Q}_\varepsilon = \hat{\mathsf{P}}\big[\delta_\mu(X, \bar{X}) \geqslant \varepsilon\big];$$
$$\hat{Q}_{\varepsilon,m} = \hat{\mathsf{P}}\big[\delta_\mu(X, \bar{X}) \geqslant \varepsilon\big]\big[n(\mu X, \mathbb{X}) = m\big].$$

Определение 3.1. *Эффективным локальным профилем расслоения будем называть последовательность значений*

$$\hat{\Delta}_m(\varepsilon) = \frac{\hat{Q}_{\varepsilon,m}}{H_L^{\ell,m}\left(s_m^-(\varepsilon)\right)}, \quad m = 0, \ldots, L.$$

Значение $\hat{\Delta}_m(\varepsilon)$ показывает, каким приблизительно должен был бы быть локальный коэффициент разнообразия m-го слоя $\Delta_m(\mu, \mathbb{X})$, чтобы оценка (3.1) не была завышенной и обращалась в точное равенство.

Очевидно, локальный коэффициент разнообразия метода μ на выборке $\mathbb{X}$ представляется в виде суммы локальных коэффициентов разнообразия всех слоёв:

$$\Delta_L^\ell(\mu, \mathbb{X}) = \Delta_0(\mu, \mathbb{X}) + \cdots + \Delta_L(\mu, \mathbb{X}).$$

Естественное предположение, что и для эффективных коэффициентов должно быть справедливо аналогичное представление, приводит к следующему определению.

Определение 3.2. *Эффективным локальным коэффициентом разнообразия (ЭЛКР) называется величина*

$$\hat{\Delta}_L^\ell(\varepsilon) = \hat{\Delta}_0(\varepsilon) + \hat{\Delta}_1(\varepsilon) + \cdots + \hat{\Delta}_L(\varepsilon).$$

Оценивание ЭЛКР является обратной задачей по отношению к основной (прямой) задаче оценивания вероятности переобучения. Оценки ЭЛКР невозможно использовать для решения прямой задачи. Цель их введения другая — выделить и сравнить различные факторы завышенности VC-оценок и показать экспериментально, к каким численным значениям коэффициентов разнообразия нужно стремиться в теоретических исследованиях.

Эффективные коэффициенты разнообразия, в отличие от обычных, могут принимать нецелые значения. Кроме того, они зависят от параметра точности ε. Для обоснованного выбора ε сначала задаётся надёжность (уровень значимости) η_0 или диапазон значений надёжности $[\eta_1, \eta_2]$, например, $\eta_0 = 0.05$ и $[\eta_1, \eta_2] = [0.01, 0.1]$. Точность ε и надёжность η связаны невозрастающей зависимостью $\eta(\varepsilon) = Q_\varepsilon \approx \hat{Q}_\varepsilon$. Это позволяет вычислить соответствующее значение точности $\varepsilon = \eta^{-1}(\eta_0)$ или диапазон значений точности $[\varepsilon_1, \varepsilon_2] = [\eta^{-1}(\eta_2), \eta^{-1}(\eta_1)]$, который, в свою очередь, определяет диапазон значений коэффициента разнообразия:

$$\hat{\Delta}_L^\ell \in \Big[\min_{\varepsilon\in[\varepsilon_1,\varepsilon_2]} \hat{\Delta}_L^\ell(\varepsilon), \max_{\varepsilon\in[\varepsilon_1,\varepsilon_2]} \hat{\Delta}_L^\ell(\varepsilon)\Big].$$

Именно это выражение и будет использоваться далее для вычисления интервальных эмпирических оценок ЭЛКР. Дополнительно будет строиться зависимость $\hat{\Delta}_L^\ell(\varepsilon)$ от параметра точности ε.

3.1.2 Эмпирическое измерение факторов завышенности

Обозначим для краткости $H_m(\varepsilon) = H_L^{\ell,m}\left(s_m^-(\varepsilon)\right)$, $\Gamma(\varepsilon) = \max\limits_m H_m(\varepsilon)$.

Запишем ещё раз цепочку оценок из доказательства Теоремы 1.12 и Следствия 1.12.2, стр. 60, используя Леммы 3.1 и 3.2:

$$Q_\varepsilon = \sum_{m=0}^{L} Q_{\varepsilon,m} \leqslant \sum_{m=0}^{L} \Delta_m H_m(\varepsilon) \leqslant \Delta_L^\ell\, \Gamma(\varepsilon) \leqslant \Delta^A(L)\, \Gamma(\varepsilon) \leqslant \Delta^A(L) \cdot \tfrac{3}{2} e^{-\varepsilon^2 \ell}. \qquad (3.2)$$

Степенью завышенности оценки $Q \leqslant Q'$ будем называть число $r = Q'/Q$.

Рассмотрим основные факторы завышенности VC-оценки (3.2) и способы их эмпирического измерения.

1. Пренебрежение эффектом локализации. Локальный коэффициент разнообразия Δ_L^ℓ может оказаться много меньше функции роста $\Delta^A(L)$.

Степень завышенности определяется отношением

$$r_1 = \frac{\Delta^A(L)}{\Delta_L^\ell}.$$

Теоретические оценки функции роста $\Delta^A(L)$ известны для многих семейств алгоритмов. Труднее оценить локальный коэффициент Δ_L^ℓ. Тривиальной оценкой для него служит число разбиений $|N|$. Оно должно быть достаточно большим, чтобы эмпирическая оценка $\hat{Q}_\varepsilon$ была достаточно близка к истинному значению Q_ε, и чтобы множество алгоритмов $\hat{A} = \{a = \mu X \colon (X, \bar{X}) \in N\}$ было достаточно представительным. С другой стороны, число разбиений не должно быть слишком большим, чтобы алгоритмы из $\hat{A}$ имели попарно различные векторы ошибок. На практике оценка $\Delta_L^\ell \approx |N|$ оказывается, как правило, сильно заниженной.

2. Применение принципа равномерной сходимости и неравенства Буля в доказательстве Леммы 3.2 обусловливает завышенность оценки (3.1).

Степень завышенности определяется отношением

$$r_2(\varepsilon) = \frac{\Delta_L^\ell}{\hat{\Delta}_L^\ell(\varepsilon)}.$$

Способ оценивания $\hat{\Delta}_L^\ell(\varepsilon)$ описан в предыдущем параграфе.

3. Оценивание гипергеометрической функции распределения $H_m(\varepsilon) \leqslant \Gamma(\varepsilon)$ приводит к замене профиля расслоения $\{\Delta_m\}_{m=0}^L$ скалярным коэффициентом разнообразия $\Delta_L^\ell = \sum_{m=0}^L \Delta_m$.

Связанную с этим степень завышенности оценим отношением

$$r_3(\varepsilon) = \frac{\sum\limits_{m=0}^{L} \hat{\Delta}_m(\varepsilon)\Gamma(\varepsilon)}{\sum\limits_{m=0}^{L} \hat{\Delta}_m(\varepsilon) H_m(\varepsilon)} = \frac{\hat{\Delta}_L^\ell(\varepsilon)\Gamma(\varepsilon)}{\hat{Q}_\varepsilon}.$$

4. Экспоненциальная аппроксимация функции гипергеометрического распределения $\Gamma(\varepsilon)$. Данный шаг оправдан только стремлением получить оценку в простом аналитическом виде. Однако для вычисления функции гипергеометрического распределения и обратной к ней существуют эффективные численные методы. Если предполагать, что все вычисления выполняются на компьютере, экспоненциальная оценка лишается практической целесообразности.

Степень завышенности определяется отношением

$$r_4(\varepsilon) = \frac{\frac{3}{2}\, e^{-\varepsilon^2 \ell}}{\Gamma(\varepsilon)}.$$

Все введённые отношения больше единицы, и их произведение равно степени завышенности VC-оценки (3.2):

$$r_1 \cdot r_2(\varepsilon) \cdot r_3(\varepsilon) \cdot r_4(\varepsilon) = \frac{\Delta^A(L) \cdot \frac{3}{2}\, e^{-\varepsilon^2 \ell}}{\hat{Q}_\varepsilon}.$$

3.1.3 О понятии эффективной ёмкости по Вапнику

Эффект локализации связан с фиксацией целевой зависимости y, метода обучения μ и выборки $\mathbb{X}$. В работах В. Н. Вапника [216, 107] вводится понятие *эффективной ёмкости* (effective VC-dimension), которая учитывает μ и $\mathbb{X}$, но не учитывает y. Следовательно, отношение эффективной (но не локальной) функции роста к эффективному локальному коэффициенту разнообразия оценивает ту долю степени завышенности r_1, которая связана с игнорированием целевой зависимости y.

Следуя [216], ограничимся задачей классификации с двумя классами, $\mathbb{Y} = \{0, 1\}$, зададим функцию потерь $\mathscr{L}(y, y') = [y \neq y']$ и положим $\ell = k$.

Рассмотрим VC-оценку функционала равномерной сходимости:

$$\widetilde{Q}_\varepsilon = \mathsf{P}\Big[\max_{a \in A} \delta(a, X, \bar{X}) \geqslant \varepsilon\Big] \leqslant \Delta^A(L) \cdot \Gamma(\varepsilon).$$

Эффективная функция роста определяется как значение функции роста $\Delta^A(L)$, при котором эта VC-оценка становится точной, т. е. обращается в равенство:

$$\hat{\Delta}^A_{\text{eff}}(L, \varepsilon) = \frac{\widetilde{Q}_\varepsilon}{\Gamma(\varepsilon)}.$$

Эффективная ёмкость h определяется как величина, связанная с эффективной функцией роста формулой (1.58): $\hat{\Delta}^A_{\text{eff}}(L, \varepsilon) = \frac{3}{2}\frac{L^h}{h!}$.

Для измерения эффективной ёмкости в [216] предлагается оценивать $\hat{\Delta}^A_{\text{eff}}$ при различных L и подбирать такое значение h, при котором зависимость $\hat{\Delta}^A_{\text{eff}}$ от L наиболее точно аппроксимируется функцией $\frac{3}{2}\frac{L^h}{h!}$. Неплохая точность аппроксимации, наблюдаемая в экспериментах, свидетельствует о корректности методики.

Для поиска алгоритма $\tilde{a}_X \in A$, на котором достигается максимум отклонения частот $\delta(a, X, \bar{X})$, предлагается следующий оригинальный приём. Нетрудно видеть, что задача максимизации $\delta(a, X, \bar{X})$ эквивалентна задаче минимизации эмпирического риска $\nu(a, \tilde{\mathbb{X}})$ по модифицированной генеральной выборке $\tilde{\mathbb{X}}$. Модификация заключается в том, что на всех контрольных объектах $x_i \in \bar{X}$ правильный ответ y_i заменяется ошибочным ответом $1 - y_i$:

$$\tilde{a}_X = \arg\max_{a \in A} \delta(a, X, \bar{X}) = \arg\min_{a \in A} \left(\frac{1}{\ell} \sum_{x_i \in X} \big[a(x_i) \neq y_i\big] + \frac{1}{k} \sum_{x_i \in \bar{X}} \big[a(x_i) = y_i\big] \right),$$

Для получения алгоритма $\tilde{a}_X$ достаточно применить тот же метод обучения μ к модифицированной генеральной выборке $\tilde{\mathbb{X}}$. Алгоритм $\tilde{a}_X$ фактически обучается делать ошибки на случайной половине объектов (если выбрано $\ell = k$). Тем самым

устраняется фиксация целевой зависимости y и связанная с этим часть эффекта локализации.

Степень завышенности, связанная с игнорированием целевой зависимости y:

$$r'_1(\varepsilon) = \frac{\hat{\mathsf{P}}\big[\delta(\tilde{a}_X, X, \bar{X}) \geqslant \varepsilon\big]}{\hat{\mathsf{P}}\big[\delta(\mu X, X, \bar{X}) \geqslant \varepsilon\big]} = \frac{\hat{\Delta}^A_{\text{eff}}(L,\varepsilon) \cdot \Gamma(\varepsilon)}{\hat{Q}_\varepsilon} = r_3(\varepsilon)\frac{\hat{\Delta}^A_{\text{eff}}(L,\varepsilon)}{\hat{\Delta}^\ell_L(\varepsilon)}.$$

Приведём ещё две интерпретации коэффициента $r'_1(\varepsilon)$.

1. *Эффективная функция роста* определяется через функционал равномерной сходимости $\widetilde{Q}_\varepsilon$, который сам является заведомо завышенной оценкой. *Эффективный локальный коэффициент разнообразия* определяется через функционал вероятности переобучения Q_ε. Отсюда следует, что $r'_1(\varepsilon)$ — это степень завышенности, возникающая в результате применения принципа равномерной сходимости.

2. Эксперименты с линейным пороговым классификатором, описанные в [216], дали вполне ожидаемый результат: эффективная ёмкость приблизительно равна размерности подпространства, в котором сосредоточены объекты выборки. Коэффициент $r'_1(\varepsilon)$ показывает, во сколько раз завышена эта оценка.

3.2 Эксперименты на реальных данных

Логические алгоритмы классификации, основанные на *индукции правил* (rule induction), особенно удобны для проведения экспериментов по эмпирическому измерению факторов завышенности VC-оценок.

Во-первых, для них известны оценки функции роста.

Во-вторых, они основаны на явном переборе большого количества элементарных классификаторов (называемых также *правилами* или *закономерностями*), что позволяет более аккуратно оценивать локальные коэффициенты разнообразия.

В-третьих, эти алгоритмы широко применяются на практике, поэтому проблема переобученности как самого алгоритма, так и составляющих его правил, представляет значительный практический интерес.

В отличие от описанной выше методики факторы завышенности будут измеряться не для алгоритмов классификации, а для составляющих их закономерностей. Для этого придётся обобщить определения метода обучения μ и вероятности переобучения Q_ε. Эти обобщения носят не принципиальный, а, скорее, технический характер.

3.2.1 Логические алгоритмы классификации

Рассматриваются задачи классификации, $\mathbb{Y}$ — конечное множество классов.

Говорят, что предикат $\varphi\colon \mathbb{X} \to \{0,1\}$ *выделяет* (covers) объект x, если $\varphi(x) = 1$. Предикат φ характеризуется относительно класса $y \in \mathbb{Y}$ и выборки X двумя величинами: числом положительных примеров p_y (выделяемых объектов класса y) и числом

отрицательных примеров n_y (выделяемых объектов других классов):

$$p_y(\varphi, X) = \#\big\{x_i \in X \;\big|\; \varphi(x_i) = 1,\ y_i = y\big\};$$
$$n_y(\varphi, X) = \#\big\{x_i \in X \;\big|\; \varphi(x_i) = 1,\ y_i \neq y\big\};$$

Понятие закономерности. *Закономерностью или правилом* (rule) класса $y \in \mathbb{Y}$ называется предикат $\varphi\colon \mathbb{X} \to \{0, 1\}$, выделяющий достаточно много объектов класса y и достаточно мало объектов всех остальных классов:

$$p_y(\varphi, X) \geqslant p_y^{\min};$$
$$n_y(\varphi, X) \leqslant n_y^{\max};$$

где $p_y^{\min}$ и $n_y^{\max}$ — заданные пороговые константы. Данное определение надо рассматривать, скорее, как неформальное, поскольку не существует строгого критерия, с помощью которого назначались бы значения $p_y^{\min}$ и $n_y^{\max}$.

Понятие информативности. Качество предиката φ относительно класса y удобно характеризовать не парой показателей (p_y, n_y), а одним показателем информативности $I(p_y, n_y)$. К сожалению, однозначно лучшего показателя не существует. В обзорной работе [135] приведено около 20 различных эвристических критериев информативности, представляющих собой разного рода функции от пары величин p_y, n_y. На практике часто используется энтропийный критерий *информационного выигрыша* (information gain), статистические критерии χ^2, ω^2, точный тест Фишера [175], критерий бустинга $\sqrt{p_y} - \sqrt{n_y}$ [197, 122], и другие.

В нашем методе синтеза информативных закономерностей используются два критерия: доля ошибочно выделенных объектов

$$E_y(\varphi) = \frac{n_y}{p_y + n_y}$$

и статистический критерий, называемый *точным тестом Фишера* [175]:

$$I_y(\varphi) = \ln \frac{C_{P_y+N_y}^{p_y+n_y}}{C_{P_y}^{p_y} C_{N_y}^{n_y}},$$

где P_y — число объектов класса y, N_y — число объектов всех остальных классов в обучающей выборке X.

Синтез логических закономерностей. В данной работе используется метод поиска информативных закономерностей в форме конъюнкций, объединяющий в себе три эвристики: усечённый поиск в ширину [62], бустинг закономерностей [122], и точный тест Фишера в роли критерия информативности [175]. Алгоритм реализован Д. Кочедыковым и А. Ивахненко и применяется в системе кредитного скоринга Forecsys ScoringAce® [60, 29, 54, 27]. Здесь приводится его упрощённое описание.

Пусть объекты $x \in \mathbb{X}$ описываются n дискретными признаками $f_j\colon \mathbb{X} \to D_j$, $j = 1, \ldots, n$. Номинальные признаки порождают *элементарные предикаты (термы)* двух видов: $\beta_j(x) = [f_j(x) = c]$ и $\beta_j(x) = [f_j(x) \neq c]$ при всевозможных $c \in D_j$.

Алгоритм 3.2.1. Обучение конъюнкций методом усечённого поиска в ширину.

Вход:

X — обучающая выборка;
$y \in \mathbb{Y}$ — класс, для которого строится список конъюнкций;
K — максимальный ранг конъюнкций;
T_1 — число лучших конъюнкций, отбираемых на каждом шаге;
T_0 — число лучших конъюнкций, отбираемых на последнем шаге, $T_0 \leqslant T_1$;
$I_{\min}$ — порог информативности;
$E_{\max}$ — порог допустимой доли ошибок;

Выход:

список конъюнкций $R_y = \{\varphi_y^t(x) | \, t = 1, \ldots, T_y\}$;

$R_y := \varnothing$;
для всех $\beta \in \mathscr{B}_j$, $j = 1, \ldots, n$
  `Добавить_в_список` (R_y, β);
для всех $k = 2, \ldots, K$
  для всех конъюнкций $\varphi \in R_y$ ранга $(k-1)$
    для всех $\beta \in \mathscr{B}_j$, $j = 1, \ldots, n$
      если признака f_j нет в конъюнкции φ и $I_y(\varphi \wedge \beta) \geqslant I_{\min}$ **то**
        `Добавить_в_список` $(R_y, \varphi \wedge \beta)$;
оставить в R_y не более T_0 конъюнкций с наибольшими $I_y(\varphi)$ и $E_y(\varphi) \leqslant E_{\max}$;

ПРОЦЕДУРА `Добавить_в_список` (R_y, φ);
если $|R_y| < T_1$ **то**
  $R_y := R_y \cup \{\varphi\}$
иначе
  найти худшую конъюнкцию в списке: $\psi := \arg\min\limits_{\psi \in R_y} I_y(\psi)$;
  если $I_y(\varphi) > I_y(\psi)$ **то**
    заменить в списке R_y худшую конъюнкцию ψ на φ;

Порядковые признаки, в дополнение к этим двум, порождают ещё два вида термов: $\beta_j(x) = [f_j(x) \leqslant c]$ и $\beta_j(x) = [f_j(x) \geqslant c]$, $c \in D_j$. Обозначим через $\mathscr{B}_j$ множество всех термов, порождаемых признаком f_j. Поиск закономерностей производится среди конъюнкций ранга не выше K, составленных из термов различных признаков:

$$\Phi[K] = \Big\{ \varphi(x) = \bigwedge_{j \in J} \beta_j(x) \;\Big|\; \beta_j \in \mathscr{B}_j,\ J \subseteq \{1, \ldots, n\},\ |J| \leqslant K \Big\}.$$

Алгоритм 3.2.1 начинает поиск закономерностей с построения конъюнкций ранга 1. Для этого отбираются не более T_1 самых информативных термов. На всех последующих шагах к каждой из имеющихся конъюнкций добавляется один терм всеми возможными способами. Получается расширенное множество конъюнкций, из которых снова отбираются T_1 самых информативных. Наращивание конъюнкций прекращается либо при достижении максимального ранга K, либо когда ни одну из конъ-

юнкций не удаётся улучшить путём добавления терма. Лучшие конъюнкции, собранные со всех шагов, заносятся в списки R_y. Параметр T_1 позволяет управлять *шириной поиска* и находить компромисс между качеством и скоростью работы алгоритма.

После выполнения Алгоритма 3.2.1 в выборке могут остаться объекты, не выделенные ни одной закономерностью из списков R_y, либо ошибочно выделенные закономерностями «чужих» классов. Этим объектам назначаются бóльшие веса согласно формулам бустинга [122] и Алгоритм 3.2.1 запускается заново. Веса объектов учитываются при вычислении критерия $I_y(\varphi)$, что позволяет находить новые закономерности, существенно отличающиеся от найденных на предыдущих итерациях.

Алгоритм классификации строится как линейная композиция закономерностей:

$$a(x) = \arg\max_{y\in\mathbb{Y}} \sum_{\varphi_y^t\in R_y} w_y^t \varphi_y^t(x),$$

где $\varphi_y^t(x)$ — закономерности класса y, w_y^t — веса закономерностей. В таком виде представимы многие логические алгоритмы, в частности, взвешенное голосование (weighted voting) правил [122], решающие списки (decision list) [190], решающие деревья (decision tree) [186], машины покрывающих множеств (set covering machine) [174].

3.2.2 Измерение факторов завышенности VC-оценок для случая логических закономерностей

Эмпирическое измерение факторов завышенности VC-оценок удобнее производить не для алгоритмов классификации, а для составляющих их закономерностей. Для этого придётся немного изменить основные определения. Модификации носят технический характер. Формулировки основных теорем остаются практически теми же и здесь даже не приводятся.

Методом обучения закономерностей класса y называется отображение μ_y, которое по обучающей выборке X строит набор R_y закономерностей класса y:

$$\mu_y X = R_y \equiv \big\{\varphi_y^t(x) \;\big|\; t = 1,\ldots,T_y\big\}.$$

Введём индикатор ошибки предиката $\varphi\colon \mathbb{X}\to\{0,1\}$ относительно класса y:

$$I_y(\varphi, x_i) = \big[\varphi(x_i) \neq [y_i = y]\big].$$

Тогда *частота ошибок* предиката φ на выборке X относительно класса y определяется стандартным образом как

$$\nu_y(\varphi, X) = \frac{1}{\ell}\sum_{x_i\in X}^{\ell} I_y(\varphi, x_i) = \frac{1}{\ell}\big(n_y(\varphi, X) + P_y(X) - p_y(\varphi, X)\big), \tag{3.3}$$

где $P_y(X)$ — число объектов класса y в выборке X.

В случае $\varphi(x_i) = 1$ и $y_i \neq y$ закономерность ошибочно выделяет объект чужого класса. В случае $\varphi(x_i) = 0$ и $y_i = y$ закономерность ошибочно не выделяет

объект своего класса. Ошибки второго типа менее существенны в логических алгоритмах, так как пропущенный объект может быть выделен другими закономерностями. Тем не менее, мы будем учитывать ошибки обоих типов с одинаковым весом. Возможно, имело бы смысл назначить им различные веса (что привело бы к небинарной функции потерь), либо исследовать частоты ошибок первого и второго типов по-отдельности. В данной работе эти постановки задачи не рассматриваются.

Переобученностью закономерности $\varphi \in \mu_y X$ при заданной контрольной выборке $\bar{X}$ называется разность частот её ошибок на контроле и на обучении:

$$\delta_y(\varphi, X, \bar{X}) = \nu_y(\varphi, \bar{X}) - \nu_y(\varphi, X).$$

Вероятность переобучения $Q_\varepsilon(\mu_y, \mathbb{X})$ определим как долю переобученных закономерностей среди всех закономерностей класса y, построенных методом μ_y по всевозможным подвыборкам $X \subset \mathbb{X}$:

$$Q_\varepsilon(\mu_y, \mathbb{X}) = \mathsf{P}\,\frac{1}{|\mu_y X|}\sum_{\varphi\in\mu_y X}\big[\delta_y(\varphi, X, \bar{X}) \geqslant \varepsilon\big].$$

Коэффициент разнообразия $\Delta(\Phi, \mathbb{X})$ множества предикатов Φ относительно класса y — это число различных векторов ошибок $\vec{\varphi} = \big(I_y(\varphi, x_i)\big)_{i=1}^L$, порождаемых всевозможными $\varphi \in \Phi$ на выборке $\mathbb{X}$.

Локальный коэффициент разнообразия метода μ_y — это коэффициент разнообразия $\Delta_L^\ell = \Delta(\Phi_L^\ell, \mathbb{X})$ множества закономерностей, получаемых методом μ_y по всевозможным обучающим подвыборкам: $\Phi_L^\ell = \bigcup\limits_{X\in[\mathbb{X}]^\ell} \mu_y X$.

Локальный профиль расслоения метода μ_y — это последовательность коэффициентов разнообразия $\Delta_m = \Delta(\Phi_m, \mathbb{X})$, $m = 0, \ldots, L$, где Φ_m есть m-й слой множества закономерностей Φ_L^ℓ относительно класса y:

$$\Phi_m = \big\{\varphi \in \Phi_L^\ell \;\big|\; \nu_y(\varphi, \mathbb{X}) = \tfrac{m}{L}\big\},\ m = 0, \ldots, L.$$

Таким образом, множество закономерностей Φ_L^ℓ разбивается на $L+1$ подмножеств — слоёв Φ_m, состоящих из закономерностей с фиксированным числом ошибок m на генеральной выборке $\mathbb{X}$. Очевидно, что $\Delta_L^\ell = \Delta_0 + \cdots + \Delta_L$.

Наряду с функционалом Q_ε определим функционал $Q_{\varepsilon,m}$ как долю переобученных закономерностей, допускающих m ошибок на $\mathbb{X}$:

$$Q_{\varepsilon,m}(\mu_y, \mathbb{X}) = \mathsf{P}\,\frac{1}{|\mu_y X|}\sum_{\varphi\in\mu_y X}\big[\delta_y(\varphi, X, \bar{X}) \geqslant \varepsilon\big]\big[\nu_y(\varphi, \mathbb{X}) = \tfrac{m}{L}\big].$$

В этих обозначениях VC-оценка (3.2) и методика эмпирического измерения факторов её завышенности могут быть переписаны практически без изменений. Модификация коснулась, главным образом, определения локального множества алгоритмов A_L^ℓ — теперь его роль играет локальное множество закономерностей Φ_L^ℓ. Смысл модификации предельно прост: надо учитывать все закономерности φ_y^t, построенные при всех разбиениях $(X, \bar{X})$. Кроме того, теперь все величины Q_ε, $Q_{\varepsilon,m}$, $\hat{\Delta}_L^\ell$, $\hat{\Delta}_m$ определяются относительно каждого класса $y \in \mathbb{Y}$ в отдельности, и, вообще говоря, могут сильно отличаться для разных классов.

Функция роста $\Delta^{\Phi[K]}(L)$ множества $\Phi[K]$ не превосходит его мощности. Пусть j-й признак порождает $d_j = |\mathscr{B}_j|$ термов, $j = 1, \ldots, n$. Тогда число конъюнкций ранга r, построенных из признаков подмножества $J = \{1, \ldots, j\}$, не превосходит

$$H_{r,j} = \sum_{\substack{J' \subseteq J \\ |J'|=r}} \prod_{j \in J'} d_j.$$

Непосредственное вычисление по этой формуле не эффективно. Однако возможно быстрое вычисление чисел $H_{r,j}$ за $O(Kn)$ операций, если воспользоваться рекуррентными соотношениями:

$$\begin{aligned} H_{0,j} &= 1; \\ H_{r,j} &= 0, \quad r > j; \\ H_{r,j+1} &= H_{r,j} + d_j H_{r-1,j}, \quad j = 1, \ldots, n, \quad r = 1, \ldots, K. \end{aligned}$$

Функция роста не превосходит общего числа конъюнкций, построенных из всех признаков $1, \ldots, n$, и имеющих ранги $1, \ldots, K$:

$$\Delta^{\Phi[K]}(L) \leqslant H_{1,n} + \cdots + H_{K,n}.$$

Локальный коэффициент разнообразия Δ_L^ℓ оценивается суммарным числом конъюнкций, попавших в списки R_y по всем обучающим выборкам X из отобранного множества разбиений N:

$$\underline{\Delta}_L^\ell = \sum_{(X,\bar{X}) \in N} |\mu_y X| \leqslant |N| \cdot T_0.$$

Данная оценка может оказаться сильно заниженной, поскольку $|N| \ll C_L^\ell$. Более адекватной оценкой является число $\underline{\Delta}_L^\ell$ всех конъюнкций, удовлетворяющих критериям высокой информативности $I_y(\varphi) \geqslant I_{\min}$ и низкой доли ошибок $E_y(\varphi) \leqslant E_{\max}$, которые были синтезированы и оценены, но не попали в списки R_y. Можно полагать, что эти конъюнкции могли бы попасть в списки при других обучающих подвыборках X, не вошедших в N. Число этих конъюнкций легко подсчитать в ходе перебора.

Подсчитаем также число $\bar{\Delta}_L^\ell$ всех конъюнкций φ, для которых в ходе перебора оценивались характеристики p_y и n_y. Тривиальная и несколько завышенная оценка

$$\bar{\Delta}_L^\ell \leqslant |N|(T_1 K - T_1 + 1)(d_1 + \cdots + d_n).$$

Для Алгоритма 3.2.1 возможно также оценить степень завышенности, связанную с игнорированием целевой зависимости y. Это отношение числа всех проанализированных конъюнкций к числу конъюнкций, оказавшихся закономерностями:

$$r_1''(\varepsilon) = \frac{\bar{\Delta}_L^\ell}{\underline{\Delta}_L^\ell}.$$

Поскольку Алгоритм 3.2.1 ведёт направленный поиск наиболее информативных конъюнкций, это отношение может оказаться несколько заниженным.

3.2.3 Эксперименты и выводы

Условия эксперимента. Измерения факторов завышенности VC-оценки проводилось на 7 реальных задачах классификации из репозитория UCI [94]. Число классов во всех задачах равнялось двум. Выборка разбивалась 20 раз случайным образом на две равные части, $\ell = k$, со стратификацией классов. При каждом разбиении сначала первая половина выборки была обучающей, вторая — контрольной, затем они менялись ролями. Таким образом, $|N| = 40$.

В Таблице 3.1 показаны характеристики задач и средний процент ошибок на контрольных данных. Данные по алгоритмам C4.5, C5.0, RIPPER и SLIPPER взяты из работ [121, 122], где скользящий контроль производился аналогичным образом. Численные результаты показывают, что качество нашего алгоритма сопоставимо с аналогами, и он вполне подходит для решения практических задач (доказательство его превосходства не является целью данного эксперимента).

Результаты эксперимента. В Таблице 3.2 показаны оценки коэффициентов разнообразия, вычисленные в процессе работы Алгоритма 3.2.1. В двух правых столбцах приведены оценки ЭЛКР (эффективного локального коэффициента разнообразия), вычисленные согласно 3.1.1.

Рис. 3.1 иллюстрирует вычисление интервальной оценки ЭЛКР по заданному диапазону значений надёжности. На каждом графике показана зависимость ЭЛКР $\hat{\Delta}_L^\ell$ от точности ε. Убывающая кривая показывает зависимость надёжности $\hat{Q}_\varepsilon$ от ε. Интервал возможных значений $\hat{\Delta}_L^\ell(\varepsilon)$ определяется за три шага:

1) на правой вертикальной оси графика откладывается интервал значений надёжности, в данном случае $\hat{Q}_\varepsilon \in [\eta_1, \eta_2] = [0.01, 0.1]$;

2) по кривой $\eta(\varepsilon) = \hat{Q}_\varepsilon$ этот интервал переводится в интервал значений точности $[\varepsilon_1, \varepsilon_2] = [\eta^{-1}(\eta_2), \eta^{-1}(\eta_1)]$, откладываемый по горизонтальной оси;

3) определяется минимальное и максимальное значение ЭЛКР $\hat{\Delta}_L^\ell(\varepsilon)$ на интервале точности $[\varepsilon_1, \varepsilon_2]$; оно откладывается по левой вертикальной оси графика.

Таблица 3.3 является итоговой. В ней сведены все факторы завышенности, вычисленные при фиксированном значении надёжности $\hat{Q}_\varepsilon = 0.05$.

Интерпретации и выводы. Численные значения ЭЛКР имеют порядок 10^0–10^2, тогда как функция роста имеет порядки 10^5–10^9. Это означает, что «эффективное количество алгоритмов», реально задействованных при решении каждой конкретной задачи, на много порядков меньше числа алгоритмов во всём семействе. Ни один из известных на сегодня подходов, включая наиболее точные [164, 185], не даёт оценок коэффициентов разнообразия таких порядков.

Эффективный локальный коэффициент разнообразия во всех задачах не превосходит длины выборки L. Попытка использовать эту величину для определения *эффективной локальной ёмкости* по формуле (1.58) приводит к вырожденному результату: на практике эффективная локальная ёмкость не превышает единицу.

Задача	L	n	$d_1 \cdots d_n$	C4.5	C5.0	RIPPER	SLIPPER	Forecsys
crx	690	15	$2^4\,3^2\,4^1\,9^1\,14^1\,20^6$	15.5	14.0	15.2	15.7	14.3 ± 0.2
german	1000	20	$2^2\,3^3\,4^3\,5^5\,10^1\,11^1\,20^5$	27.0	28.3	28.7	27.2	28.5 ± 1.0
hepatitis	155	19	$2^{13}\,6^4\,8^1\,9^1$	18.8	20.1	23.2	17.4	16.7 ± 1.7
horse-colic	300	25	$2^3\,3^2\,4^6\,5^5\,6^2\,20^7$	16.0	15.3	16.3	15.0	16.4 ± 0.5
hypothyroid	3163	25	$2^{18}\,20^7$	0.4	0.4	0.9	0.7	0.8 ± 0.04
liver	345	6	$12^1\,20^5$	37.5	31.9	31.3	32.2	29.2 ± 1.6
promoters	106	57	57^4	18.1	22.7	19.0	18.9	12.0 ± 2.0

Таблица 3.1. Характеристики задач: длина выборки L; число признаков n; число порождаемых термов d_j, где запись 20^5 означает, что имеется 5 признаков, порождающих по 20 термов; процент ошибок на контроле для четырёх стандартных алгоритмов из [121, 122]; процент ошибок на контроле для Алгоритма 3.2.1.

Задача	T_1	K	y	$\lvert\Phi[K]\rvert$	$\frac{1}{\lvert N\rvert}\bar{\Delta}_L^\ell$	$\frac{1}{\lvert N\rvert}\underline{\Delta}_L^\ell$	$\frac{1}{\lvert N\rvert}\underline{\underline{\Delta}}_L^\ell$	$\hat{\Delta}_L^\ell[\varepsilon_1,\varepsilon_2]$	$\hat{\Delta}_L^\ell(\varepsilon_0)$
crx	50	4	0	$1.4 \cdot 10^7$	$2.1 \cdot 10^4$	380	5	$[10; 41]$	24
			1			490	6	$[11; 180]$	12
german	50	5	1	$5.2 \cdot 10^8$	$3.0 \cdot 10^4$	1370	14	$[38; 530]$	54
			2			330	3	$[1.0; 2.2]$	1.9
hepatitis	50	4	0	$5.6 \cdot 10^5$	$0.9 \cdot 10^4$	570	7	$[11; 148]$	83
			1			240	3	$[12; 27]$	15
horse-colic	50	5	1	$1.9 \cdot 10^6$	$3.8 \cdot 10^4$	630	7	$[2; 9]$	7
			2			330	3	$[3; 6]$	6
hypothyroid	100	5	0	$5.3 \cdot 10^8$	$6.3 \cdot 10^4$	210	7	$[3; 220]$	21
			1			80	3	$[2; 44]$	30
liver	50	4	0	$1.9 \cdot 10^6$	$1.1 \cdot 10^4$	700	7	$[4; 21]$	12
			1			650	7	$[3; 12]$	5
promoters	50	3	0	$1.0 \cdot 10^8$	$2.2 \cdot 10^4$	480	5	$[36; 230]$	72
			1			300	3	$[9; 22]$	18

Таблица 3.2. Параметры метода: ширина поиска T_1; максимальный ранг конъюнкций K; номер класса в кодировке UCI. Оценки коэффициентов разнообразия: функция роста $\lvert\Phi[K]\rvert$; среднее число проанализированных конъюнкций $\frac{1}{\lvert N\rvert}\bar{\Delta}_L^\ell$; среднее число информативных конъюнкций $\frac{1}{\lvert N\rvert}\underline{\Delta}_L^\ell$; среднее число конъюнкций, отобранных в списки $\frac{1}{\lvert N\rvert}\underline{\underline{\Delta}}_L^\ell$; эффективный локальный коэффициент разнообразия $\hat{\Delta}_L^\ell(\varepsilon)$, соответствующий диапазону $\hat{Q}_\varepsilon \in [0.01, 0.1]$ и значению $\hat{Q}_\varepsilon = 0.05$.

Среди четырёх факторов завышенности первые два наиболее значимы:

1) r_1 — пренебрежение эффектом локализации;

2) r_2 — применение принципа равномерной сходимости и неравенства Буля.

На устранение первого фактора направлено большинство современных работ, посвящённых *оценкам, зависящим от задачи* (data-dependent bounds) [159, 164, 103, 108, 185]. Однако все эти оценки по-прежнему имеют сомножитель, описывающий «сложность» некоторого множества алгоритмов, пусть даже и локального. Большие значения r_2 говорят о том, что «проклятие завышенности» присуще самой струк-

Задача	y	r_1	$r'_1(\varepsilon)$	$r''_1(\varepsilon)$	$r_2(\varepsilon)$	$r_3(\varepsilon)$	$r_4(\varepsilon)$
crx	0	890	20	55	680	3.1	32.6
	1	690	21	43	1700	1.6	11.6
german	1	8 950	18	22	1500	1.7	10.9
	2	37 000	22	92	9000	1.2	9.9
hepatitis	0	23	20	16	280	13.4	9.5
	1	55	20	37	680	2.4	22.5
horse-colic	1	72	19	60	4500	2.1	7.2
	2	140	20	115	3400	3.6	7.3
hypothyroid	0	61 000	21	310	400	32.2	16.5
	1	153 000	15	770	460	3.8	28.7
promoters	0	94	16	46	340	5.9	9.8
	1	150	23	73	790	3.4	6.9

Таблица 3.3. Факторы завышенности VC-оценок при значении точности ε, соответствующей надёжности $\hat{Q}_\varepsilon = 0.05$.

туре «сложностных» оценок. Коэффициенты разнообразия не учитывают степень сходства и неравномерность распределения алгоритмов.

Коэффициенты r'_1 и r''_1 оценивают влияние целевой зависимости y на степень завышенности r_1. Оба они занижены, поэтому можно утверждать только, что соответствующая потеря точности составляет два порядка или более. Введённое В. Н. Вапником понятие *эффективной ёмкости* не учитывает этот фактор, поскольку основывается на принципе равномерной сходимости.

Фактор r_3 в большинстве случаев сравнительно мал. При значениях числа ошибок $m = L\nu_y(\varphi, \mathbb{X})$, характерных для закономерностей, значения функции $H(m) = H_L^{\ell,m}(s_m^-(\varepsilon))$ не сильно отличаются от максимума, см. рис. 3.2. Однако при $m \to 0$ функция $H(m)$ стремится к нулю быстрее геометрической прогрессии. Поэтому для обычных алгоритмов классификации и «хороших» задач с частотой ошибок (ориентировочно) менее 10% фактор r_3 может оказаться значительным.

Фактор r_4 показывает, что экспоненциальная аппроксимация гипергеометрического распределения неточна, и на практике от неё следует отказаться.

3.3 Эксперименты на модельных данных

Эмпирический анализ факторов завышенности VC-оценки, проведённый в предыдущем параграфе, показал, что наиболее существенны два фактора. Пренебрежение эффектом расслоения завышает оценку в 10^3–10^5 раз. Пренебрежение эффектом сходства алгоритмов завышает оценку в 10^3–10^4 раз. Остальные факторы совместно дают завышенность в 10^1–10^2 раз и относительно легко устраняются.

Ниже описаны ещё два эксперимента на модельных семействах алгоритмов.

В 3.3.1 рассматривается двухэлементное семейство, и для него выводится точная комбинаторная оценка вероятности переобучения. Оказывается, что явление переобучения возникает даже в этом простейшем случае, причём эффекты расслоения и сходства снижают вероятность переобучения.

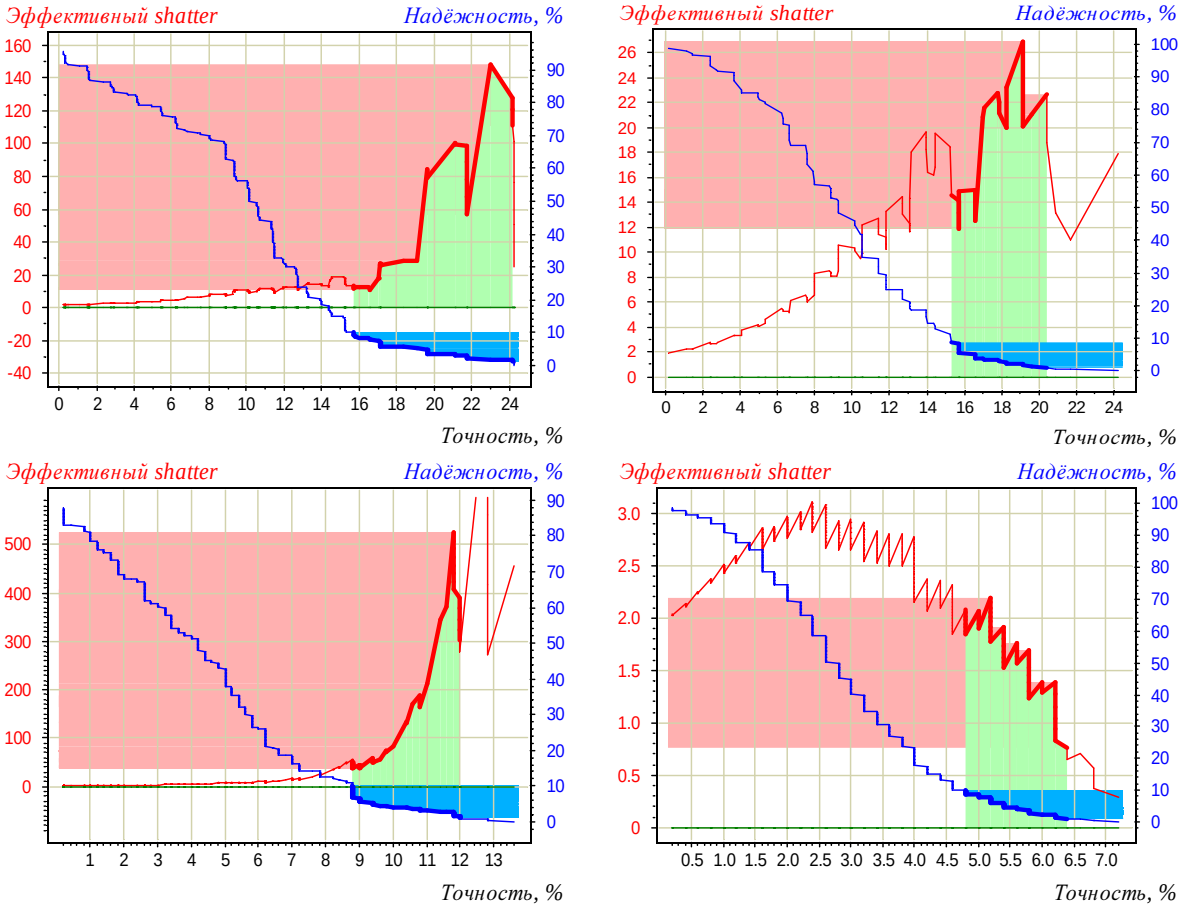


Рис. 3.1. Зависимость эффективного локального коэффициента разнообразия $\hat{\Delta}_L^\ell$ и надёжности $\hat{Q}_\varepsilon$ от точности ε для задач hepatitis (вверху, $y = 0, 1$), german (внизу, $y = 1, 2$). Полосы показывают определение диапазона возможных значений $\hat{\Delta}_L^\ell$ по заданному диапазону надёжности $\hat{Q}_\varepsilon \in [0.01, 0.1]$.

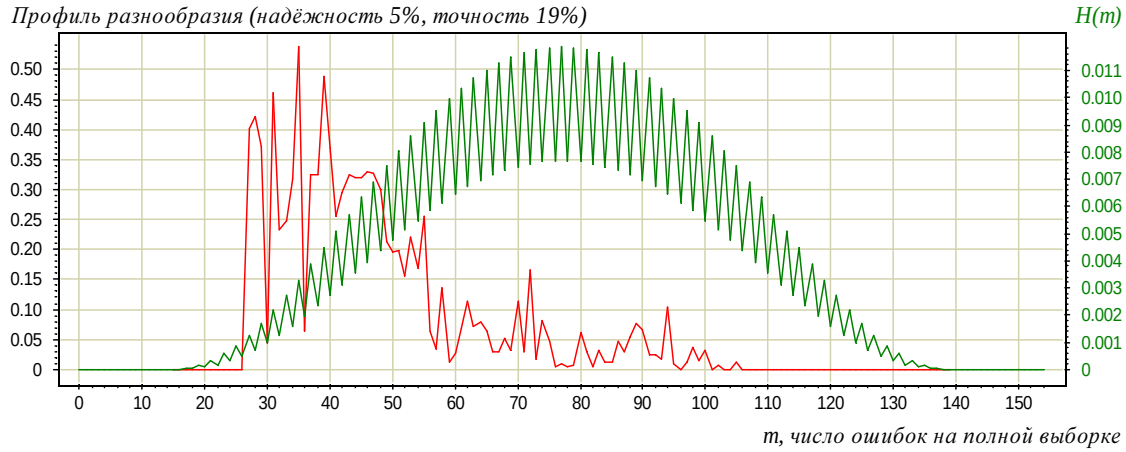


Рис. 3.2. Зависимость эффективного локального профиля расслоения $\hat{D}_m(\varepsilon)$ и функции $H(m)$ от числа ошибок $m = L\nu_y(\varphi, \mathbb{X})$, для задачи hepatitis, $y = 0$.

В 3.3.2 рассматривается монотонная цепочка алгоритмов — простейшее семейство, обладающее свойствами расслоения и связности. Эксперименты показывают, что если одно из этих двух свойств отсутствует, то переобучение уже неприемлемо велико. Точные оценки вероятности переобучения для монотонной цепочки выводятся в следующей главе.

3.3.1 Семейство из двух алгоритмов

Точная оценка вероятности переобучения. Рассмотрим семейство из двух алгоритмов $A = \{a_1, a_2\}$ и метод *минимизации эмпирического риска*:

$$\mu X = \arg\min_{a \in A} \nu(a, X).$$

Для определённости договоримся, что в случае неоднозначности, когда минимум частоты ошибок достигается на обоих алгоритмах, $\nu(a_1, X) = \nu(a_2, X)$, метод μ будет выбирать алгоритм с бóльшим числом ошибок на полной выборке.

Теорема 3.3. *Пусть в выборке $\mathbb{X}$ имеется m_0 объектов, на которых оба алгоритма допускают ошибку; m_1 объектов, на которых только a_1 допускает ошибку; m_2 объектов, на которых только a_2 допускает ошибку; m_3 остальных объектов (на которых оба алгоритма не допускают ошибку), и для определённости $m_1 \leqslant m_2$:*

$$a_1 = (1, \ldots, 1,\ 1, \ldots, 1,\ 0, \ldots, 0,\ 0, \ldots, 0);$$
$$a_2 = (\underbrace{1, \ldots, 1}_{m_0},\ \underbrace{0, \ldots, 0}_{m_1},\ \underbrace{1, \ldots, 1}_{m_2},\ \underbrace{0, \ldots, 0}_{m_3}).$$

Тогда для любого $\varepsilon \in [0, 1)$ вероятность переобучения есть

$$Q_\varepsilon = \sum_{\substack{s_0=0 \\ s_0+s_1+s_2+s_3=\ell}}^{m_0} \sum_{s_1=0}^{m_1} \sum_{s_2=0}^{m_2} \sum_{s_3=0}^{m_3} \frac{C_{m_0}^{s_0} C_{m_1}^{s_1} C_{m_2}^{s_2} C_{m_3}^{s_3}}{C_L^\ell} \times \Big(\big[s_1 < s_2\big] \big[s_0 + s_1 \leqslant \tfrac{\ell}{L}(m_0 + m_1 - \varepsilon k)\big] + \\ + \big[s_1 \geqslant s_2\big] \big[s_0 + s_2 \leqslant \tfrac{\ell}{L}(m_0 + m_2 - \varepsilon k)\big] \Big).$$

Доказательство. Метод минимизации эмпирического риска выбирает алгоритм a_1 при $\nu(a_1, X) < \nu(a_2, X)$ и алгоритм a_2 в противном случае. Следовательно,

$$Q_\varepsilon = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \big[\nu(a_1, X) < \nu(a_2, X)\big] \big[\nu(a_1, \bar{X}) - \nu(a_1, X) \geqslant \varepsilon\big] + \\ + \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \big[\nu(a_1, X) \geqslant \nu(a_2, X)\big] \big[\nu(a_2, \bar{X}) - \nu(a_2, X) \geqslant \varepsilon\big].$$

Разобьём множество $\mathbb{X}$ на 4 подмножества: X_0 — объекты, на которых ошибаются оба алгоритма; X_1 — объекты, на которых ошибается только a_1; X_2 — объекты, на которых ошибается только a_2; X_3 — все остальные объекты. Очевидно, $m_i = |X_i|$. Положим $s_i = |X_i \cap X|$ — число объектов из X_i, попавших в обучающую выборку X.

В этих обозначениях частоты ошибок алгоритмов a_1, a_2 на выборках X, $\bar{X}$ есть

$$\nu(a_1, X) = \tfrac{s_0+s_1}{\ell}; \qquad \nu(a_1, \bar{X}) = \tfrac{m_0+m_1-s_0-s_1}{k};$$
$$\nu(a_2, X) = \tfrac{s_0+s_2}{\ell}; \qquad \nu(a_2, \bar{X}) = \tfrac{m_0+m_2-s_0-s_2}{k}.$$

Число разбиений, при которых реализуется набор значений (s_0, s_1, s_2, s_3), есть

$$\sum_{X\in[\mathbb{X}]^\ell} \prod_{i=0}^{3} \big[s_i = |X_i \cap X|\big] = C_{m_0}^{s_0} C_{m_1}^{s_1} C_{m_2}^{s_2} C_{m_3}^{s_3}. \tag{3.4}$$

Отсюда следует, что s_0, s_1, s_2, s_3 должны удовлетворять ограничениям

$$0 \leqslant s_0 \leqslant m_0; \quad 0 \leqslant s_1 \leqslant m_1; \quad 0 \leqslant s_2 \leqslant m_2; \quad 0 \leqslant s_3 \leqslant m_3.$$

Кроме того, s_0, s_1, s_2, s_3 должны удовлетворять соотношению $s_0+s_1+s_2+s_3 = \ell$. Таким образом,

$$Q_\varepsilon = \frac{1}{C_L^\ell} \sum_{X\in[\mathbb{X}]^\ell} \sum_{s_0=0}^{m_0} \sum_{s_1=0}^{m_1} \sum_{s_2=0}^{m_2} \sum_{s_3=0}^{m_3} \big[s_0+s_1+s_2+s_3 = \ell\big] \prod_{i=0}^{3} \big[s_i = |X_i \cap X|\big] \times$$
$$\times \Big(\big[s_1 < s_2\big] \big[\tfrac{m_0+m_1-s_0-s_1}{k} - \tfrac{s_0+s_1}{\ell} \geqslant \varepsilon\big] +$$
$$\big[s_1 \geqslant s_2\big] \big[\tfrac{m_0+m_2-s_0-s_2}{k} - \tfrac{s_0+s_2}{\ell} \geqslant \varepsilon\big] \Big).$$

Переставляя знаки суммирования и подставляя (3.4), получаем требуемое. ■

Вычисление ЭЛКР. В отличие от 3.1.1, определим *эффективный локальный коэффициент разнообразия* (ЭЛКР) как значение коэффициента разнообразия Δ_L^ℓ, при котором оценка (1.50) не является завышенной. Сопоставляя (1.48) и (1.50), получим двустороннюю оценку для ЭЛКР:

$$\frac{\mathsf{P}\big[\delta_\mu(X) \geqslant \varepsilon\big]}{\max\limits_{a\in A} \mathsf{P}\big[\delta(a, X) \geqslant \varepsilon\big]} = \underline{\Delta}_L^\ell \leqslant \hat{\Delta}_L^\ell \leqslant \bar{\Delta}_L^\ell = \frac{\mathsf{P}\big[\delta_\mu(X) \geqslant \varepsilon\big]}{\min\limits_{a\in A} \mathsf{P}\big[\delta(a, X) \geqslant \varepsilon\big]}.$$

Верхняя оценка ЭЛКР $\bar{\Delta}_L^\ell$ имеет естественную содержательную интерпретацию. Она показывает, во сколько раз вероятность переобучения метода μ превышает вероятность большого отклонения частот для наилучшего алгоритма в семействе.

Чтобы знаменатель в верхней оценке ЭЛКР $\bar{\Delta}_L^\ell$ не обращался в нуль, минимум берётся только по тем алгоритмам $a \in A$, для которых $n(a, \mathbb{X}) \geqslant \varepsilon k$.

Очевидно, в случае двухэлементного семейства алгоритмов $1 \leqslant \bar{\Delta}_L^\ell \leqslant 2$.

Вычислительный эксперимент. На рис. 3.3, 3.4 показана зависимость вероятности переобучения Q_ε и верхней оценки ЭЛКР $\bar{\Delta}_L^\ell$ от различности алгоритмов при $\ell = k = 100$, $\varepsilon = 0.05$. В качестве естественной меры различности взято хэммингово расстояние между векторами ошибок $\rho(a_1, a_2) = m_1+m_2$. Тонкими линиями показаны оценки ЭЛКР, вычисленные *методом Монте-Карло* по 1000 случайных разбиений.

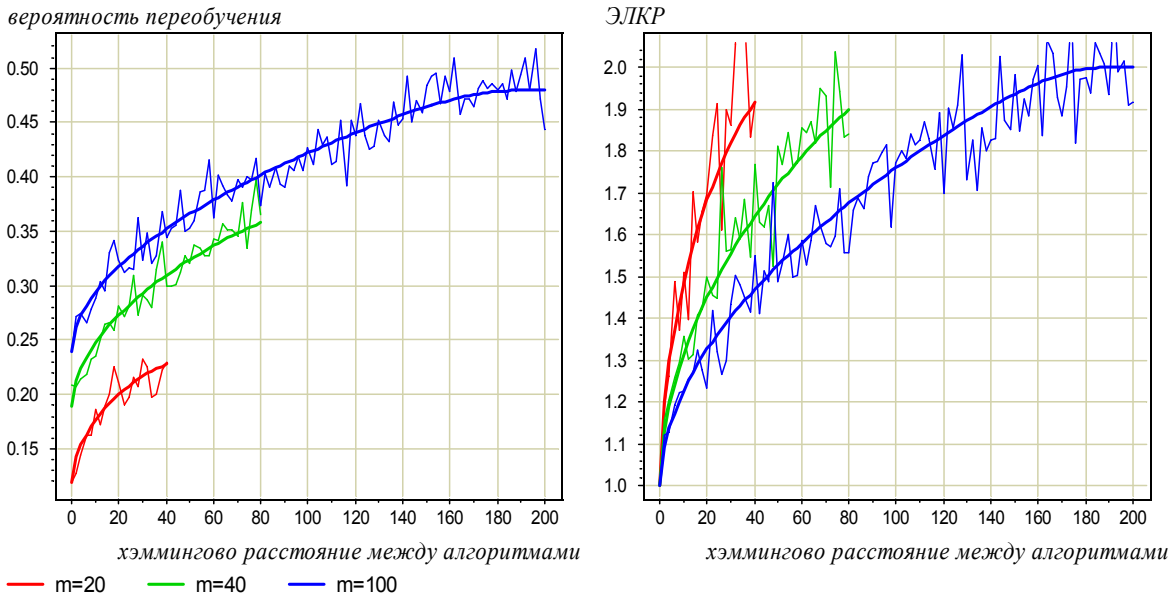


Рис. 3.3. Зависимость вероятности переобучения Q_ε и верхней оценки ЭЛКР $\bar{\Delta}_L^\ell$ от различности алгоритмов, когда они допускают одинаковое число ошибок, $m_1 = m_2$. Три графика соответствуют трём значениям числа ошибок на полной выборке $m = n(a_i, \mathbb{X}) = m_i + m_0 \in \{20, 40, 100\}$. Тонкими линиями показаны эмпирические оценки методом Монте-Карло по 1000 случайных разбиений.

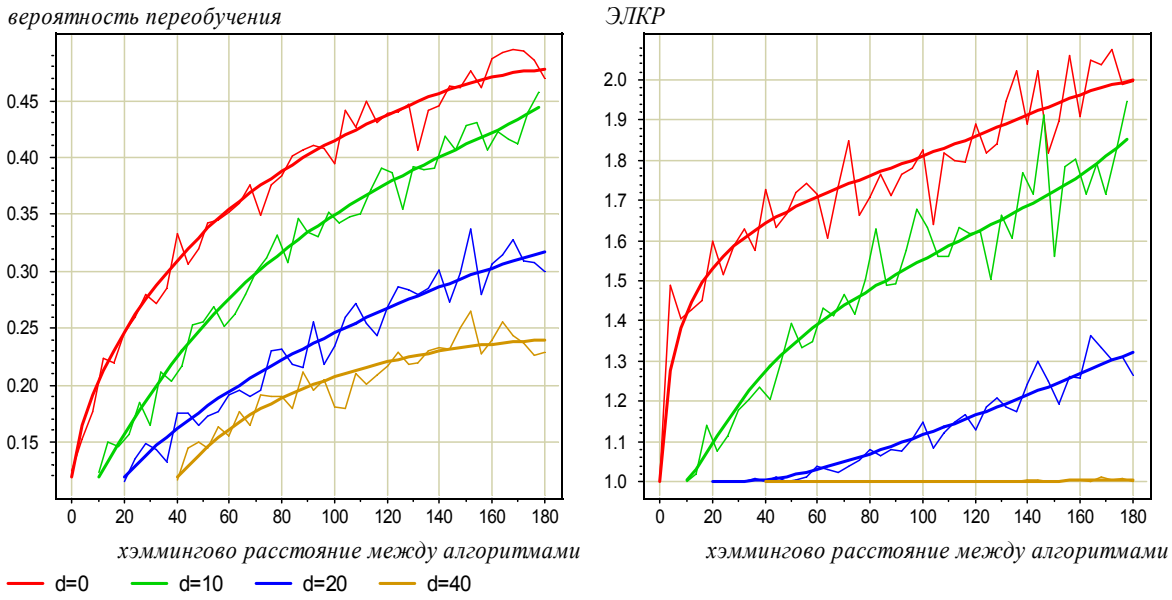


Рис. 3.4. Зависимость верхней оценки ЭЛКР $\bar{\Delta}_L^\ell$ от различности алгоритмов, когда $m_0 = 20$ и второй алгоритм допускает на d ошибок больше, $m_2 = m_1 + d$. Четыре графика соответствуют четырём разным значениям $d \in \{0, 10, 20, 40\}$. Тонкими линиями показаны эмпирические оценки методом Монте-Карло по 1000 случайных разбиений.

Графики позволяют сделать следующие выводы.

1. VC-оценка $\bar{\Delta}_L^\ell = 2$ достигается лишь в том случае, когда нет расслоения (алгоритмы допускают на $\mathbb{X}$ одинаковое число ошибок, $m_1 = m_2$) и нет сходства (алгоритмы максимально различны, $m_0 = m_3 = 0$).

2. Чем сильнее расслоение ($d = m_2 - m_1 > 0$), тем меньше вероятность переобучения и тем ближе ЭЛКР к 1. В данном эксперименте при $d = 40$ худший из двух алгоритмов уже практически никогда не выбирается.

3. Чем сильнее сходство ($m_1, m_2 \to 0$), тем меньше вероятность переобучения и тем ближе ЭЛКР к 1. С точки зрения переобучения два схожих алгоритма ведут себя практически как один алгоритм.

Таким образом, даже в простейшем случае, когда алгоритмов только два, уже возникает явление переобучения, и уже проявляются эффекты расслоения и сходства, снижающие вероятность переобучения.

3.3.2 Монотонная цепочка алгоритмов

Последовательность алгоритмов $\{a_1, \ldots, a_D\}$ будем называть *цепочкой*, если *хэммингово расстояние* между векторами ошибок a_{t-1} и a_t равно 1 для всех $t = 2, \ldots, D$. Цепочка алгоритмов является простейшим примером связного семейства алгоритмов [205]. Её можно рассматривать как модель однопараметрического семейства классификаторов с непрерывной по параметру дискриминантной функцией, см. пример 154 на стр. 154.

В модельном эксперименте будем генерировать цепочки алгоритмов как последовательности векторов ошибок $a_1, \ldots, a_D$ и исследовать зависимости вероятности переобучения Q_ε от длины цепочки D.

Рассмотрим модельные цепочки двух типов.

1. *Цепочка с расслоением*. Лучший алгоритм a_1 допускает m ошибок на полной выборке. Каждый следующий вектор ошибок a_t получается из a_{t-1} путём инверсии одной случайно выбранной координаты. Если цепочка достаточно длинная ($D \gg L$), то большинство алгоритмов допускают число ошибок m, близкое к $L/2$.

2. *Цепочка без расслоения*. Число ошибок алгоритмов на полной выборке, чередуясь, принимает значения m и $m + 1$.

На рис. 3.5 показаны зависимости частоты ошибок $\nu(a_t, \mathbb{X})$ от порядкового номера t алгоритма для цепочки с расслоением (слева) и без расслоения (справа).

Для произвольной цепочки $a_1, \ldots, a_D$ можно построить соответствующую ей *нецепочку* $a'_1, \ldots, a'_D$ с таким же распределением частот ошибок: $\nu(a'_t, \mathbb{X}) = \nu(a_t, \mathbb{X})$ для всех $t = 1, \ldots, D$. Чтобы соседние алгоритмы a'_{t-1}, a'_t существенно различались, векторы ошибок a'_t предлагается генерировать случайным образом, но при соблюдении требования, чтобы число единиц в векторе равнялось $\nu(a_t, \mathbb{X})$.

Итого, в модельном эксперименте строится четыре конечных семейства алгоритмов. Они задаются непосредственно своими $L \times D$-матрицами ошибок и имеют одинаковые значения параметров D и m.

Рис. 3.5. Распределение алгоритмов по частоте ошибок на генеральной выборке в модельных цепочках с расслоением и без расслоения. Зависимость $\nu(a_t, \mathbb{X})$ от t при $\ell = k = 100,\ m = 10$.

Сопоставление этих четырёх случаев позволяет разделить влияние *связности* (цепочки или не-цепочки) и *расслоения* (m ошибок у всех алгоритмов или только у лучшего) на вероятность переобучения.

На рис. 3.6 и рис. 3.7 показаны зависимости вероятности переобучения Q_ε и ЭЛКР $\bar{\Delta}_L^\ell$ от числа алгоритмов D для четырёх типов семейств, при $\ell = k = 100$, $\varepsilon = 0.05,\ m \in \{10, 50\}$. Значения Q_ε вычислялись *методом Монте-Карло* по 10000 случайных разбиений. Условные обозначения на графиках: $+Ц$ — цепочка, $-Ц$ — не-цепочка, $+P$ — с расслоением, $-P$ — без расслоения.

Графики $Q_\varepsilon(D)$, $\bar{\Delta}_L^\ell(D)$ позволяют сделать следующие выводы.

1. По мере увеличения D как вероятность переобучения, так и ЭЛКР выходят на горизонтальную асимптоту и перестают зависеть от D. В то же время, VC-оценка линейна по D и вообще не имеет горизонтальной асимптоты. VC-оценка достигается только для не-цепочек и только при малых D; в данном эксперименте — при $D < 10$.

2. Связность заметно снижает темп роста зависимости $Q_\varepsilon(D)$.

3. Расслоение понижает уровень горизонтальной асимптоты $Q_\varepsilon(D)$, особенно для «лёгкой задачи» с меньшим значением m, рис. 3.6. В случае расслоения вероятность переобучения Q_ε может вообще не достигать 1. Это связано с тем, что метод обучения крайне редко выбирает алгоритмы из верхних слоёв. Сколько там алгоритмов — практически не имеет значения.

4. Для относительно простых задач, когда существует алгоритм с низким уровнем ошибок, расслоение сильнее, чем связность, сказывается на уменьшении вероятности переобучения, рис. 3.6. При увеличении сложности задачи влияние расслоения уменьшается, рис. 3.7. При этом свойство расслоения в отдельности уже практически не влияет на оценку, однако в совокупности со свойством связности, опять-таки, приводит к существенному уменьшению вероятности переобучения Q_ε.

5. При больших D только одновременное наличие расслоения и связности позволяет избежать сильного переобучения (нижние кривые на каждом из графиков).

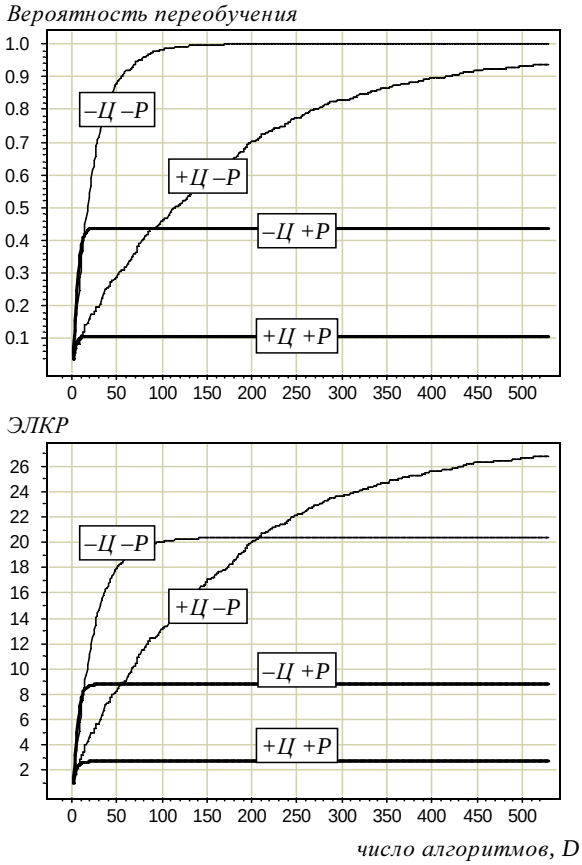


Рис. 3.6. Зависимость вероятности переобучения Q_ε и ЭЛКР $\bar{\Delta}_L^\ell$ от числа алгоритмов D. Простая задача: $\nu(a_1, \mathbb{X}) = 0.05$.

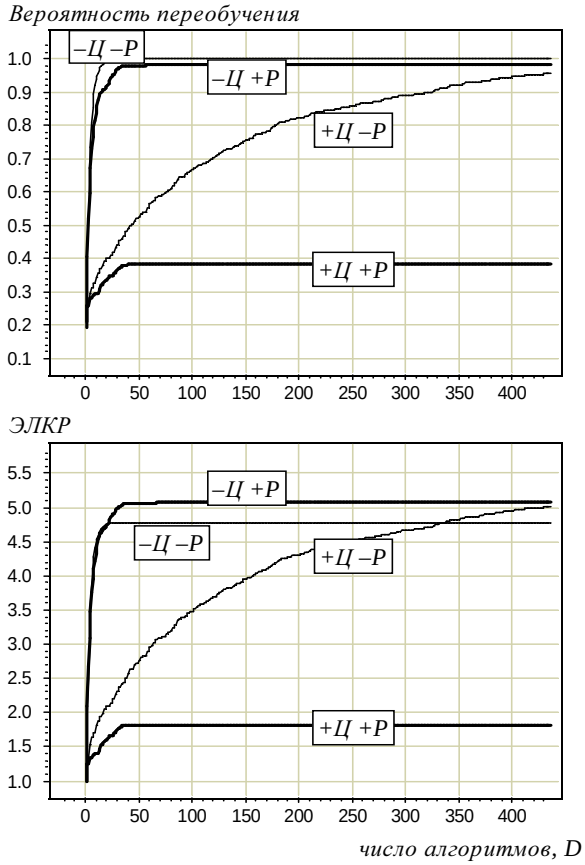


Рис. 3.7. Зависимость вероятности переобучения Q_ε и ЭЛКР $\bar{\Delta}_L^\ell$ от числа алгоритмов D. Трудная задача: $\nu(a_1, \mathbb{X}) = 0.25$.

3.4 Основные выводы

1. Предлагается *эмпирическая методика* измерения факторов завышенности VC-оценок, основанная на скользящем контроле.
2. Эксперименты с логическими алгоритмами классификации на реальных задачах из репозитория UCI показывают, что наиболее значимых причин завышенности две — это эффекты *расслоения* и *связности* семейства алгоритмов, которые никак не учитываются в VC-оценках. Каждый из них может приводить к завышенности оценок вероятности переобучения в 10^3–10^6 раз.
3. Эксперимент на *модельном семействе из двух алгоритмов* показывает, что даже в столь простом случае возникает переобучение, а эффекты расслоения и сходства снижают вероятность переобучения. Для произвольного двухэлементного семейства выводится точная оценка вероятности переобучения.
4. Простой эксперимент на четырёх модельных семействах алгоритмов, обладающих и не обладающих свойствами *расслоения* и *связности*, является поворотным моментом во всём исследовании. Оказывается, что если семейство алгоритмов

не обладает одновременно свойствами расслоения и связности, то вероятность переобучения может достигать $\frac{1}{2}$ уже при нескольких десятках алгоритмов в семействе. Отсюда вытекают два концептуально важных вывода. Во-первых, реальные семейства, содержащие миллиарды различных алгоритмов, с необходимостью являются расслоенными и связными. По всей видимости, этот случай наиболее распространён, и именно для него в первую очередь необходимо получать оценки. Во-вторых, получение точных оценок вероятности переобучения возможно только при совместном учёте обоих этих двух свойств.

5. Известные в теории статистического обучения подходы не дают точных оценок вероятности переобучения ни для общего случая расслоенного связного семейства, ни даже для такого простого частного случая, как *цепочка алгоритмов с расслоением*. В следующей главе такие оценки будут получены в рамках слабой аксиоматики, чисто комбинаторными методами.

Глава 4

Точные оценки вероятности переобучения

В данной главе выводятся точные оценки вероятности переобучения, основанные на предположении, что для каждого алгоритма $a \in A$ возможно в явном виде выписать условия, при которых a является результатом обучения: $\mu X = a$.

В параграфе 4.1 выводятся общие оценки вероятности переобучения двух типов.

Оценки первого типа основаны на понятиях порождающих и запрещающих множеств объектов. *Порождающее множество* алгоритма a — это множество объектов, которые обязательно должны присутствовать в обучающей выборке, чтобы данный алгоритм a был выбран в результате обучения. Здесь напрашивается аналогия с такими известными методами классификации, как метод опорных векторов SVM [123, 118] или метод отбора эталонов STOLP [51, 6]. В этих методах происходит выделение подмножества «опорных» или «эталонных» объектов, на основе которых и формируется результат обучения. Порождающее множество отличается тем, что оно зависит не только от обучающей выборки, но ещё и от алгоритма $a \in A$. Аналогично, *запрещающее множество* алгоритма a — это множество объектов, которых не должно быть в обучающей выборке, чтобы алгоритм a был выбран методом обучения μ. В некотором смысле это «шумовые» объекты или «выбросы», которые мешают данному алгоритму быть лучше других на обучающей выборке. Опять-таки, множество запрещающих объектов у каждого алгоритма своё. Доказываются теоремы о том, что множества порождающих и запрещающих объектов можно указать всегда, а если они указаны, то можно выписать точную формулу как для вероятности получить каждый из алгоритмов $P_a = \mathsf{P}[\mu X{=}a]$, так и для вероятности переобучения Q_ε.

Оценки второго типа основаны на разбиении генеральной выборки на блоки. Эти оценки эффективны только при малом числе алгоритмов в семействе.

В параграфе 4.2 рассматриваются модельные семейства алгоритмов: пара алгоритмов, слой и интервал булева куба, монотонные и унимодальные цепочки, единичная окрестность. С помощью инструментария, введённого в предыдущем параграфе, доказываются точные оценки вероятности переобучения для этих семейств.

В параграфе 4.3 рассматривается рекуррентная процедура вычисления вероятности переобучения. Идея заключается в том, чтобы добавлять алгоритмы в семейство по одному, и после каждого добавления корректировать множества порождающих и запрещающих объектов для всех алгоритмов семейства. Непосредственное вычисление вероятности переобучения по этому алгоритму не очень эффективно. Доказывается, что если в алгоритме пропускать определённые шаги, то он будет выполняться гораздо быстрее и с гарантией даст либо верхнюю, либо нижнюю оценку вероятности переобучения. При максимальном упрощении алгоритма верхняя оценка вероятности переобучения выписывается в явном аналитическом виде. Она похожа на VC-оценку, но зависит от профиля расслоения и связности семейства алгоритмов. Благодаря свойству связности оценка оказывается экспоненциально лучше (по размерности пространства), чем классическая VC-оценка.

4.1 Общие оценки вероятности переобучения

4.1.1 О разновидностях минимизации эмпирического риска

Будем полагать, что A — конечное множество, и все алгоритмы имеют попарно различные векторы ошибок.

Обозначим через $A(X)$ множество алгоритмов с минимальным числом ошибок на обучающей выборке X:

$$A(X) = \operatorname{Arg}\min_{a\in A} n(a, X). \tag{4.1}$$

Определение 4.1. *Метод обучения μ называется минимизацией эмпирического риска, МЭР (empirical risk minimization, ERM), если $\mu X \in A(X)$ при всех $X \in [\mathbb{X}]^\ell$.*

Если множество $A(X)$ содержит более одного элемента, то в методе μ возникает проблема неоднозначности выбора алгоритма. Рассмотрим сначала два крайних случая — когда выбирается наилучший или наихудший алгоритм из $A(X)$.

Определение 4.2. *Метод минимизации эмпирического риска μ называется оптимистичным, если $\mu X = \arg\min\limits_{a\in A(X)} n(a, \bar{X})$.*

Оптимистичная минимизация эмпирического риска на практике не реализуема, так как скрытая контрольная выборка $\bar{X}$ не может быть использована на этапе обучения. Теоретически оптимистичная МЭР интересна тем, что даёт неулучшаемую точную нижнюю оценку вероятности переобучения.

В некоторых случаях оптимистичная МЭР приводит к выбору глобально лучшего алгоритма $a_0 = \arg\min\limits_{a\in A} n(a, \mathbb{X}) = \mu X$ при любой обучающей выборке X. Например, это происходит в случае монотонной цепочки алгоритмов, которая будет определена ниже, стр. 153. В таких случаях вероятность переобучения Q_ε всего семейства A совпадает с Q_ε одноэлементного семейства $\{a_0\}$ и равна, согласно Теореме 1.11,

$$Q_\varepsilon = H_L^{\ell,m}\left(s_m^-(\varepsilon)\right), \quad m = n(a_0, \mathbb{X}).$$

Однако оптимистичная МЭР далеко не всегда тривиальна. Легко строится пример двухэлементного семейства $A = \{a_1, a_2\}$, в котором $n(a_1, \mathbb{X}) > n(a_2, \mathbb{X})$, и такого разбиения $(X, \bar{X})$, что $n(a_1, X) < n(a_2, X)$. Следовательно, любой метод МЭР, включая оптимистичный, выберет алгоритм a_1, который не является глобально лучшим.

Определение 4.3. *Метод минимизации эмпирического риска μ называется пессимистичным, если* $\mu X = \arg\max\limits_{a \in A(X)} n(a, \bar{X})$.

Пессимистичная минимизация эмпирического риска также не реализуема на практике. Тем не менее, она представляет значительный теоретический интерес, поскольку даёт точную верхнюю оценку вероятности переобучения. При любых других способах разрешения неоднозначности в методе МЭР вероятность переобучения может оказаться только меньше.

Более практичным представляется способ разрешения неоднозначности, основанный на случайном выборе алгоритма из множества $A(X)$.

Определение 4.4. *Метод минимизации эмпирического риска μ называется рандомизированным, если μX — это произвольный алгоритм, выбранный случайно и равновероятно из конечного множества алгоритмов $A(X)$.*

Существенная особенность рандомизированной МЭР состоит в том, что в задаче появляется второй источник случайности. Если ранее предполагалось, что случайным является только разбиение $X \sqcup \bar{X}$, то теперь случаен также и выбор алгоритма a из множества $A(X)$. Соответствующим образом изменяется и определение вероятности переобучения:

$$Q_\varepsilon = \mathsf{P} \frac{1}{|A(x)|} \sum_{a \in A(X)} \big[\nu(a, \bar{X}) - \nu(a, X) \geqslant \varepsilon\big]. \tag{4.2}$$

Большинство получаемых далее оценок основаны на пессимистичной МЭР и предположении, что результат обучения μX является детерминированной функцией от обучающей выборки X. Но в некоторых случаях будут даны также и оценки функционала (4.2) для рандомизрованной МЭР.

Заметим, что фактор завышенности, связанный с неоднозначностью выбора в МЭР и использованием пессимистичной МЭР, как правило, очень мал в сравнении с другими факторами завышенности VC-оценок. Эксперименты показывают, что оптимистичные, пессимистичные и рандомизированные оценки, как правило, быстро сходятся к одной и той же величине с ростом длины выборки L.

4.1.2 Порождающие и разрушающие множества объектов

Гипотеза 4.1. *Пусть множество A, выборка $\mathbb{X}$ и метод μ таковы, что для каждого алгоритма $a \in A$ можно указать пару непересекающихся подмножеств $X_a \subset \mathbb{X}$ и $X'_a \subset \mathbb{X}$, удовлетворяющую условию*

$$\big[\mu X{=}a\big] = \big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big], \quad \forall X \in [\mathbb{X}]^\ell. \tag{4.3}$$

Множество X_a будем называть *порождающим*, множество X'_a — *запрещающим* алгоритм a. Гипотеза 4.1 означает, что метод μ выбирает алгоритм a тогда и только тогда, когда в обучающей выборке X находятся все порождающие объекты и ни одного запрещающего. Все остальные объекты $\mathbb{X}\backslash X_a\backslash X'_a$ будем называть *нейтральными* для алгоритма a. Наличие или отсутствие нейтральных объектов в обучающей выборке не влияет на результат обучения. В следующих параграфах будут приведены нетривиальные примеры семейств, для которых гипотеза 4.1 выполняется.

Лемма 4.1. *Для любой выборки X справедливо тождество*

$$\sum_{a\in A}\big[\mu X{=}a\big] = 1. \tag{4.4}$$

Доказательство с очевидностью вытекает из того, что для любой выборки X метод μ выбирает один и только один алгоритм. Тождество (4.4) может использоваться для проверки того, что условия в правой части (4.3) сформулированы корректно.

Для произвольного $a \in A$ обозначим через L_a число нейтральных объектов, через ℓ_a — число нейтральных объектов, попадающих в обучающую выборку:

$$L_a = L - |X_a| - |X'_a|;$$
$$\ell_a = \ell - |X_a|.$$

Лемма 4.2. *Если гипотеза 4.1 справедлива, то вероятность получить в результате обучения алгоритм a равна*

$$P_a = \mathsf{P}\big[\mu X{=}a\big] = \frac{C^{\ell_a}_{L_a}}{C^{\ell}_{L}}.$$

Доказательство. Согласно гипотезе 4.1

$$\mathsf{P}\big[\mu X{=}a\big] = \mathsf{P}\big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big].$$

Это есть доля разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$ таких, что множество объектов X_a целиком лежит в X, а множество объектов X'_a целиком лежит в $\bar{X}$. Число таких разбиений равно числу способов отобрать ℓ_a из L_a нейтральных объектов в обучающую подвыборку $X\backslash X_a$, которое, очевидно, равно $C^{\ell_a}_{L_a}$. Общее число разбиений равно C^{ℓ}_{L}, а их отношение как раз и есть P_a. ■

Вероятность переобучения Q_ε выражается по формуле полной вероятности, если для каждого алгоритма a из A известна вероятность P_a получить его в результате обучения и условная вероятность большого отклонения частот $\mathsf{P}\big(\delta(a,X) \geqslant \varepsilon \mid a\big)$ при условии, что получен алгоритм a:

$$Q_\varepsilon = \sum_{a\in A} P_a\, \mathsf{P}\big(\delta(a,X) \geqslant \varepsilon \mid a\big).$$

Условная вероятность даётся Леммой 1.11, если учесть, что при фиксированном алгоритме a подмножества X_a и X'_a не участвуют в разбиениях. Рассматривая

L_a нейтральных объектов и всевозможные их разбиения на ℓ_a обучающих и $L_a - \ell_a$ контрольных, получим:

$$\mathsf{P}\big(\delta(a,X) \geqslant \varepsilon \mid a\big) = H_{L_a}^{\ell_a, m_a}\left(s_a(\varepsilon)\right),$$

где m_a — число ошибок алгоритма a на нейтральных объектах; $s_a(\varepsilon)$ — наибольшее число ошибок алгоритма a на нейтральных обучающих объектах $X \setminus X_a$, при котором имеет место большое отклонение частот ошибок, $\delta(a,X) \geqslant \varepsilon$:

$$\begin{aligned} m_a &= n(a, \mathbb{X} \setminus X_a \setminus X'_a); \\ s_a(\varepsilon) &= \tfrac{\ell}{L}\big(n(a,\mathbb{X}) - \varepsilon k\big) - n(a, X_a). \end{aligned}$$

Более строгий комбинаторный вывод точной оценки Q_ε представлен ниже.

Теорема 4.3. *Если гипотеза 4.1 справедлива, то вероятность переобучения вычисляется по формуле*

$$Q_\varepsilon = \sum_{a \in A} P_a H_{L_a}^{\ell_a, m_a}\left(s_a(\varepsilon)\right).$$

Доказательство. Рассмотрим функционал Q_ε. Введём в (**??**) под знак суммирования по X ещё два вспомогательных суммирования: первый — по всем алгоритмам a из A при условии $\mu X = a$, второй — по всем значениям s числа ошибок алгоритма a на подвыборке $X \setminus X_a$. Очевидно, значение Q_ε от этого не изменится:

$$Q_\varepsilon = \mathsf{P}\big[\delta_\mu(X) \geqslant \varepsilon\big] = \mathsf{P} \sum_{a \in A} \big[\mu X = a\big] \sum_{s=0}^{\ell_a} \big[n(a, X \setminus X_a) = s\big] \big[\delta(a,X) \geqslant \varepsilon\big]. \tag{4.5}$$

Число ошибок алгоритма a на обучающей подвыборке X равно $s + n(a, X_a)$, поэтому отклонение частот ошибок выражается в виде

$$\delta(a,X) = \frac{n(a,\mathbb{X}) - s - n(a,X_a)}{k} - \frac{s + n(a,X_a)}{\ell},$$

следовательно,

$$\big[\delta(a,X) \geqslant \varepsilon\big] = \big[s \leqslant \tfrac{\ell}{L}\big(n(a,\mathbb{X}) - \varepsilon k\big) - n(a,X_a)\big] = \big[s \leqslant s_a(\varepsilon)\big].$$

Подставим полученное выражение в (4.5), затем заменим $\big[\mu X = a\big]$ правой частью равенства (4.3) и переставим знаки суммирования (очевидно, P также можно рассматривать как суммирование):

$$Q_\varepsilon = \sum_{a \in A} \sum_{s=0}^{\ell_a} \underbrace{\mathsf{P}\big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big]\big[n(a, X \setminus X_a) = s\big]}_{N(a)} \big[s \leqslant s_a(\varepsilon)\big]. \tag{4.6}$$

Выделенное в данной формуле выражение $N(a)$ есть доля разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$ таких, что множество объектов X_a целиком лежит в X,

множество объектов X'_a целиком лежит в $\bar{X}$, и в подвыборку $X\backslash X_a$ длины ℓ_a попадает ровно s объектов, на которых алгоритм a допускает ошибку.

Для наглядности представим вектор ошибок a разбитым на шесть блоков:

$$\vec{a} = \Big(\underbrace{X_a; \overbrace{1,\ldots,1}^{s}; \underbrace{0,\ldots,0}_{X\backslash X_a}}_{X}; \underbrace{X'_a; \overbrace{1,\ldots,1}^{m_a-s}; \underbrace{0,\ldots,0}_{\bar{X}\backslash X'_a}}_{\bar{X}} \Big).$$

Число ошибок алгоритма a на объектах, не попадающих ни в X_a, ни в X'_a, равно m_a. Существует $C^s_{m_a}$ способов выбрать из них s объектов, которые попадут в $X\backslash X_a$. Для каждого из этих способов имеется ровно $C^{\ell_a-s}_{L_a-m_a}$ способов выбрать $\ell_a - s$ объектов, на которых алгоритм a не допускает ошибку, и которые также попадут в $X\backslash X_a$. Тем самым однозначно определяется состав выборки $X\backslash X_a$, а, значит, и состав выборки $\bar{X}\backslash X'_a$. Таким образом, $N(a) = C^s_{m_a} C^{\ell_a-s}_{L_a-m_a} / C^\ell_L$. Подставим это выражение в (4.6) и выделим в нём формулу гипергеометрической функции вероятности:

$$Q_\varepsilon = \sum_{a\in A} \frac{C^{\ell_a}_{L_a}}{C^\ell_L} \sum_{s=s_0}^{\ell_a} \big[s \leqslant s_a(\varepsilon)\big] \frac{C^s_{m_a} C^{\ell_a-s}_{L_a-m_a}}{C^{\ell_a}_{L_a}} = \sum_{a\in A} P_a H^{\ell_a,m_a}_{L_a}\left(s_a(\varepsilon)\right).$$

Теорема доказана. ■

Ослабление гипотезы о порождающих и запрещающих объектах. Гипотеза 4.1 накладывает слишком сильные ограничения на выборку $\mathbb{X}$, семейство A и метод μ. Поэтому Теорему 4.3 удаётся применять лишь в некоторых специальных случаях. Рассмотрим естественное обобщение гипотезы 4.1. Предположим, что для каждого алгоритма a существуют различные варианты выделения порождающих и запрещающих множеств.

Гипотеза 4.2. *Пусть множество A, выборка $\mathbb{X}$ и метод μ таковы, что для каждого алгоритма $a \in A$ можно указать конечное множество индексов V_a, и для каждого индекса $v \in V_a$ можно указать порождающее множество $X_{av} \subset \mathbb{X}$, запрещающее множество $X'_{av} \subset \mathbb{X}$ и коэффициент $c_{av} \in \mathbb{R}$, удовлетворяющие условиям*

$$\big[\mu X{=}a\big] = \sum_{v\in V_a} c_{av} \big[X_{av} \subseteq X\big] \big[X'_{av} \subseteq \bar{X}\big], \quad \forall X \in [\mathbb{X}]^\ell. \tag{4.7}$$

Гипотеза 4.1 является частным случаем гипотезы 4.2, когда все множества V_a одноэлементные и $c_{av} = 1$. Примеры задач, удовлетворяющих гипотезам 4.1 или 4.2, будут рассмотрены в следующих параграфах.

Следующая теорема утверждает, что гипотеза 4.2 верна всегда.

Теорема 4.4. *Для любых $\mathbb{X}$, A и μ существуют множества V_a, X_{av}, X'_{av}, при которых справедливо представление* (4.7)*, причём $c_{av} = 1$ для всех $a \in A$, $v \in V_a$.*

Доказательство. Зафиксируем произвольный алгоритм $a \in A$. Возьмём в качестве индексного множества V_a множество всех подвыборок $v \in [\mathbb{X}]^\ell$, при которых $\mu v = a$. Для каждого $v \in V_a$ положим $X_{av} = v$, $X'_{av} = \mathbb{X} \setminus v$, $c_{av} = 1$. Тогда для любого $X \in [\mathbb{X}]^\ell$ справедливо представление, имеющее вид (4.7):

$$\big[\mu X{=}a\big] = \sum_{v \in V_a} \big[v{=}X\big] = \sum_{v \in V_a} \big[v{=}X\big]\big[\mathbb{X}\backslash v{=}\mathbb{X}\backslash X\big] = \sum_{v \in V_a} \big[v \subseteq X\big]\big[\mathbb{X}\backslash v \subseteq \bar{X}\big],$$

причём, если $\mu X = a$, то ровно одно слагаемое в этой сумме равно единице, остальные равны нулю; если же $\mu X \neq a$, то все слагаемые равны нулю. ■

Теорема 4.4 является типичной теоремой существования. Использованный при её доказательстве способ построения индексных множеств V_a требует явного перебора всех разбиений выборки, что приводит к вычислительно неэффективным оценкам вероятности переобучения. Однако представление (4.7) в общем случае не единственно. Отдельной проблемой является поиск такого представления, в котором мощности множеств $|V_a|$, $|X_{av}|$, $|X'_{av}|$ были бы как можно меньше. Эффективные вычислительные алгоритмы, эксплуатирующие свойства расслоения и сходства в семействах алгоритмов, рассматриваются в следующих параграфах.

Хотя гипотеза 4.2 верна всегда, мы будем продолжать называть её «гипотезой», имея в виду предположение о существовании некоторого представления вида (4.7), более эффективного, чем использованное в Теореме 4.4.

Введём для каждого алгоритма $a \in A$ и каждого индекса $v \in V_a$ обозначения:

$$\begin{aligned} L_{av} &= L - |X_{av}| - |X'_{av}|; \\ \ell_{av} &= \ell - |X_{av}|; \\ m_{av} &= n(a, \mathbb{X}) - n(a, X_{av}) - n(a, X'_{av}); \\ s_{av}(\varepsilon) &= \tfrac{\ell}{L}\big(n(a, \mathbb{X}) - \varepsilon k\big) - n(a, X_{av}). \end{aligned}$$

В условиях гипотезы 4.2 справедливы соответствующие обобщения леммы о вероятностях получения алгоритмов и теоремы о вероятности переобучения.

Лемма 4.5. *Если гипотеза 4.2 справедлива, то для всех $a \in A$ вероятность получить в результате обучения алгоритм a равна*

$$P_a = \mathsf{P}\big[\mu X{=}a\big] = \sum_{v \in V_a} c_{av} P_{av}; \tag{4.8}$$

$$P_{av} = \mathsf{P}\big[X_{av} \subseteq X\big]\big[X'_{av} \subseteq \bar{X}\big] = \frac{C_{L_{av}}^{\ell_{av}}}{C_L^\ell}. \tag{4.9}$$

Доказательство. Достаточно применить операцию P к левой и правой частям (4.7). Дальнейшие рассуждения аналогичны доказательству Леммы 4.2. ■

Теорема 4.6. *Если гипотеза 4.2 справедлива, то вероятность переобучения вычисляется по формуле*

$$Q_\varepsilon = \sum_{a \in A} \sum_{v \in V_a} c_{av} P_{av} H_{L_{av}}^{\ell_{av}, m_{av}}\left(s_{av}(\varepsilon)\right). \tag{4.10}$$

Доказательство. Аналогично доказательству Теоремы 4.3, вероятность переобучения приводится к выражению, которое отличается от (4.6) появлением знака суммирования по v, коэффициентов c_{av} и двойных индексов ${}_{av}$ вместо одинарных ${}_{a}$:

$$Q_\varepsilon = \sum_{a\in A}\sum_{v\in V_a}\sum_{s=0}^{\ell} c_{av}\,\mathsf{P}\big[X_{av}\subseteq X\big]\big[X'_{av}\subseteq \bar{X}\big]\big[n(a,X\backslash X_{av})=s\big]\big[s\leqslant s_{av}(\varepsilon)\big],$$

В остальном доказательство аналогично доказательству Теоремы 4.3. ■

Гипотеза корректности. Алгоритм a_0, не допускающий ошибок на выборке $U\subseteq\mathbb{X}$, называется *корректным на выборке* U. Формула (4.10) сильно упрощается, если в семействе A содержится алгоритм, корректный на всей генеральной выборке.

Теорема 4.7. *Пусть гипотеза 4.2 справедлива, метод μ является минимизацией эмпирического риска, множество A содержит алгоритм a_0 такой, что $n(a_0,\mathbb{X})=0$. Тогда вероятность переобучения принимает более простой вид:*

$$Q_\varepsilon = \sum_{a\in A}\big[n(a,\mathbb{X})\geqslant\varepsilon k\big]P_a. \tag{4.11}$$

Доказательство. Рассмотрим произвольный алгоритм $a\in A$ и произвольный индекс $v\in V_a$. Если некоторый объект, на котором a допускает ошибку, содержится в обучающей выборке X, то метод μ не сможет выбрать данный алгоритм, так как существует корректный алгоритм a_0, не допускающий ошибок на X. Следовательно, множество объектов, на которых алгоритм a допускает ошибку, целиком содержится в X'_{av}. Значит, алгоритм a не допускает ошибок на нейтральных объектах и $m_{av}=0$. В этом случае гипергеометрическая функция $H_{L_{av}}^{\ell_{av},0}\big(s_{av}(\varepsilon)\big)$ является вырожденной: при $s_{av}(\varepsilon)\geqslant 0$ она представляет собой сумму из одного слагаемого, равного 1; при $s_{av}(\varepsilon)<0$ число слагаемых равно нулю и вся сумма равна нулю:

$$H_{L_{av}}^{\ell_{av},0}\big(s_{av}(\varepsilon)\big) = \big[s_{av}(\varepsilon)\geqslant 0\big] = \big[\tfrac{\ell}{L}\big(n(a,\mathbb{X})-\varepsilon k\big)-n(a,X_{av})\geqslant 0\big] = \big[n(a,\mathbb{X})\geqslant\varepsilon k\big].$$

Подставляя это выражение в (4.10), получаем (4.11). ■

Оценки функционала R_ε доказываются полностью аналогично, с той лишь разницей, что выражение

$$s_{av}(\varepsilon) = \tfrac{\ell}{L}\big(n(a,\mathbb{X})-\varepsilon k\big)-n(a,X_{av})$$

всюду заменяется на

$$s'_{av}(\varepsilon) = n(a,\mathbb{X})-\varepsilon k-n(a,X_{av}).$$

Оценка (4.11) для корректного случая остаётся справедлива для функционала R_ε в неизменном виде.

4.1.3 Блочное вычисление вероятности переобучения

Допустим, что векторы ошибок всех алгоритмов из множества $A = \{a_1, \ldots, a_D\}$ известны и попарно различны.

Будем полагать, что μ — это пессимистичный метод минимизации эмпирического риска. Согласно определению 4.3, когда минимум $n(a, X)$ достигается на нескольких алгоритмах, μ выбирает алгоритм с бóльшим $n(a, \bar{X})$. Если же и таких алгоритмов несколько, то μ выбирает алгоритм с бóльшим порядковым номером.

Значения $I(a_d, x_i)$ образуют бинарную $L \times D$-матрицу ошибок, столбцы которой являются векторами ошибок алгоритмов, а строки соответствуют объектам. Обозначим через $b = (b_1, \ldots, b_D)$ произвольный бинарный вектор размерности D. Выборка $\mathbb{X}$ разбивается на непересекающиеся *блоки* $U_b \subseteq \mathbb{X}$ так, что всем объектам в блоке соответствует одна и та же строка $b = (b_1, \ldots, b_D)$ в матрице ошибок:

$$U_b = \big\{x_i \in \mathbb{X} \;\big|\; I(a_d, x_i) = b_d,\ d = 1, \ldots, D\big\}.$$

Обозначим через B множество бинарных векторов b, которым соответствуют непустые блоки U_b. Очевидно, $|B| \leqslant \min\{L, 2^D\}$.

Обозначим $m_b = |U_b|$.

Каждой обучающей выборке $X \in [\mathbb{X}]^\ell$ поставим в соответствие целочисленный вектор $(s_b)_{b \in B}$ такой, что $s_b = |X \cap U_b|$ — число объектов из блока U_b, попадающих в обучающую выборку. Множество всех таких векторов, соответствующих всевозможным обучающим выборкам, обозначим через S. Очевидно, S можно также определить и другим способом:

$$S = \Big\{s = (s_b)_{b \in B} \;\Big|\; s_b = 0, \ldots, m_b,\ \sum_{b \in B} s_b = \ell\Big\}.$$

Запишем число ошибок алгоритма a_d на обучающей выборке X и контрольной выборке $\bar{X}$ в виде суммы по блокам:

$$n(a_d, X) = \sum_{b \in B} b_d\, |X \cap U_b| = \sum_{b \in B} b_d s_b;$$
$$n(a_d, \bar{X}) = \sum_{b \in B} b_d\, |\bar{X} \cap U_b| = \sum_{b \in B} b_d (m_b - s_b).$$

Таким образом, выбор алгоритма методом μ зависит только от того, сколько объектов s_b из каждого блока попадёт в обучающую выборку, но не зависит от того, какие именно это будут объекты. Определим функцию $d^*\colon S \to \{1, \ldots, D\}$ как номер алгоритма, выбранного методом μ по обучающей выборке. Если μ — пессимистичная минимизация риска, то положим

$$A(s) = \operatorname*{Arg\,min}_{d=1,\ldots,D} \sum_{b \in B} b_d s_b,$$
$$A'(s) = \operatorname*{Arg\,max}_{d \in A(s)} \sum_{b \in B} b_d (m_b - s_b),$$
$$d^*(s) = \max\big\{d\colon d \in A'(s)\big\}, \tag{4.12}$$

где через $\underset{d=1,\ldots,D}{\operatorname{Arg\,min}} f(d)$ обозначается множество значений d, при которых функция $f(d)$ достигает минимального значения.

Теорема 4.8. *Пусть μ — пессимистичная минимизация эмпирического риска, векторы ошибок всех алгоритмов $a \in A$ попарно различны. Тогда вероятность получить алгоритм a_d в результате обучения:*

$$\mathsf{P}\big[\mu X{=}a_d\big] = \frac{1}{C_L^\ell} \sum_{s\in S} \Big(\prod_{b\in B} C_{m_b}^{s_b} \Big) \big[\, d^*(s){=}d \,\big]; \tag{4.13}$$

вероятность переобучения:

$$Q_\varepsilon = \frac{1}{C_L^\ell} \sum_{s\in S} \Big(\prod_{b\in B} C_{m_b}^{s_b} \Big) \Big[\sum_{b\in B} b_{d^*(s)}\big(m_b\ell - s_b L\big) \geqslant \varepsilon k\ell \,\Big]. \tag{4.14}$$

Доказательство.

Произвольному набору значений $(s_b)_{b\in B}$ из S соответствует множество выборок $X \in [\mathbb{X}]^\ell$ таких, что $|X \cap U_b| = s_b$. Число таких выборок равно произведению $\prod\limits_{b\in B} C_{m_b}^{s_b}$, так как для каждого блока U_b существует $C_{m_b}^{s_b}$ способов отобрать s_b объектов в подвыборку $X \cap U_b$.

Поскольку условия $\mu X = a_d$ и $d^*(s) = d$ равносильны, вероятность получить алгоритм a_d в результате обучения выражается в следующем виде:

$$\begin{aligned}\mathsf{P}\big[\mu X{=}a_d\big] &= \frac{1}{C_L^\ell} \sum_{X\in[\mathbb{X}]^\ell} \big[\mu X{=}a_d\big] = \\ &= \frac{1}{C_L^\ell} \sum_{X\in[\mathbb{X}]^\ell} \big[d^*(s){=}d\big] = \frac{1}{C_L^\ell} \sum_{s\in S} \Big(\prod_{b\in B} C_{m_b}^{s_b} \Big) \big[\, d^*(s){=}d \,\big].\end{aligned}$$

Теперь запишем вероятность переобучения:

$$Q_\varepsilon = \mathsf{P}\big[\delta_\mu(X) \geqslant \varepsilon\big] = \mathsf{P} \sum_{d=1}^{D} \big[\mu X{=}a_d\big] \big[\delta(a_d, X) \geqslant \varepsilon\big].$$

Распишем отклонение частот ошибок алгоритма a_d в виде суммы по блокам:

$$\delta(a_d, X) = \frac{1}{k} \sum_{b\in B} b_d(m_b - s_b) - \frac{1}{\ell} \sum_{b\in B} b_d s_b = \frac{1}{\ell k} \sum_{b\in B} b_d\big(m_b\ell - s_b L\big).$$

Тогда выражение для вероятности переобучения примет вид:

$$Q_\varepsilon = \frac{1}{C_L^\ell} \sum_{s\in S} \Big(\prod_{b\in B} C_{m_b}^{s_b} \Big) \sum_{d=1}^{D} \big[\, d^*(s){=}d \,\big] \Big[\sum_{b\in B} b_d\big(m_b\ell - s_b L\big) \geqslant \varepsilon\ell k \,\Big].$$

Отсюда немедленно вытекает требуемое выражение (4.14). ■

Замечание 4.1. Если множество B содержит векторы b, соответствующие пустым блокам U_b, то формулы (4.13) и (4.14) остаются в силе, поскольку тогда $m_b = s_b = 0$.

Следствие 4.8.1. *Аналогичная оценка справедлива для функционала R_ε. Пусть μ — пессимистичная минимизация эмпирического риска, векторы ошибок всех алгоритмов $a \in A$ попарно различны. Тогда*

$$R_\varepsilon = \frac{1}{C_L^\ell} \sum_{s \in S} \Big(\prod_{b \in B} C_{m_b}^{s_b} \Big) \Big[\sum_{b \in B} b_{d^*(s)} (m_b - s_b) \geqslant \varepsilon k \Big].$$

Теорема 4.8 позволяет, зная векторы ошибок всех алгоритмов, выписать вероятности получения каждого из алгоритмов и вероятность переобучения Q_ε. Однако непосредственные вычисления по формулам (4.13) и (4.14) могут потребовать значительных затрат времени, экспоненциальных по длине выборки L. В худшем случае, когда все блоки U_b одноэлементные, множество S состоит из всевозможных булевых векторов длины L, содержащих ровно ℓ единиц. При этом число слагаемых в (4.13) и (4.14) равно C_L^ℓ.

Вычисления по Теореме 4.8 эффективны только когда число блоков $|B|$ невелико, в частности, при малом числе алгоритмов.

4.2 Модельные семейства алгоритмов

В данном параграфе рассматриваются специальные семейства алгоритмов, для которых удаётся в явном виде выписать точные оценки вероятности переобучения. Все эти семейства являются «искусственными» в том смысле, что они задаются непосредственно бинарной матрицей ошибок, а не каким-либо реальным семейством алгоритмов и реальной выборкой. В некоторых случаях удаётся строить примеры выборок, для которых порождаются данные матрицы ошибок. Однако ясно, что число таких случаев исчезающе мало в сравнении с числом всевозможных матриц, порождаемых реальными задачами обучения. Все модельные семейства отличаются некоторой «регулярностью» или симметрией, которой, как правило, не обладают реальные семейства. Тем не менее, изучение модельных семейств представляется перспективным по нескольким причинам.

Во-первых, они хорошо иллюстрируют эффекты расслоения и связности.

Во-вторых, на них отрабатываются математические приёмы, которые могут оказаться полезными при получении оценок более общего вида.

В-третьих, рассмотрение большого числа разнообразных частных случаев ведёт к постепенному обобщению модельных семейств и получению оценок, неплохо аппроксимирующих реальные семейства. Такой путь развития комбинаторной теории переобучения представляется наиболее реалистичным.

4.2.1 Семейство из двух алгоритмов

Рассмотрим семейство из двух алгоритмов $A = \{a_1, a_2\}$. Точная оценка вероятности переобучения для данного случая уже была получена в 3.3.1. Здесь мы представим более короткое доказательство. Чтобы воспользоваться блочной оценкой, положим $B = (11, 10, 01, 00)$.

Пусть в выборке $\mathbb{X}$ имеется m_{11} объектов, на которых оба алгоритма допускают ошибку; m_{10} объектов, на которых только a_1 допускает ошибку; m_{01} объектов, на которых только a_2 допускает ошибку; $m_{00} = L - m_{11} - m_{10} - m_{01}$ объектов, на которых оба алгоритма дают верный ответ:

$$\vec{a}_1 = (\,1,\ldots,1,\ 1,\ldots,1,\ 0,\ldots,0,\ 0,\ldots,0\,);$$
$$\vec{a}_2 = (\underbrace{1,\ldots,1}_{m_{11}},\ \underbrace{0,\ldots,0}_{m_{10}},\ \underbrace{1,\ldots,1}_{m_{01}},\ \underbrace{0,\ldots,0}_{m_{00}}).$$

Теорема 4.9. *Пусть μ — пессимистичная минимизация эмпирического риска, и семейство состоит из двух алгоритмов $A = \{a_1, a_2\}$. Тогда при любом $\varepsilon \in [0, 1)$ справедлива точная оценка:*

$$Q_\varepsilon = \sum_{s_{11}=0}^{m_{11}} \sum_{s_{10}=0}^{m_{10}} \sum_{s_{01}=0}^{m_{01}} \frac{C_{m_{11}}^{s_{11}} C_{m_{10}}^{s_{10}} C_{m_{01}}^{s_{01}} C_{L-m_{11}-m_{10}-m_{01}}^{\ell-s_{11}-s_{10}-s_{01}}}{C_L^\ell} \times$$
$$\times \Big(\big[s_{10} < s_{01}\big] \big[s_{11} + s_{10} \leqslant \tfrac{\ell}{L}(m_{11} + m_{10} - \varepsilon k)\big] +$$
$$+ \big[s_{10} \geqslant s_{01}\big] \big[s_{11} + s_{01} \leqslant \tfrac{\ell}{L}(m_{11} + m_{01} - \varepsilon k)\big] \Big). \tag{4.15}$$

Доказательство. Воспользуемся Теоремой 4.8.

Множество S состоит из целочисленных векторов $s = (s_{11}, s_{10}, s_{01}, s_{00})$, для которых $s_{11} + s_{10} + s_{01} + s_{00} = \ell$. Поэтому сумма $\sum\limits_{s \in S}$ преобразуется в тройную сумму $\sum\limits_{s_{11}=0}^{m_{11}} \sum\limits_{s_{10}=0}^{m_{10}} \sum\limits_{s_{01}=0}^{m_{01}}$, при этом s_{00} выражается через остальные компоненты вектора s.

Номер $d^*(s)$ алгоритма, выбранного методом μ по обучающей выборке, равен 1 при $s_{10} < s_{01}$ и 2 при $s_{10} \geqslant s_{01}$.

Теперь подставим значения m_b, s_b, $d^*(s)$ в (4.14):

$$\left[\sum_{b \in B} b_{d^*(s)} \big(m_b \ell - s_b L \big) \geqslant \varepsilon \ell k \right] = \big[d^*(s) = 1\big] \big[(m_{10} + m_{11})\ell - (s_{10} + s_{11})L \geqslant \varepsilon \ell k\big] +$$
$$+ \big[d^*(s) = 2\big] \big[(m_{01} + m_{11})\ell - (s_{01} + s_{11})L \geqslant \varepsilon \ell k\big].$$

Отсюда следует требуемое равенство (4.15). ■

Заметим, что возможен и третий способ доказательства — через порождающие и запрещающие множества, с использованием Гипотезы 4.2 и Теоремы 4.6. Он является наиболее громоздким, поэтому мы не будем его здесь приводить.

4.2.2 Слой булева куба

Рассмотрим множество A, состоящее из всех C_L^m алгоритмов, допускающих ровно m ошибок на полной выборке $\mathbb{X}$, и имеющих попарно различные векторы ошибок. Поскольку все возможные векторы ошибок образуют булев куб размерности L, то векторы ошибок множества A — это m-й слой булева куба.

Теорема 4.10. *Пусть μ — произвольный метод минимизации эмпирического риска, A — m-й слой булева куба. Тогда для любого $\varepsilon \in [0,1]$*

$$Q_\varepsilon = \big[\varepsilon k \leqslant m \leqslant \ell - \varepsilon\ell\big].$$

Доказательство. Согласно Теореме 1.14, в случае одного слоя алгоритмов минимизация эмпирического риска эквивалентна принципу равномерной сходимости. Следовательно,

$$Q_\varepsilon = \mathsf{P}\left[\max_{a\in A}\big(\nu(a,\bar{X}) - \nu(a,X)\big) \geqslant \varepsilon\right].$$

Если $m \leqslant k$, то максимум достигается на таком $a \in A$, у которого в контроль попадают все m ошибок, а в обучение — ни одной. Тогда $\nu(a,\bar{X}) = \frac{m}{k}$, $\nu(a,X) = 0$ и

$$Q_\varepsilon = \mathsf{P}\big[\tfrac{m}{k} - 0 \geqslant \varepsilon\big] = [m \geqslant \varepsilon k] = [\varepsilon k \leqslant m \leqslant k].$$

Если $m > k$, то максимум достигается на таком $a \in A$, который допускает ошибки на всех контрольных объектах. Тогда

$$Q_\varepsilon = \mathsf{P}\big[1 - \tfrac{m-k}{\ell} \geqslant \varepsilon\big] = [m \leqslant L - \varepsilon\ell] = [k < m \leqslant L - \varepsilon\ell].$$

Объединяя два несовместных случая $m \leqslant k$ и $m > k$, получаем требуемое. ■

Таким образом, точная оценка вероятности переобучения является вырожденной и принимает значения либо 0, либо 1. Хотя этот результат тривиальный и в определённом смысле отрицательный, он позволяет сделать несколько важных выводов.

Во-первых, алгоритмы самых нижних слоёв, $m < \lceil\varepsilon k\rceil$, не вносят вклад в переобучение.

Во-вторых, никакой слой уровней $m = \lceil\varepsilon k\rceil, \ldots, \lfloor L - \varepsilon\ell\rfloor$ не должен полностью содержаться в семействе алгоритмов. Бессмысленно решать задачи обучения методом МЭР, используя слишком богатое семейство алгоритмов, в частности, множество всех возможных алгоритмов или любое его подмножество, целиком включающее в себя один из слоёв выше $\lceil\varepsilon k\rceil$.

4.2.3 Интервал булева куба

Предположим, что векторы ошибок всех алгоритмов из A попарно различны и образуют интервал ранга m в L-мерном булевом кубе. Это означает, что объекты делятся на три группы: m_0 «внутренних» объектов, на которых ни один из алгоритмов не допускает ошибок; m_1 «шумовых» объектов, на которых все алгоритмы

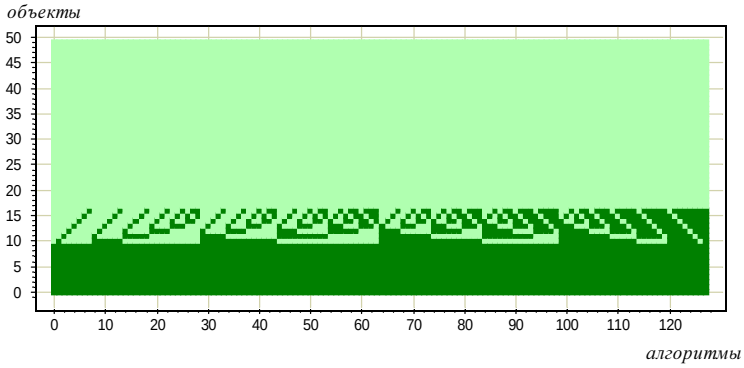


Рис. 4.1. Матрица ошибок интервала булева куба при $L = 50,\ m = 7,\ m_1 = 10$. Число алгоритмов в семействе $D = 2^m = 128$.

допускают ошибки; и m «пограничных» объектов, на которых реализуются все 2^m вариантов допустить ошибки. Других объектов нет: $m_0 + m_1 + m = L$. На рис. 4.1 показан пример матрицы ошибок для интервала булева куба ранга $m = 7$.

Интервал булева куба обладает свойствами расслоения и связности и может рассматриваться как модель реальных семейств. Число алгоритмов в нём равно 2^m. Алгоритмы допускают от m_1 до $m_1 + m$ ошибок. Ни один из слоёв булева куба не содержится целиком в A, за исключением неинтересного частного случая, когда $m = L$ и A совпадает с булевым кубом. Параметр m может рассматриваться как характеристика сложности или «размерности» данного семейства.

Пессимистичная оценка вероятности переобучения.

Теорема 4.11. *Пусть μ — пессимистичная минимизация эмпирического риска, A — интервал булева куба с m пограничными и m_1 шумовыми объектами. Тогда для любого $\varepsilon \in [0, 1]$ вероятность переобучения есть*

$$Q_\varepsilon = \sum_{s=0}^{m} \sum_{s_1=0}^{m_1} \frac{C_m^s C_{m_1}^{s_1} C_{L-m-m_1}^{\ell-s-s_1}}{C_L^\ell} \big[s_1 + \tfrac{\ell}{L} s \leqslant \tfrac{\ell}{L}(m_1 + m - \varepsilon k) \big].$$

Доказательство. Обозначим через X_0, X_1, S соответственно множества всех внутренних, шумовых и пограничных объектов; а через s_0, s_1, s соответственно — число внутренних, шумовых и пограничных объектов, попавших в обучающую выборку X.

Поскольку метод μ пессимистичный, он всегда будет выбирать из A алгоритм, который не ошибается на всех обучающих пограничных объектах, но ошибается на всех контрольных пограничных объектах. Поэтому

$$\nu(\mu X, X) = \frac{s_1}{\ell}; \qquad \nu(\mu X, \bar{X}) = \frac{(m_1 - s_1) + (m - s)}{k}.$$

Число разбиений $X \sqcup \bar{X}$, при которых $|X_0 \cap X| = s_0,\ |X_1 \cap X| = s_1,\ |S \cap X| = s$, равно $C_{m_0}^{s_0} C_{m_1}^{s_1} C_m^s$. Следовательно, вероятность переобучения представима в виде

$$Q_\varepsilon = \sum_{\substack{s_0=0 \\ s_0+s_1+s=\ell}}^{m_0} \sum_{s_1=0}^{m_1} \sum_{s=0}^{m} \frac{C_{m_0}^{s_0} C_{m_1}^{s_1} C_m^s}{C_L^\ell} \left[\frac{(m_1 - s_1) + (m - s)}{k} - \frac{s_1}{\ell} \geqslant \varepsilon \right].$$

Чтобы получить утверждение теоремы, достаточно воспользоваться соотношениями $m_0 + m_1 + m = L$ и $s_0 + s_1 + s = \ell$ и преобразовать неравенство в квадратных скобках к виду $s_1 + \frac{\ell}{L}s \leqslant \frac{\ell}{L}(m_1 + m - \varepsilon k)$. ■

Следствие 4.11.1. *Если μ — пессимистичная минимизация эмпирического риска, A — интервал булева куба, то вероятность переобучения $Q_\varepsilon(\mu, \mathbb{X})$ совпадает с функционалом равномерной сходимости:*

$$P_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\Big[\max_{a \in A} \delta(a, X, \bar{X}) \geqslant \varepsilon\Big].$$

Доказательство. Это следует из того, что пессимистичный метод μ всегда выбирает из A алгоритм, который не ошибается на всех обучающих пограничных объектах, но ошибается на всех контрольных пограничных объектах. ■

Вычислительный эксперимент. Для проверки полученной формулы и анализа зависимости вероятности переобучения Q_ε от параметра точности ε был выполнен численный эксперимент. Точная оценка из Теоремы 4.11 сравнивалась с результатом эмпирического измерения $\hat{Q}_\varepsilon$ *методом Монте-Карло* по $N = 1000$ случайных разбиений. Графики на рис. 4.2 получены при $\ell = k = 100$ и $m = m_1 = 10$, то есть когда шумовые и пограничные объекты составляют по 5% генеральной выборки.

Точная оценка практически совпадает с эмпирической оценкой для пессимистичной МЭР. Оценка оптимистичной МЭР заметно занижена. Оценка рандомизированной МЭР проходит несколько ближе к пессимистичной.

В эксперименте вычислялась также эмпирическая оценка функционала равномерной сходимости $\hat{P}_\varepsilon$. Согласно следствию 4.11.1, она совпадает с пессимистичной оценкой, и на рис. 4.2 их графики накладываются.

В следующем эксперименте анализировалось влияние эффекта расслоения на вероятность переобучения. Для этого все алгоритмы из A были перенумерованы в порядке неубывания числа ошибок на полной выборке, и вероятность переобучения вычислялась для подсемейств $A_D = \{a_1, \ldots, a_D\}$, где D пробегало целые значения от 1 до 2^m. На рис. 4.3 показана зависимость вероятности переобучения от мощности подсемейства D. Для пессимистичной МЭР вероятность переобучения резко возрастает с переходом на каждый следующий слой, затем «выходит на насыщение». Рандомизированная оценка проходит существенно ниже и растёт более плавно, слабо реагируя на появление новых алгоритмов из более высоких слоёв.

Верхняя кривая на графике соответствует точной оценке для t нижних слоёв, которая будет выведена в следующем параграфе.

Рандомизированная оценка вероятности переобучения. Полученная выше пессимистичная оценка Q_ε представляется завышенной. Она совпадает с функционалом равномерной сходимости, а в эксперименте принимает довольно высокие значения. Рандомизированная МЭР представляется более реалистичной моделью обучения. Поэтому получение точной оценки вероятности переобучения для рандомизированной минимизации эмпирического риска представляется важной задачей.

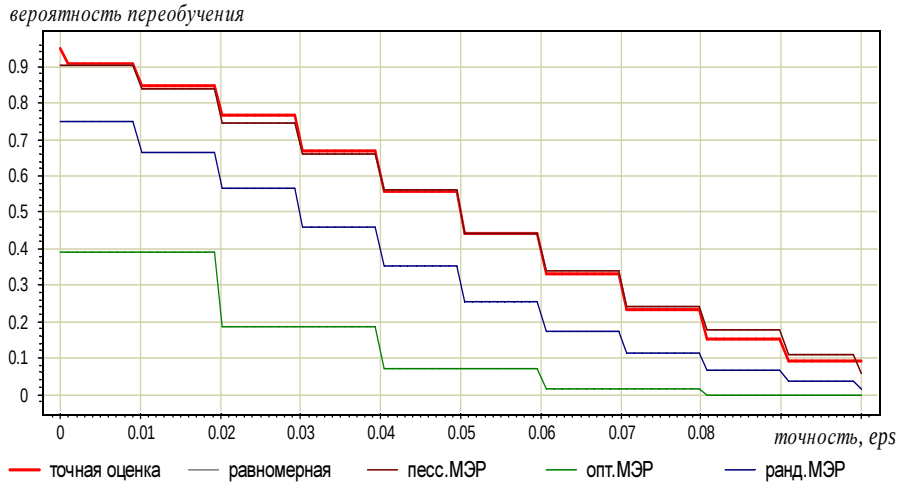


Рис. 4.2. Зависимость оценок вероятности переобучения Q_ε от ε: точная оценка из Теоремы 4.11 и оценки, вычисленные методом Монте-Карло по 1000 случайных разбиений: для пессимистичной, оптимистичной и рандомизированной МЭР. Графики построены при $\ell = k = 100,\ m = m_1 = 10$.

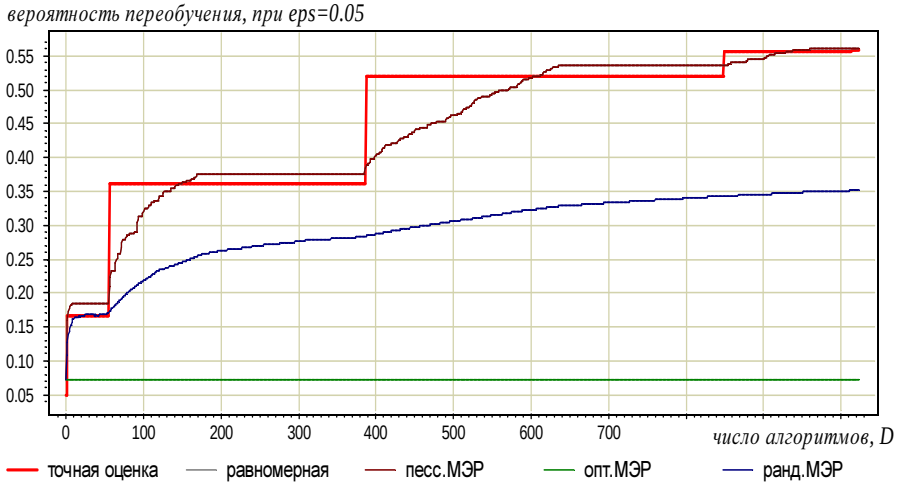


Рис. 4.3. Зависимость вероятности переобучения от числа алгоритмов, взятых из интервала булева куба в порядке возрастания числа ошибок на полной выборке. Графики построены при $\ell = k = 100,\ m = m_1 = 10,\ \varepsilon = 0.05$.

Теорема 4.12. *Пусть μ — рандомизированная минимизация эмпирического риска, A — интервал булева куба с m пограничными и m_1 шумовыми объектами. Тогда для любого $\varepsilon \in [0,1]$ вероятность переобучения* (4.2) *есть*

$$Q_\varepsilon = \sum_{s=0}^{m} \sum_{s_1=0}^{m_1} \frac{C_m^s C_{m_1}^{s_1} C_{L-m-m_1}^{\ell-s-s_1}}{C_L^\ell} \sum_{t=0}^{m-s} \frac{C_{m-s}^t}{2^{m-s}} \big[\, s_1 \leqslant \tfrac{\ell}{L}(m_1 + t - \varepsilon k)\big].$$

Доказательство. Как и в доказательстве предыдущей теоремы, обозначим через X_0, X_1, S соответственно множества всех внутренних, шумовых и пограничных объектов; а через s_0, s_1, s соответственно — число внутренних, шумовых и пограничных объектов, попавших в обучающую выборку X.

Вероятность переобучения для рандомизированного метода μ определяется согласно (4.2):

$$Q_\varepsilon = \mathsf{P}\,\frac{1}{|A(x)|}\sum_{a\in A(X)}\big[\nu(a,\bar{X}) - \nu(a,X) \geqslant \varepsilon\big],$$

где $A(X) = \operatorname{Arg}\min\limits_{a\in A} n(a,X)$.

Все алгоритмы из $A(X)$ допускают одинаковое число ошибок на обучающей выборке, так как для любого $a \in A(X)$, по определению множеств X_0, X_1, S,

$$\begin{aligned}
&n(a, X_0 \cap X) = 0,\\
&n(a, X_1 \cap X) = |X_1 \cap X| = s_1,\\
&n(a,\ \mathrm{S}\ \cap X) = 0.
\end{aligned}$$

Число ошибок t на $S \cap \bar{X}$ может варьироваться от 0 до $|S \cap \bar{X}| = m - s$, поэтому на контрольной выборке алгоритмы из $A(X)$ допускают различное число ошибок:

$$\begin{aligned}
&n(a, X_0 \cap \bar{X}) = 0,\\
&n(a, X_1 \cap \bar{X}) = |X_1 \cap \bar{X}| = m_1 - s_1,\\
&n(a,\ \mathrm{S}\ \cap \bar{X}) = t \in \{0, \ldots, m-s\}.
\end{aligned}$$

Число алгоритмов, допускающих t ошибок на $S \cap \bar{X}$, равно числу способов выбрать t объектов из множества $S \cap \bar{X}$ мощности $m - s$, то есть C^t_{m-s}.

Общее число алгоритмов $|A(X)| = C^0_{m-s} + \cdots + C^{m-s}_{m-s} = 2^{m-s}$.

Для алгоритмов $a \in A(X)$, допускающих t ошибок на $S \cap \bar{X}$,

$$\nu(a,X) = \frac{s_1}{\ell}; \qquad \nu(a,\bar{X}) = \frac{(m_1 - s_1) + t}{k}.$$

Число разбиений $X \sqcup \bar{X}$, при которых $|X_0 \cap X| = s_0,\ |X_1 \cap X| = s_1,\ |S \cap X| = s$, равно $C^{s_0}_{m_0} C^{s_1}_{m_1} C^s_m$.

Итак, вероятность переобучения представима в виде

$$Q_\varepsilon = \sum_{\substack{s_0=0\\ s_0+s_1+s=\ell}}^{m_0}\sum_{s_1=0}^{m_1}\sum_{s=0}^{m} \frac{C^{s_0}_{m_0} C^{s_1}_{m_1} C^s_m}{C^\ell_L}\,\frac{1}{2^{m-s}}\sum_{t=0}^{m-s} C^t_{m-s}\left[\frac{(m_1 - s_1) + t}{k} - \frac{s_1}{\ell} \geqslant \varepsilon\right]. \tag{4.16}$$

Чтобы получить утверждение теоремы, достаточно воспользоваться соотношениями $m_0 + m_1 + m = L$ и $s_0 + s_1 + s = \ell$ и преобразовать неравенство в квадратных скобках к виду $s_1 \leqslant \frac{\ell}{L}(m_1 + t - \varepsilon k)$. ■

Сравнивая оценки, полученные в двух последних теоремах, легко заметить, что Q_ε рандомизированной минимизации эмпирического риска гарантированно не превосходит Q_ε пессимистической минимизации эмпирического риска.

4.2.4 Расслоение интервала булева куба

Рассмотрим снова интервал ранга m в L-мерном булевом кубе, но теперь будем полагать, что множество A состоит только из тех алгоритмов, которые допускают не более t ошибок на пограничных объектах. Другими словами, рассматривается модельное семейство, образованное t нижними слоями интервала булева куба. Число различных векторов ошибок в A равно $C_m^0 + C_m^1 + \cdots + C_m^t$. Алгоритмы допускают от m_1 до $m_1 + t$ ошибок. Параметр t может принимать значения $0, \ldots, m$

Это семейство интересно тем, что оно позволяет исследовать влияние эффекта расслоения на вероятность переобучения, построив зависимость Q_ε от числа нижних слоёв t. В крайнем частном случае $t = 0$ имеем единственный алгоритм, допускающий m_1 ошибок на генеральной выборке. В другом крайнем частном случае $t = m$ имеем весь интервал — этот случай рассмотрен в предыдущем параграфе.

Точные оценки вероятности переобучения для случая произвольного t получаются путём некоторой модификации Теоремы 4.11.

Теорема 4.13. *Пусть μ — пессимистичная минимизация эмпирического риска, A — нижние t слоёв интервала булева куба с m пограничными и m_1 шумовыми объектами. Тогда для любого $\varepsilon \in [0, 1]$ вероятность переобучения есть*

$$Q_\varepsilon = \sum_{s=0}^{m} \sum_{s_1=0}^{m_1} \frac{C_m^s C_{m_1}^{s_1} C_{L-m-m_1}^{\ell-s-s_1}}{C_L^\ell} \left[s_1 \leqslant \tfrac{\ell}{L}(m_1 + \min\{t, m-s\} - \varepsilon k) \right].$$

Доказательство. Обозначим через X_0, X_1, S соответственно множества всех внутренних, шумовых и пограничных объектов; а через s_0, s_1, s соответственно — число внутренних, шумовых и пограничных объектов, попавших в обучающую выборку X.

Поскольку метод μ пессимистичный, он всегда будет выбирать из A алгоритм, который не ошибается на всех обучающих пограничных объектах, но ошибается на всех контрольных пограничных объектах. Поэтому

$$\nu(\mu X, X) = \frac{s_1}{\ell}; \qquad \nu(\mu X, \bar{X}) = \frac{(m_1 - s_1) + \min\{t, m-s\}}{k}.$$

Число разбиений $X \sqcup \bar{X}$, при которых $|X_0 \cap X| = s_0$, $|X_1 \cap X| = s_1$, $|S \cap X| = s$, равно $C_{m_0}^{s_0} C_{m_1}^{s_1} C_m^s$. Следовательно, вероятность переобучения представима в виде

$$Q_\varepsilon = \sum_{\substack{s_0=0 \\ s_0+s_1+s=\ell}}^{m_0} \sum_{s_1=0}^{m_1} \sum_{s=0}^{m} \frac{C_{m_0}^{s_0} C_{m_1}^{s_1} C_m^s}{C_L^\ell} \left[\frac{(m_1 - s_1) + \min\{t, m-s\}}{k} - \frac{s_1}{\ell} \geqslant \varepsilon \right].$$

Чтобы получить утверждение теоремы, достаточно воспользоваться соотношениями $m_0 + m_1 + m = L$ и $s_0 + s_1 + s = \ell$ и преобразовать неравенство в квадратных скобках к виду $s_1 \leqslant \frac{\ell}{L}(m_1 + \min\{t, m-s\} - \varepsilon k)$. ■

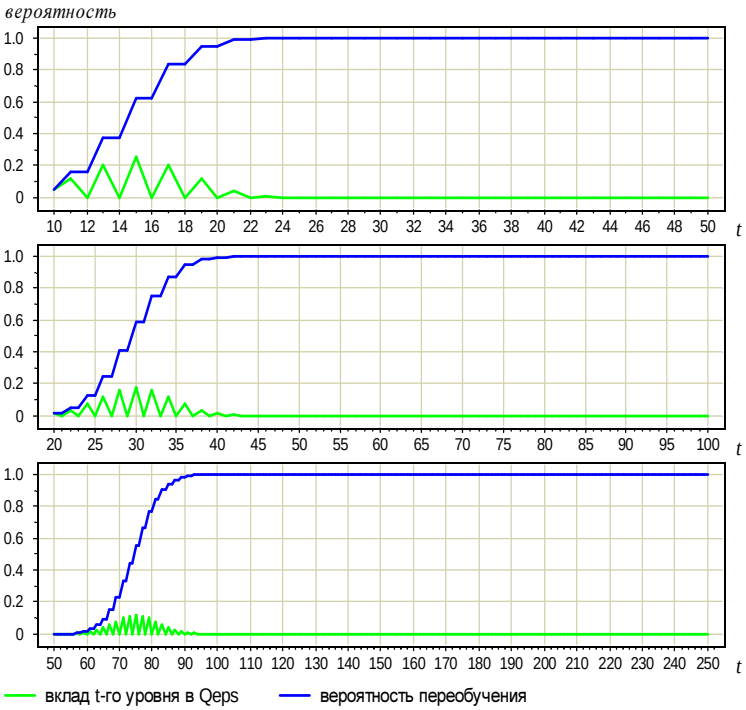


Рис. 4.4. Зависимость вероятности переобучения Q_ε от уровня числа ошибок, при использовании пессимистичной МЭР и $\varepsilon = 0.05$. Верхний график: $\ell = k = 100$, $m_1 = 10$, $m = 40$. Средний график: $\ell = k = 200$, $m_1 = 20$, $m = 80$. Нижний график: $\ell = k = 500$, $m_1 = 50$, $m = 200$.

Вычислительный эксперимент. На рис. 4.4 представлены графики зависимости вероятности переобучения Q_ε от уровня ошибок $t = m_1, \ldots, m_1+m$. Три эксперимента отличались длиной генеральной выборки (200, 400, 1000), при этом сохранялись пропорции $\frac{m}{L} = 0.2$ и $\frac{m_1}{L} = 0.05$, иными словами, генеральная выборка всегда содержала 20% пограничных и 5% шумовых объектов.

На графиках также показаны вклады слоёв в значение функционала Q_ε. Только нижние слои дают ненулевые вклады. Зубчатость графиков вкладов связана с тем, что в силу отношения $\frac{\ell}{L} = \frac{1}{2}$ каждый второй слой вообще не вносит вклад в Q_ε.

В данном эксперименте оказалось, что 20% пограничных объектов — это настолько мощный интервал, что при всех трёх значениях L вероятность переобучения быстро достигает значения 1. Вероятность переобучения близка к нулю только если брать самые нижние слои интервала (не более 2% от длины выборки).

Рандомизированная оценка вероятности переобучения. Итак, вероятность переобучения пессимистичной МЭР в интервале булева куба достаточно высока, даже если брать только нижние слои интервала. Попробуем получить аналогичную оценку для рандомизированной МЭР, в надежде, что она даст лучшие результаты.

Теорема 4.14. *Пусть μ — рандомизированная минимизация эмпирического риска, A — нижние t слоёв интервала булева куба с m пограничными и m_1 шумовыми объ-*

ектами. Тогда для любого $\varepsilon \in [0,1]$ *вероятность переобучения* (4.2) *есть*

$$Q_\varepsilon = \sum_{s=0}^{m} \sum_{s_1=0}^{m_1} \frac{C_m^s C_{m_1}^{s_1} C_{L-m-m_1}^{\ell-s-s_1}}{C_L^\ell} \frac{\sum\limits_{\tau=0}^{\min\{t,m-s\}} C_{m-s}^{\tau} \big[\, s_1 \leqslant \frac{\ell}{L}(m_1+\tau-\varepsilon k)\big]}{\sum\limits_{\tau=0}^{\min\{t,m-s\}} C_{m-s}^{\tau}}.$$

Доказательство. Аналогично доказательству Теоремы 4.12, введём множества X_0, X_1, S и числа s_0, s_1, s. Заметим, что все алгоритмы из $A(X)$ не допускают ошибок на пограничных объектах из $S \cap X$. Число τ их ошибок на $S \cap \bar{X}$ может варьироваться от 0 до $\min\{t, m-s\}$.

Общее число алгоритмов $|A(X)| = C_{m-s}^0 + \cdots + C_{m-s}^{\min\{t,m-s\}}$.

Таким образом, в формуле (4.16), соответствующей полному интервалу булева куба, достаточно сделать лишь два небольших исправления. Во-первых, вместо суммы $\sum\limits_{t=0}^{m-s}$ взять сумму $\sum\limits_{\tau=0}^{\min\{t,m-s\}}$. Во-вторых, общее число алгоритмов 2^{m-s} заменить суммой $\sum\limits_{\tau=0}^{\min\{t,m-s\}} C_{m-s}^{\tau}$. Тогда вероятность переобучения представляется в виде

$$Q_\varepsilon = \sum_{\substack{s_0=0 \\ s_0+s_1+s=\ell}}^{m_0} \sum_{s_1=0}^{m_1} \sum_{s=0}^{m} \frac{C_{m_0}^{s_0} C_{m_1}^{s_1} C_m^s}{C_L^\ell} \frac{\sum\limits_{\tau=0}^{\min\{t,m-s\}} C_{m-s}^{\tau} \big[\frac{(m_1-s_1)+\tau}{k} - \frac{s_1}{\ell} \geqslant \varepsilon\big]}{\sum\limits_{\tau=0}^{\min\{t,m-s\}} C_{m-s}^{\tau}}.$$

Чтобы получить утверждение теоремы, достаточно воспользоваться соотношениями $m_0 + m_1 + m = L$ и $s_0 + s_1 + s = \ell$ и преобразовать неравенство в квадратных скобках к виду $s_1 \leqslant \frac{\ell}{L}(m_1 + \tau - \varepsilon k)$. ■

Вычислительный эксперимент выполнялся при тех же условиях, что и предыдущий, только вместо пессимистичной оценки использовалась рандомизированная, полученная в Теореме 4.14. На рис. 4.5 представлены графики зависимости вероятности переобучения Q_ε от уровня ошибок. Они почти не отличаются от соответствующих графиков на рис. 4.5: кривые рандомизированных оценок немного более гладкие и проходят лишь немного ниже.

Интерпретации и выводы. Приходится констатировать отрицательный, по сути, результат. Интервал булева куба с «разумными» на первый взгляд параметрами m_1, m оказывается слишком богатым семейством алгоритмов. Оценки вероятности переобучения Q_ε для интервала близки к нулю лишь при очень низких значениях m_1, m. Отсюда следуют два вывода.

Во-первых, хорошая обобщающая способность вряд ли возможна, если в выборке есть значительное количество пограничных объектов, на которых алгоритмы семейства допускают ошибки всеми возможными способами. Фактически, доля таких объектов добавляется к величине переобученности, причём рандомизированная оценка практически столь же плоха, как и пессимистическая.

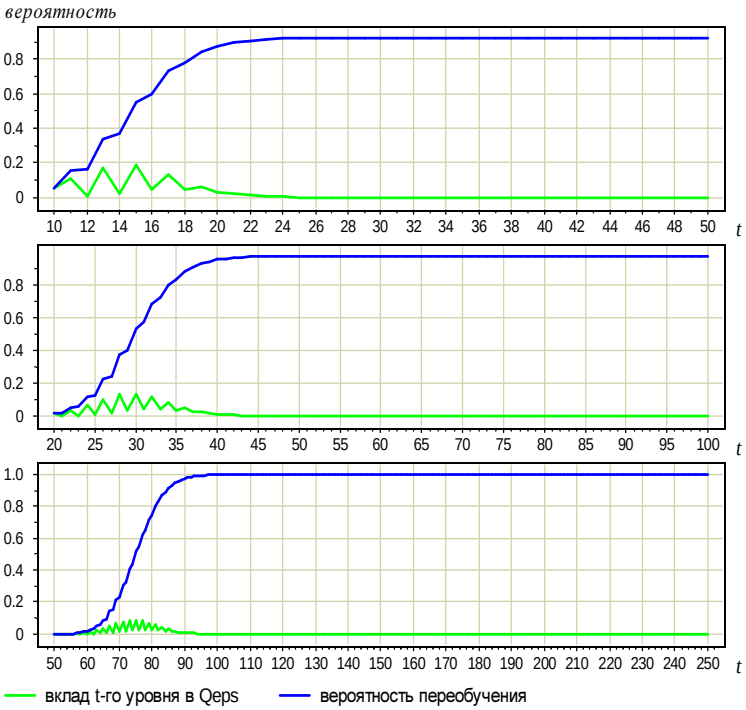


Рис. 4.5. Зависимость вероятности переобучения Q_ε от уровня числа ошибок, при использовании рандомизированной МЭР и $\varepsilon = 0.05$. Верхний график: $\ell = k = 100$, $m_1 = 10$, $m = 40$. Средний график: $\ell = k = 200$, $m_1 = 20$, $m = 80$. Нижний график: $\ell = k = 500$, $m_1 = 50$, $m = 200$.

Во-вторых, использовать интервал булева куба как модель реальных семейств вряд ли целесообразно. Гипотеза о существовании слоя пограничных объектов представляется достаточно разумной. Однако в реальных задачах алгоритмами семейства реализуются, по всей видимости, далеко не все способы допустить ошибки на пограничных объектах. Возможно, более адекватной была бы модель, в которой тем или иным способом вводится характеристика «степени граничности» объектов и оценивается распределение этой характеристики в выборке.

4.2.5 Монотонная цепочка алгоритмов

Монотонная цепочка алгоритмов — это простейшая модель однопараметрического *связного семейства алгоритмов*, предполагающая, что при непрерывном удалении некоторого параметра от оптимального значения число ошибок на полной выборке только увеличивается.

Определим расстояние между алгоритмами как *расстояние Хэмминга* между их векторами ошибок:

$$\rho(a, a') = \sum_{i=1}^{L} \big|I(a, x_i) - I(a', x_i)\big|, \quad \forall a, a' \in A.$$

Определение 4.5. *Множество алгоритмов* $A = \{a_0, a_1, \ldots, a_D\}$ *называется цепочкой алгоритмов, если* $\rho(a_{d-1}, a_d) = 1$ *для всех* $d = 1, \ldots, D$.

В экспериментах [220] было показано, что вероятность переобучения цепочки может оказаться существенно ниже, чем у произвольного несвязного множества алгоритмов с такими же частотами ошибок $\nu(a_d, \mathbb{X})$. Ниже выводятся точные оценки вероятности переобучения для некоторых специальных видов цепочек.

Определение 4.6. *Цепочка алгоритмов $A = \{a_0, a_1, \ldots, a_D\}$ называется монотонной, если $n(a_d, \mathbb{X}) = m + d$ для всех $d = 0, \ldots, D$, при некотором $m \geqslant 0$.*

Алгоритм a_0 называется лучшим в цепочке.

Утверждение 4.15. *Множество алгоритмов $A = \{a_0, a_1, \ldots, a_D\}$ является монотонной цепочкой, если выполнены два условия:*

1) $I(a_{d-1}, x_i) \leqslant I(a_d, x_i)$ для всех $x_i \in \mathbb{X}$, $d = 1, \ldots, D$;

2) $n(a_d, \mathbb{X}) = m + d$ для всех $d = 0, \ldots, D$ при некотором $m \geqslant 0$.

Пример 4.1. Пусть $\mathbb{X}$ — множество точек в $\mathbb{R}^n$; A — семейство *линейных алгоритмов классификации* — параметрических отображений из $\mathbb{X}$ в $\{-1, +1\}$ вида

$$a(x, w) = \operatorname{sign}(x_1 w_1 + \cdots + x_n w_n), \quad x = (x_1, \ldots, x_n) \in \mathbb{R}^n$$

с параметром $w \in \mathbb{R}^n$. Пусть функция потерь имеет вид $I(a, x) = \big[a(x, w) \neq y(x)\big]$, где $y(x)$ — истинная классификация объекта x, и множество объектов $\mathbb{X}$ линейно разделимо, т. е. существует $w^* \in \mathbb{R}^n$, при котором алгоритм $a(x, w^*)$ не допускает ошибок на $\mathbb{X}$. Тогда, при некоторых дополнительных предположениях технического характера, множество алгоритмов $\big\{a(x, w^* + t\delta) \colon t \in [0, +\infty)\big\}$ образует монотонную цепочку для любого направляющего вектора $\delta \in \mathbb{R}^n$, за исключением некоторого конечного множества векторов. При этом $m = 0$.

Теорема 4.16. *Пусть $A = \{a_0, a_1, \ldots, a_D\}$ — монотонная цепочка; $L \geqslant m + D$. Тогда в случае $D \geqslant k$*

$$Q_\varepsilon = \sum_{d=0}^{k} P_d H_{L-d-1}^{\ell-1,m}(s_d(\varepsilon)); \qquad P_d = \frac{C_{L-d-1}^{\ell-1}}{C_L^\ell},\ d = 0, \ldots, k;$$

в случае $D < k$

$$Q_\varepsilon = \sum_{d=0}^{D-1} P_d H_{L-d-1}^{\ell-1,m}(s_d(\varepsilon)) \ + \ P_D H_{L-D}^{\ell,m}(s_D(\varepsilon));$$

$$P_d = \frac{C_{L-d-1}^{\ell-1}}{C_L^\ell},\ d = 0, \ldots, D-1; \qquad P_D = \frac{C_{L-D}^{\ell}}{C_L^\ell},$$

где P_d — вероятность получить алгоритм a_d методом μ; $s_d(\varepsilon) = \frac{\ell}{L}(m + d - \varepsilon k)$.

Доказательство. Перенумеруем объекты таким образом, чтобы каждый из алгоритмов a_d, $d = 1, \ldots, D$ допускал ошибку на объектах $x_1, \ldots, x_d$. Очевидно, лучший алгоритм a_0 не допускает ошибку ни на одном из этих объектов. Нумерация остальных объектов не имеет значения, так как алгоритмы не различимы на них.

Для наглядности представим выборку $\mathbb{X}$ разбитой на три блока:

$$\begin{array}{rccccccccc}
 & x_1 & x_2 & x_3 & & x_D & & & \overbrace{\qquad}^{m} & \\
\vec{a}_0 = (& 0, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & & 1,\ldots,1 &); \\
\vec{a}_1 = (& 1, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & & 1,\ldots,1 &); \\
\vec{a}_2 = (& 1, & 1, & 0, & \ldots & 0, & 0,\ldots,0, & & 1,\ldots,1 &); \\
\vec{a}_3 = (& 1, & 1, & 1, & \ldots & 0, & 0,\ldots,0, & & 1,\ldots,1 &); \\
\ldots & & & & \ldots & & \ldots & & \ldots & \\
\vec{a}_D = (& 1, & 1, & 1, & \ldots & 1, & 0,\ldots,0, & & 1,\ldots,1 &);
\end{array}$$

При рассмотрении алгоритма a_d возможны три случая.

1. Если $k < d$, то число ошибок алгоритма a_d на объектах $\{x_1, \ldots, x_d\}$ превышает длину контрольной выборки. Часть ошибок обязательно окажется в обучающей подвыборке X, и метод μ выберет другой алгоритм. В этом случае

$$[\mu X{=}a_d] = 0.$$

2. Если $d = D < k$, то метод μ выберет наихудший алгоритм в цепочке a_D тогда и только тогда, когда все объекты $\{x_1, \ldots, x_D\}$ будут находиться в контрольной подвыборке $\bar{X}$. В этом случае

$$[\mu X{=}a_d] = [x_1, \ldots, x_D \in \bar{X}].$$

3. Во всех остальных случаях метод μ выберет алгоритм a_d, если только все объекты $\{x_1, \ldots, x_d\}$ будут находиться в контрольной подвыборке $\bar{X}$, а объект x_{d+1} — в обучающей подвыборке X. В этом случае

$$[\mu X{=}a_d] = [x_{d+1} \in X][x_1, \ldots, x_d \in \bar{X}].$$

Теперь можно применить Теорему 4.3.

Если $D \geqslant k$, то алгоритму a_d соответствуют следующие значения параметров (для упрощения обозначений вместо двойных индексов L_{a_d} будем использовать одинарные L_d): $L_d = L - d - 1$, $\ell_d = \ell - 1$, $m_d = m + d - d = m$, $s_d(\varepsilon) = \frac{\ell}{L}(m + d - \varepsilon k)$. Отсюда получаем утверждение теоремы для случая $D \geqslant k$.

Если $D < k$, то алгоритмам $a_0, \ldots, a_{D-1}$ соответствуют те же значения параметров, что и при $D \geqslant k$. Для наихудшего алгоритма a_D отличается только параметр $\ell_D = \ell$. Отсюда получаем утверждение теоремы для случая $D < k$. ■

Замечание 4.2. В ходе доказательства полезно проверить, что вероятности P_d вычислены корректно и в сумме дают единицу. Для случая $D \geqslant k$ проверка сводится к применению известного комбинаторного тождества:

$$\sum_{d=0}^{D} P_d = \sum_{d=0}^{k} P_d + \sum_{d=k+1}^{D} 0 = \frac{1}{C_L^\ell}\big(C_{L-1}^{\ell-1} + C_{L-2}^{\ell-1} + \cdots + C_{\ell-1}^{\ell-1}\big) = 1.$$

Для случая $D < k$ то же самое тождество приходится применить дважды, заметив, что $C_{L-D}^{\ell} = C_{L-D-1}^{\ell-1} + \cdots + C_{\ell-1}^{\ell-1}$:

$$\sum_{d=0}^{D} P_d = \frac{1}{C_L^{\ell}}\left(C_{L-1}^{\ell-1} + \cdots + C_{L-D}^{\ell-1} + C_{L-D}^{\ell}\right) = 1.$$

Вычислительный эксперимент. Для проверки полученной формулы и анализа зависимости вероятности переобучения Q_ε от параметра точности ε и числа алгоритмов в цепочке D был выполнен численный эксперимент. Точная оценка из Теоремы 4.16 сравнивалась с результатом эмпирического измерения $\hat{Q}_\varepsilon$ *методом Монте-Карло* по $N = 1000$ случайных разбиений. Показанные на рис. 4.6 экспериментальные результаты получены при $\ell = k = 100$ и $m = 20$, то есть когда лучший алгоритм в цепочке допускает 10% ошибок на полной выборке.

Оказалось, что для монотонной цепочки пессимистичная и рандомизированная МЭР дают почти одинаковые оценки Q_ε, а оптимистичная МЭР даёт заметно заниженную оценку, см. левый график на рис. 4.6. Это означает, что для данного семейства пессимистичная оценка почти не завышена и является «более разумной» по сравнению с оптимистичной.

В эксперименте вычислялась также эмпирическая оценка функционала равномерной сходимости,

$$\hat{P}_\varepsilon = \hat{\mathsf{P}}\left[\max_{a\in A} \delta(a, X, \bar{X}) \geqslant \varepsilon\right].$$

Он является завышенной верхней оценкой вероятности переобучения, $Q_\varepsilon \leqslant P_\varepsilon$, см. стр. 60. Правый график на рис. 4.6 убедительно показывает, что эта оценка может быть завышенной в сотни раз. Хотя, при малых значениях точности ε завышенность $\hat{P}_\varepsilon$ не столь катастрофична.

Левый график на рис. 4.7 показывает, что с ростом числа алгоритмов в монотонной цепочке функционал равномерной сходимости P_ε продолжает возрастать, тогда как вероятность переобучения Q_ε после 5–8 алгоритмов выходит на горизонтальную асимптоту. Согласно Теореме 1.14, оценка равномерной сходимости завышена из-за того, что она не учитывает эффект расслоения. Поэтому кривую $\hat{P}_\varepsilon$ (верхняя кривая на левом графике рис. 4.7) можно рассматривать как оценку вероятности переобучения для *цепочки без расслоения*. Только совместное проявление эффектов расслоения и связности понижает вероятность переобучения до приемлемо малых значений. Этот же вывод был сделан и в 3.3.2 по результатам экспериментов на рандомизированных цепочках.

Вкладом $Q_\varepsilon(a)$ алгоритма a в вероятность переобучения Q_ε будем называть слагаемое под знаком суммы $\sum\limits_{a\in A}$ в общей формуле вероятности переобучения (4.10):

$$Q_\varepsilon = \sum_{a\in A} \underbrace{\sum_{v\in V_a} c_{av} P_{av} H_{L_{av}}^{\ell_{av},m_{av}}\left(s_{av}(\varepsilon)\right)}_{Q_\varepsilon(a)}.$$

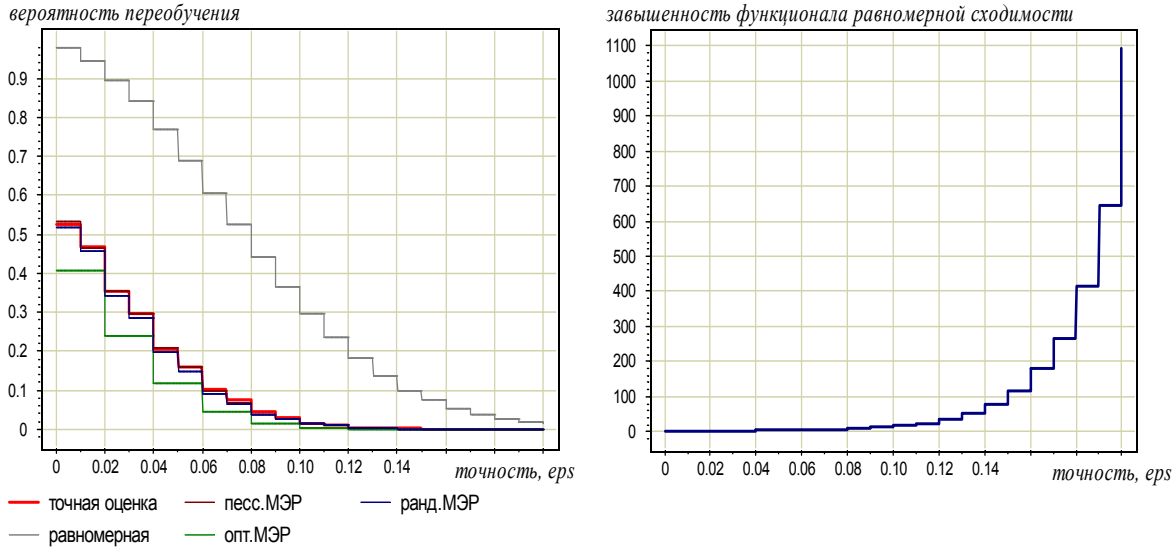


Рис. 4.6. Слева: зависимость оценок вероятности переобучения Q_ε от ε: точная оценка из Теоремы 4.16 и четыре оценки, вычисленные методом Монте-Карло по 1000 случайных разбиений: для пессимистичной, оптимистичной и рандомизированной МЭР. Верхняя кривая соответствует оценке по функционалу равномерной сходимости $\hat{P}_\varepsilon$. Справа: степень завышенности функционала равномерной сходимости $\hat{P}_\varepsilon/Q_\varepsilon$. Все графики построены при $\ell = k = 100,\ m = 20$.

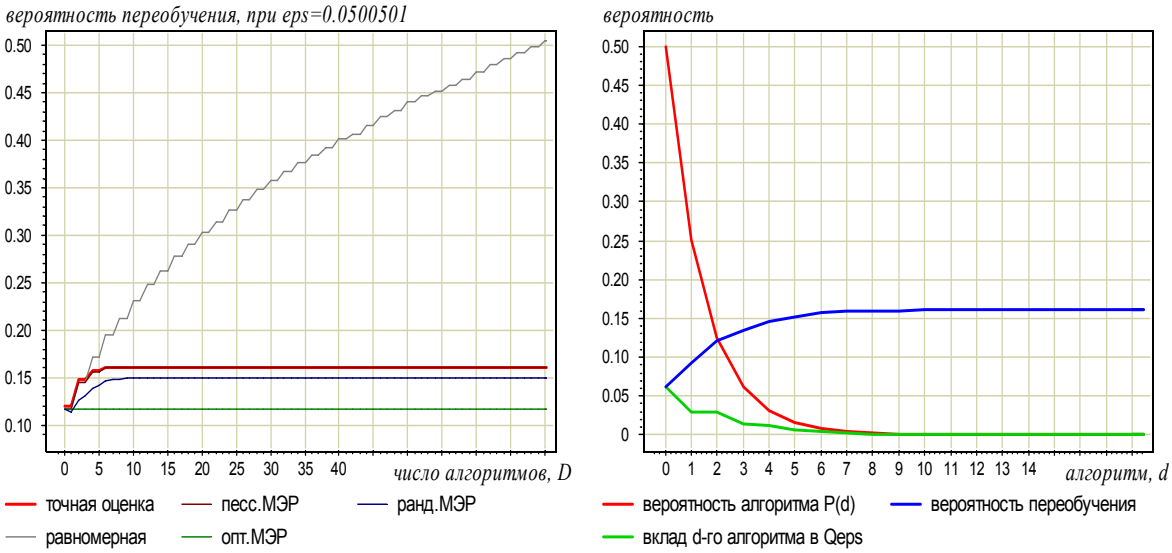


Рис. 4.7. Слева: зависимость оценок вероятности переобучения Q_ε от числа алгоритмов в цепочке D. Справа: вероятность получения каждого из алгоритмов $P_d = \mathsf{P}[\mu X{=}a_d]$, вклад каждого алгоритма в вероятность переобучения Q_ε, значение Q_ε для пессимистичной МЭР по подмножеству алгоритмов $\{a_0, \ldots, a_d\}$ как функция от числа алгоритмов d. Все графики построены при $\ell = k = 100,\ m = 20$, $\varepsilon = 0.05$.

Рис. 4.7 (справа) показывает, что существенные вклады в вероятность переобучения вносят только алгоритмы 5–8 нижних уровней. По всей видимости, аналогичный вывод справедлив не только для монотонных цепочек, но и для многих других семейств алгоритмов.

Основной вывод заключается в том, что монотонная цепочка алгоритмов почти не переобучается. Этот факт служит обоснованием для процедур одномерной оптимизации, которые часто применяются в машинном обучении для выбора некоторого критически важного параметра по отложенной выборке (hold-out model selection). Например, это выбор константы регуляризации, выбор ширины окна сглаживания, выбор числа нейронов в скрытом слое нейронной сети, выбор степени аппроксимирующего полинома, и так далее.

4.2.6 Унимодальная цепочка алгоритмов

Унимодальная цепочка является более реалистичной моделью однопараметрического *связного семейства*, по сравнению с монотонной цепочкой. Если мы имеем лучший алгоритм a_0 с оптимальным значением некоторого вещественного параметра, то отклонение значения этого параметра как в бóльшую, так и в меньшую, сторону будет приводить, как правило, к увеличению числа ошибок.

Определение 4.7. *Множество алгоритмов* $A = \{a_0, a_1, \ldots, a_D, a'_1, \ldots, a'_{D'}\}$ *называется унимодальной цепочкой, если левая ветвь* $a_0, a_1, \ldots, a_D$ *и правая ветвь* $a_0, a'_1, \ldots, a'_{D'}$ *являются монотонными цепочками. Алгоритм* a_0 *называется лучшим в цепочке.*

Пример 4.2 (продолжение примера 4.1). Пусть множество объектов $\mathbb{X} \subset \mathbb{R}^n$ линейно разделимо, т. е. существует *линейный алгоритм классификации* $a(x, w^*)$ с параметром $w^* \in \mathbb{R}^n$, не допускающий ошибок на $\mathbb{X}$. Тогда множество алгоритмов $\big\{a(x, w^* + t\delta) \colon t \in \mathbb{R}\big\}$ образует унимодальную цепочку для почти любого направляющего вектора $\delta \in \mathbb{R}^n$.

Обозначим через $m = n(a_0, \mathbb{X})$ число ошибок лучшего алгоритма.

Рассмотрим унимодальную цепочку с ветвями равной длины, $D = D'$. Перенумеруем объекты так, чтобы каждый из алгоритмов a_d, $d = 1, \ldots, D$ допускал ошибку на объектах $x_1, \ldots, x_d$; а каждый из алгоритмов a'_d, $d = 1, \ldots, D$ допускал ошибку на объектах $x'_1, \ldots, x'_d$. Будем предполагать, что множества объектов $\{x_1, \ldots, x_D\}$ и $\{x'_1, \ldots, x'_D\}$ не пересекаются. Очевидно, лучший алгоритм a_0 не допускает ошибку ни на одном из этих объектов. Нумерация остальных объектов не имеет значения, так как алгоритмы не различимы на них.

Для наглядности представим выборку $\mathbb{X}$ разбитой на четыре блока:

$$
\begin{array}{rccccccccccccc}
 & x_1 & x_2 & x_3 & & x_D & x'_1 & x'_2 & x'_3 & & x'_D & & \overbrace{\qquad}^{m} & \\
\vec{a}_0 = (& 0, & 0, & 0, & \ldots & 0, & 0, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\vec{a}_1 = (& 1, & 0, & 0, & \ldots & 0, & 0, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\vec{a}_2 = (& 1, & 1, & 0, & \ldots & 0, & 0, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\vec{a}_3 = (& 1, & 1, & 1, & \ldots & 0, & 0, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\ldots & & & & \ldots & & & & & \ldots & & \ldots & \ldots & \\
\vec{a}_D = (& 1, & 1, & 1, & \ldots & 1, & 0, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\vec{a}'_1 = (& 0, & 0, & 0, & \ldots & 0, & 1, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\vec{a}'_2 = (& 0, & 0, & 0, & \ldots & 0, & 1, & 1, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\vec{a}'_3 = (& 0, & 0, & 0, & \ldots & 0, & 1, & 1, & 1, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\ldots & & & & \ldots & & & & & \ldots & & \ldots & \ldots & \\
\vec{a}'_D = (& 0, & 0, & 0, & \ldots & 0, & 1, & 1, & 1, & \ldots & 1, & 0,\ldots,0, & 1,\ldots,1 &);
\end{array}
$$

Будем полагать, что если минимум (4.1) достигается на нескольких алгоритмах с одинаковым числом ошибок как на обучающей, так и на генеральной выборке, то метод μ выбирает алгоритм из левой ветви.

Теорема 4.17. *Пусть $A = \{a_0, a_1, \ldots, a_D, a'_1, \ldots, a'_D\}$ — унимодальная цепочка, $k \leqslant D$ и $2D + m \leqslant L$. Тогда вероятность получить каждый из алгоритмов цепочки в результате обучения есть*

$$
\begin{aligned}
P_0 &= \mathsf{P}\big[\mu X{=}a_0\big] = \frac{C_{L-2}^{\ell-2}}{C_L^\ell}; \\
P_d &= \mathsf{P}\big[\mu X{=}a_d\big] = \frac{C_{L-d-1}^{\ell-1} - C_{L-2d-2}^{\ell-1}}{C_L^\ell}; \\
P'_d &= \mathsf{P}\big[\mu X{=}a'_d\big] = \frac{C_{L-d-1}^{\ell-1} - C_{L-2d-1}^{\ell-1}}{C_L^\ell};
\end{aligned}
$$

вероятность переобучения при $s_d(\varepsilon) = \frac{\ell}{L}(m + d - \varepsilon k)$ выражается в виде

$$
\begin{aligned}
Q_\varepsilon = \frac{C_{L-2}^{\ell-2}}{C_L^\ell} H_{L-2}^{\ell-2,m}\left(s_0(\varepsilon)\right) + \sum_{d=1}^{k} \Bigg(& 2\frac{C_{L-d-1}^{\ell-1}}{C_L^\ell} H_{L-d-1}^{\ell-1,m}\left(s_d(\varepsilon)\right) - \\
& - \frac{C_{L-2d-2}^{\ell-1}}{C_L^\ell} H_{L-2d-2}^{\ell-1,m}\left(s_d(\varepsilon)\right) - \frac{C_{L-2d-1}^{\ell-1}}{C_L^\ell} H_{L-2d-1}^{\ell-1,m}\left(s_d(\varepsilon)\right) \Bigg).
\end{aligned}
$$

Доказательство. Введём вспомогательные переменные:

$$
\begin{aligned}
&\beta_d = [x_{d+1} \in X][x_1, \ldots, x_d \in \bar{X}], \quad d = 1, \ldots, D-1; \\
&\beta_D = [x_1, \ldots, x_D \in \bar{X}]; \\
&\beta'_d = [x'_{d+1} \in X][x'_1, \ldots, x'_d \in \bar{X}], \quad d = 1, \ldots, D-1; \\
&\beta'_D = [x'_1, \ldots, x'_D \in \bar{X}].
\end{aligned}
$$

Условия $\beta_1, \ldots, \beta_D$ несовместны, причём одно из них выполнено тогда и только тогда, когда $x_1 \in \bar{X}$. Следовательно,

$$[x_1 \in X] + \beta_1 + \cdots + \beta_D = 1.$$

Аналогично,

$$[x'_1 \in X] + \beta'_1 + \cdots + \beta'_D = 1.$$

Если бы левая и правая ветви рассматривать как отдельные монотонные цепочки, то можно было бы утверждать, что $[\mu X{=}a_d] = \beta_d$ и $[\mu X{=}a'_d] = \beta'_d$. Однако в случае унимодальной цепочки условия получения алгоритмов a_d, a'_d приобретают более сложный вид. Если выполнено условие β_d и одновременно одно из условий $\beta'_{d+1}, \ldots, \beta'_D$, то метод μ выберет один из алгоритмов $a'_{d+1}, \ldots, a'_D$ из правой ветви, согласно договорённости, что выбирается наихудший алгоритм из всех, допускающих одинаковое наименьшее число ошибок на X. Аналогично, если выполнено условие β'_d и одновременно одно из условий $\beta_d, \ldots, \beta_D$, то метод μ выберет один из алгоритмов $a_d, \ldots, a_D$ из левой ветви. Обратим внимание, что алгоритмы левой ветви имеют приоритет. Таким образом, условия получения всех алгоритмов унимодальной цепочки выражаются через вспомогательные переменные:

$$\begin{aligned}
&[\mu X{=}a_0] = [x_1, x'_1 \in X] = \big(1 - \beta_1 - \cdots - \beta_D\big)\big(1 - \beta'_1 - \cdots - \beta'_D\big);\\
&[\mu X{=}a_d] = \beta_d\big(1 - \beta'_{d+1} - \cdots - \beta'_D\big), \quad d = 1, \ldots, D-1;\\
&[\mu X{=}a'_d] = \beta'_d\big(1 - \beta_d - \cdots - \beta_D\big), \quad d = 1, \ldots, D-1;\\
&[\mu X{=}a_D] = \beta_D;\\
&[\mu X{=}a'_D] = \beta'_D\big(1 - \beta_D\big).
\end{aligned}$$

Непосредственной подстановкой нетрудно убедиться, что тождество (4.4) выполнено, следовательно, совокупность условий $[\mu X{=}a]$ определена корректно.

Найдём вероятности всех алгоритмов цепочки, применив Лемму 4.5.

$$\begin{aligned}
P_0 = \mathsf{P}\big[\mu X{=}a_0\big] &= \mathsf{P}[x_1, x'_1 \in X] = \frac{C_{L-2}^{\ell-2}}{C_L^\ell};\\
P_d = \mathsf{P}\big[\mu X{=}a_d\big] &= \mathsf{P}[x_{d+1} \in X][x_1, \ldots, x_d \in \bar{X}] -\\
&\quad - \sum_{t=d+1}^{k-d} \mathsf{P}[x_{d+1}, x'_{t+1} \in X][x_1, \ldots, x_d, x'_1, \ldots, x'_t \in \bar{X}] =\\
&= \frac{1}{C_L^\ell}\left(C_{L-d-1}^{\ell-1} - \sum_{t=d+1}^{k-d} C_{L-d-t-2}^{\ell-2}\right) = \frac{C_{L-d-1}^{\ell-1} - C_{L-2d-2}^{\ell-1}}{C_L^\ell};\\
P'_d = \mathsf{P}\big[\mu X{=}a'_d\big] &= \mathsf{P}[x'_{d+1} \in X][x'_1, \ldots, x'_d \in \bar{X}] -\\
&\quad - \sum_{t=d}^{k-d} \mathsf{P}[x'_{d+1}, x_{t+1} \in X][x'_1, \ldots, x'_d, x_1, \ldots, x_t \in \bar{X}] =\\
&= \frac{1}{C_L^\ell}\left(C_{L-d-1}^{\ell-1} - \sum_{t=d}^{k-d} C_{L-d-t-2}^{\ell-2}\right) = \frac{C_{L-d-1}^{\ell-1} - C_{L-2d-1}^{\ell-1}}{C_L^\ell}.
\end{aligned}$$

Теперь запишем вероятность переобучения, применив Теорему 4.6.

$$Q_\varepsilon = \frac{C_{L-2}^{\ell-2}}{C_L^\ell} H_{L-2}^{\ell-2,m}\left(s_0(\varepsilon)\right) + \\ + \sum_{d=1}^{k}\left(\frac{C_{L-d-1}^{\ell-1}}{C_L^\ell} H_{L-d-1}^{\ell-1,m}\left(s_d(\varepsilon)\right) - \sum_{t=d+1}^{k-d} \frac{C_{L-d-t-2}^{\ell-2}}{C_L^\ell} H_{L-d-t-2}^{\ell-2,m}\left(s_d(\varepsilon)\right)\right) + \\ + \sum_{d=1}^{k}\left(\frac{C_{L-d-1}^{\ell-1}}{C_L^\ell} H_{L-d-1}^{\ell-1,m}\left(s_d(\varepsilon)\right) - \sum_{t=d}^{k-d} \frac{C_{L-d-t-2}^{\ell-2}}{C_L^\ell} H_{L-d-t-2}^{\ell-2,m}\left(s_d(\varepsilon)\right)\right).$$

Полученное выражение можно упростить, заметив, что

$$\sum_{t=d+1}^{k-d} \frac{C_{L-d-t-2}^{\ell-2}}{C_L^\ell} H_{L-d-t-2}^{\ell-2,m}\left(s_d(\varepsilon)\right) = \frac{C_{L-2d-2}^{\ell-1}}{C_L^\ell} H_{L-2d-2}^{\ell-1,m}\left(s_d(\varepsilon)\right);$$

$$\sum_{t=d}^{k-d} \frac{C_{L-d-t-2}^{\ell-2}}{C_L^\ell} H_{L-d-t-2}^{\ell-2,m}\left(s_d(\varepsilon)\right) = \frac{C_{L-2d-1}^{\ell-1}}{C_L^\ell} H_{L-2d-1}^{\ell-1,m}\left(s_d(\varepsilon)\right);$$

Подставляя эти выражения в формулу для Q_ε, получим требуемую оценку. ■

Замечание 4.3. Нетрудно убедиться, что вероятности P_d, P'_d найдены корректно:

$$P_0 + \sum_{d=1}^{D}\left(P_d + P'_d\right) = \frac{C_{L-2}^{\ell-2}}{C_L^\ell} + \sum_{d=1}^{D} \frac{C_{L-d-1}^{\ell-1} - C_{L-2d-2}^{\ell-1}}{C_L^\ell} + \frac{C_{L-d-1}^{\ell-1} - C_{L-2d-1}^{\ell-1}}{C_L^\ell} = \\ = \frac{C_{L-2}^{\ell-2}}{C_L^\ell} + \frac{1}{C_L^\ell}\sum_{d=1}^{D}\left(2C_{L-d-1}^{\ell-1} - C_{L-2d-2}^{\ell-1} - C_{L-2d-1}^{\ell-1}\right) = \\ = \frac{1}{C_L^\ell}\left(C_{L-2}^{\ell-2} + 2\left(C_{L-2}^{\ell-1} + \cdots + C_{\ell-1}^{\ell-1}\right) - \left(C_{L-3}^{\ell-1} + \cdots + C_{\ell-1}^{\ell-1}\right)\right) = \\ = \frac{1}{C_L^\ell}\left(C_{L-2}^{\ell-2} + 2C_{L-1}^{\ell} - C_{L-2}^{\ell}\right) = 1.$$

4.2.7 Единичная окрестность лучшего алгоритма

Другим примером *связного семейства* является единичная окрестность лучшего алгоритма. Это искусственная постановка задачи, но она интересна по двум причинам. Во-первых, это «экстремальный» случай, когда алгоритмы максимально близки друг к другу, и классические оценки, основанные на неравенстве Буля, максимально завышены. Во-вторых, это первый шаг на пути к общим точным оценкам вероятности переобучения. Следующим шагом должно стать увеличение радиуса окрестности.

Определение 4.8. *Множество алгоритмов $A = \{a_0, a_1, \ldots, a_D\}$ называется единичной окрестностью алгоритма a_0, если все векторы ошибок a_d попарно различны, $n(a_d, \mathbb{X}) = n(a_0, \mathbb{X}) + 1$ и $\rho(a_0, a_d) = 1$ для всех $d = 1, \ldots, D$. Алгоритм a_0 называется лучшим в окрестности или центром окрестности.*

Будем полагать, что если минимум (4.1) достигается на нескольких алгоритмах с одинаковым числом ошибок как на обучающей, так и на генеральной выборке, то метод μ выбирает алгоритм с меньшим номером.

Теорема 4.18. *Пусть $A = \{a_0, a_1, \ldots, a_D\}$ — единичная окрестность алгоритма a_0; $m = n(a_0, \mathbb{X})$; $L \geqslant m + D$. Тогда*

$$Q_\varepsilon = P_0 H_{L-D}^{\ell-D,m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right) + \sum_{d=1}^{D} P_d H_{L-d}^{\ell-d+1,m}\left(\tfrac{\ell}{L}(m+1-\varepsilon k)\right);$$

$$P_0 = \frac{C_{L-D}^k}{C_L^k}; \quad P_d = \frac{C_{L-d}^{k-1}}{C_L^k},\ d = 1, \ldots, D;$$

где P_d — вероятность получить алгоритм a_d в результате обучения.

Доказательство. Перенумеруем объекты таким образом, чтобы каждый из алгоритмов a_d, $d = 1, \ldots, D$ допускал ошибку на объекте x_d. Очевидно, лучший алгоритм a_0 не допускает ошибку ни на одном из этих объектов. Нумерация остальных объектов не имеет значения, так как алгоритмы не различимы на них.

Для наглядности представим выборку $\mathbb{X}$ разбитой на три блока:

$$\begin{array}{rccccccccc}
 & x_1 & x_2 & x_3 & & x_D & & \overbrace{\qquad}^{m} & \\
\vec{a}_0 = (& 0, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\vec{a}_1 = (& 1, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\vec{a}_2 = (& 0, & 1, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\vec{a}_3 = (& 0, & 0, & 1, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 &); \\
\ldots & & & & \ldots & & \ldots & \ldots & \\
\vec{a}_D = (& 0, & 0, & 0, & \ldots & 1, & 0,\ldots,0, & 1,\ldots,1 &);
\end{array}$$

Нетрудно видеть, что множество разбиений, при которых метод μ выбирает алгоритм a_d, представляется в следующем виде:

$$\begin{aligned}
&\big[\mu X{=}a_0\big] = \big[x_1, \ldots, x_D \in X\big]; \\
&\big[\mu X{=}a_d\big] = \big[x_1, \ldots, x_{d-1} \in X\big]\big[x_d \in \bar{X}\big], \quad d = 1, \ldots, D.
\end{aligned}$$

Параметры для подстановки в формулу Теоремы 4.3:

$$\begin{aligned}
&L_0 = L - D; \quad \ell_0 = \ell - D; \quad m_0 = m; \quad s_0(\varepsilon) = \tfrac{\ell}{L}(m - \varepsilon k); \\
&L_d = L - d; \quad \ell_d = \ell - d + 1; \quad m_d = m; \quad s_d(\varepsilon) = \tfrac{\ell}{L}(m + 1 - \varepsilon k); \quad d = 1, \ldots, D.
\end{aligned}$$

Подставляя эти параметры в формулу Теоремы 4.3, получаем требуемое. ■

Замечание 4.4. Нетрудно убедиться, что вероятности P_d найдены корректно:

$$\sum_{d=0}^{D} P_d = \frac{1}{C_L^k}\big(C_{L-D}^k + \underbrace{C_{L-D}^{k-1} + \cdots + C_{L-1}^{k-1}}_{C_L^k - C_{L-D}^k}\big) = 1.$$

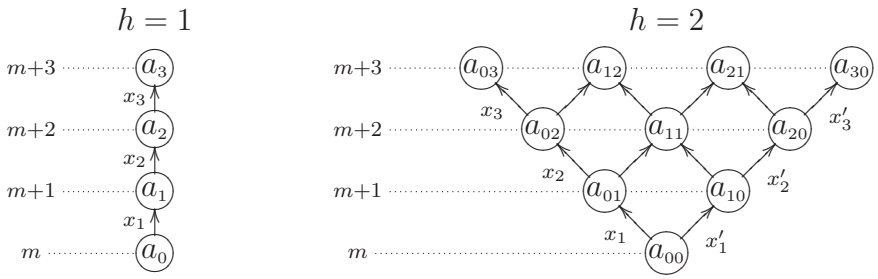


Рис. 4.8. Слева: одномерная монотонная сетка высоты 3 (монотонная цепочка). Справа: двумерная монотонная сетка высоты 3. Горизонтальные линии отмечают уровни числа ошибок $n(a_J, \mathbb{X})$ на генеральной выборке. Объекты вдоль стрелок означают, что переход от нижнего алгоритма к верхнему связан с появлением ошибки на данном объекте.

4.2.8 О некоторых других модельных семействах

В данном обзорном параграфе кратко перечисляются модельные семейства, для которых другими авторами получены точные оценки вероятности переобучения в рамках предлагаемого в данной работе комбинаторного подхода.

Многомерные монотонные и унимодальные сетки являются естественным обобщением монотонных и унимодальных цепочек алгоритмов. Они моделируют многомерные параметрические семейства алгоритмов с расслоением и связностью.

Определение 4.9. *Монотонная h-мерная сетка высоты H — это множество алгоритмов $A = \{a_J \colon J = (j_1, \ldots, j_h) \in \{0, \ldots, L\}^h;\ j_1 + \cdots + j_h \leqslant H\}$, для которого выполняются два условия:*

1) если $J \prec J'$, то $I(a_J, x_i) \leqslant I(a_{J'}, x_i)$ для всех $x_i \in \mathbb{X}$, где отношение $J \prec J'$ определяется следующим образом: $j_s \leqslant j'_s$ для всех $s = 1, \ldots, h$;

2) $n(a_J, \mathbb{X}) = m + j_1 + \cdots + j_h$ для всех $a_J \in A$ при некотором $m \geqslant 0$.

Определение 4.10. *Унимодальная h-мерная сетка высоты H — это множество алгоритмов $A = \{a_J \colon J = (j_1, \ldots, j_h) \in \{-L, \ldots, 0, \ldots, L\}^h;\ |j_1| + \cdots + |j_h| \leqslant H\}$, для которого выполняются два условия:*

1) если $J \prec J'$, то $I(a_J, x_i) \leqslant I(a_{J'}, x_i)$ для всех $x_i \in \mathbb{X}$, где отношение $J \prec J'$ определяется следующим образом: $|j_s| \leqslant |j'_s|$ и j_s, j'_s одного знака для всех $s = 1, \ldots, h$;

2) $n(a_J, \mathbb{X}) = m + |j_1| + \cdots + |j_h|$ для всех $a_J \in A$ при некотором $m \geqslant 0$.

Согласно данным определениям, h-мерная унимодальная сетка образуется 2^h монотонными сетками той же размерности, имеющими один общий алгоритм $a_{(0,\ldots,0)}$, который является лучшим алгоритмом в A.

На рис. 4.8 показаны примеры одномерной и двумерной монотонной сетки.

Точные оценки вероятности переобучения для h-мерных монотонных и унимодальных сеток получены П. Ботовым в [7]. Показано, что достаточно точные оценки можно получать, используя лишь несколько нижних слоёв сеток. Довольно неожиданным оказался тот факт, что вероятность переобучения h-мерных унимодальных

сеток неплохо аппроксимируется $2h$-мерными монотонными сетками. Экспериментально показано, что вероятность переобучения реальных семейств алгоритмов можно аппроксимировать монотонными сетками подходящей размерности h. В этом эксперименте использовались стандартные методы классификации: наивный байесовский классификатор, решающие деревья, нейронные сети. Данный факт позволяет вводить понятие эффективной размерности, которая характеризует структуру расслоения и связности локальной окрестности лучшего алгоритма.

Симметричные семейства алгоритмов предложил А. Фрей в [83]. Учёт симметрии множества алгоритмов и привлечение теоретико-групповых соображений позволяет получать более эффективные формулы для вероятности переобучения в случаях, когда используется рандомизированный метод обучения. Получены общие оценки для семейства с произвольной группой симметрий. Рассмотрен частный случай связки монотонных цепочек, которую можно рассматривать как ещё одну модель многомерного семейства, обладающего свойствами расслоения и связности. *Связкой из p монотонных цепочек* называется множество алгоритмов, полученное объединением p штук монотонных цепочек равной длины H («ветвей») с общим лучшим алгоритмом a_0 при условии, что множества объектов, на которых ошибаются алгоритмы ветвей, не пересекаются. Связка является естественным обобщением монотонных и унимодальных цепочек, а также единичной окрестности, которые были рассмотрены в предыдущих параграфах данной главы.

4.3 Рекуррентное вычисление вероятности переобучения

Допустим, что векторы ошибок всех алгоритмов из множества A известны, попарно различны, и в A есть корректный на $\mathbb{X}$ алгоритм. Пусть μ — пессимистичный метод минимизации эмпирического риска. Задача заключается в том, чтобы для каждого алгоритма $a \in A$ найти всю информацию, необходимую для вычисления вероятности переобучения Q_ε по Теореме 4.6:

$$\mathfrak{I}(a) = \langle X_{av}, X'_{av}, c_{av} \rangle_{v \in V_a},$$

где V_a — индексное множество, X_{av} — множество порождающих объектов, X'_{av} — множество запрещающих объектов, $c_{av} \in \mathbb{R}$.

4.3.1 Добавление одного алгоритма

Перенумеруем алгоритмы в порядке неубывания $n(a, \mathbb{X})$ — числа ошибок на полной выборке: $A = \{a_0, \ldots, a_D\}$. Очевидно, $n(a_0, \mathbb{X}) = 0$.

Обозначим через μ_d метод обучения, удовлетворяющий определению 4.3 и выбирающий алгоритмы только из подмножества $A_d = \{a_0, \ldots, a_d\}$. Рассмотрим процедуру последовательного добавления алгоритмов, на каждом шаге которой осуществляется переход от метода μ_{d-1} к методу μ_d.

Допустим, что для всех алгоритмов a_t, $t < d$, информация $\mathfrak{I}(a_t)$ относительно метода μ_{d-1} уже вычислена. Зададимся целью вычислить информацию $\mathfrak{I}(a_d)$ и скорректировать информацию $\mathfrak{I}(a_t)$, $t < d$, относительно метода μ_d. Заметим, что такая коррекция в общем случае необходима, поскольку алгоритм a_d может отбирать некоторые разбиения у каждого из предыдущих алгоритмов a_t.

Лемма 4.19. *Метод μ_d выбирает алгоритм a_d тогда и только тогда, когда все объекты, на которых a_d допускает ошибку, попадают в контрольную выборку:*

$$\big[\mu_d X{=}a_d\big] = \big[X'_d \subseteq \bar{X}\big], \quad X'_d = \big\{x_i \in \mathbb{X}\colon I(a_d, x_i) = 1\big\}.$$

Доказательство. Если хотя бы один объект, на котором a_d допускает ошибку, окажется в обучающей выборке X, то метод μ_d выберет алгоритм с меньшим числом ошибок на X. Такой алгоритм точно есть, например, a_0. Итак, условие $X'_d \subseteq \bar{X}$ является необходимым для того, чтобы метод μ_d выбрал алгоритм a_d. Покажем, что оно является также и достаточным. Для этого достаточно показать, что если алгоритмов из A_d, не допускающих ошибок на обучающей выборке X, окажется несколько, то из них будет выбран именно a_d. В силу упорядоченности множества A_d алгоритм a_d допускает максимальное число ошибок на $\mathbb{X}$, а среди алгоритмов с таким же числом ошибок на $\mathbb{X}$ имеет максимальный номер. Поэтому, согласно определению 4.3, алгоритм a_d будет выбран методом μ_d из A_d всегда, когда $X'_d \subseteq \bar{X}$. ■

Допустим, что непосредственно перед добавлением алгоритма a_d условия выбора каждого из предыдущих алгоритмов a_t были записаны в виде (4.7):

$$\big[\mu_{d-1} X{=}a_t\big] = \sum_{v \in V_t} c_{tv} \underbrace{\big[X_{tv} \subseteq X\big]\big[X'_{tv} \subseteq \bar{X}\big]}_{J_{tv}(d-1)}, \quad t < d.$$

После добавления алгоритма a_d эти условия изменятся. К ним добавится требование, чтобы множество X'_d не лежало целиком в контрольной выборке $\bar{X}$; иначе вместо алгоритма a_t метод μ_d выберет алгоритм a_d:

$$\begin{aligned}\big[\mu_d X{=}a_t\big] &= \big[\mu_{d-1} X{=}a_t\big]\big[X'_d \not\subseteq \bar{X}\big] = \\ &= \sum_{v \in V_t} c_{tv} \underbrace{\big[X_{tv} \subseteq X\big]\big[X'_{tv} \subseteq \bar{X}\big]\big[X'_d \not\subseteq \bar{X}\big]}_{J_{tv}(d)}, \quad t < d. \end{aligned} \tag{4.17}$$

Индукцией по d легко доказывается, что тождество (4.4) выполняется на каждом шаге, то есть что условия (4.17) определены корректно. Действительно, предполагая

$$\sum_{t=0}^{d-1} \big[\mu_{d-1} X{=}a_t\big] = 1,$$

получаем

$$\sum_{t=0}^{d} \big[\mu_d X{=}a_t\big] = \sum_{t=0}^{d-1} \big[\mu_{d-1} X{=}a_t\big]\big[X'_d \not\subseteq \bar{X}\big] + \big[X'_d \subseteq \bar{X}\big] = 1.$$

Чтобы получить правила корректировки информации $\mathfrak{I}(a_t)$, достаточно привести выражение (4.17) к виду (4.7), что и будет проделано в следующей лемме.

Лемма 4.20. *Корректировка информации $\mathfrak{I}(a_t)$, $t < d$ при добавлении алгоритма a_d сводится к проверке для каждого $v \in V_t$ такого, что $X_{tv} \cap X'_d = \varnothing$, трёх условий:*

1) *если $X'_d \setminus X'_{tv} = \{x_i\}$ — одноэлементное множество, то x_i присоединяется к X_{tv};*
2) *если $|X'_d \setminus X'_{tv}| > 1$, то множество индексов V_t пополняется ещё одним элементом (обозначим его w) и полагается $c_{tw} = -c_{tv}$, $X_{tw} = X_{tv}$, $X'_{tw} = X'_{tv} \cup X'_d$;*
3) *если $|X'_d \setminus X'_{tv}| = 0$, то из множества индексов V_t удаляется индекс v; соответственно, из $\mathfrak{I}(a_t)$ удаляется вся тройка $\langle X_{tv}, X'_{tv}, c_{tv}\rangle$.*

Доказательство. Если $X_{tv} \cap X'_d \neq \varnothing$, то из $X_{tv} \subseteq X$ следует, что множество X'_d не лежит целиком в контрольной выборке $\bar{X}$. Следовательно, никакая корректировка для тройки $\langle X_{tv}, X'_{tv}, c_{tv}\rangle$ не нужна:

$$J_{tv}(d) = \big[X_{tv} \subseteq X\big]\big[X'_{tv} \subseteq \bar{X}\big] = J_{tv}(d-1).$$

Если всё-таки $X_{tv} \cap X'_d = \varnothing$, то возможны три случая, в зависимости от мощности множества $X'_d \setminus X'_{tv}$.

Первый случай: $X'_d \setminus X'_{tv} = \{x_i\}$ — одноэлементное множество. Тогда справедлива цепочка равенств $[X'_d \not\subseteq \bar{X}] = [x_i \notin \bar{X}] = [x_i \in X]$. Подставляя в (4.17), получим

$$J_{tv}(d) = \big[X_{tv} \sqcup \{x_i\} \subseteq X\big]\big[X'_{tv} \subseteq \bar{X}\big].$$

Второй случай: $|X'_d \setminus X'_{tv}| > 1$. Тогда

$$\begin{aligned} J_{tv}(d) &= \big[X_{tv} \subseteq X\big]\big[X'_{tv} \subseteq \bar{X}\big]\big(1 - [X'_d \subseteq \bar{X}]\big) = \\ &= J_{tv}(d-1) - \big[X_{tv} \subseteq X\big]\big[X'_{tv} \cup X'_d \subseteq \bar{X}\big]. \end{aligned} \tag{4.18}$$

Таким образом, в выражении для $\big[\mu_d X{=}a_t\big]$ появится ещё одно слагаемое, что равносильно добавлению во множество V_t ещё одного индекса (обозначим его w), для которого полагается $c_{tw} = -c_{tv}$, $X_{tw} = X_{tv}$, $X'_{tw} = X'_{tv} \cup X'_d$.

Наконец, третий случай: $|X'_d \setminus X'_{tv}| = 0$. Тогда представление (4.18) остаётся в силе, однако разность $J_{tv}(d)$ оказывается равной нулю, поскольку $X'_{tv} \cup X'_d = X'_{tv}$. Обнуление $J_{tv}(d)$ равносильно удалению индекса v из множества индексов V_t вместе с удалением соответствующей тройки $\langle X_{tv}, X'_{tv}, c_{tv}\rangle$ из информации $\mathfrak{I}(a_t)$. ■

4.3.2 Вычисление вероятности переобучения

Леммы 4.19, 4.20 и Теорема 4.7 позволяют рекуррентно вычислять вероятность переобучения Q_ε. На каждом d-м шаге, $d = 0, \ldots, D$, добавляется алгоритм a_d, вычисляется информация $\mathfrak{I}(a_d)$; затем для каждого $t = 0, \ldots, d-1$ корректируется информация $\mathfrak{I}(a_t)$ и вероятности P_{tv}. На основе скорректированной информации обновляется текущая оценка Q_ε. По окончании последнего D-го шага текущая оценка Q_ε должна совпадать с точным значением вероятности переобучения. Эта рекуррентная вычислительная процедура подробно записана в виде псевдокода Алгоритма 4.3.1.

Алгоритм 4.3.1. Рекуррентное вычисление вероятности переобучения

Вход:
матрица ошибок $I(a_d, x_i)$, $d = 0, \ldots, D$, $i = 1, \ldots, L$;

Выход:
информация $\mathfrak{I}(a_d) = \langle X_{dv}, X'_{dv}, c_{dv} \rangle_{v \in V_d}$ для всех $d = 0, \ldots, D$,
вероятность переобучения Q_ε;

$Q_\varepsilon := 0$; упорядочить $A = \{a_0, \ldots, a_D\}$ по возрастанию $n(a_d, \mathbb{X})$;
для всех $d := 1, \ldots, D$
  добавить алгоритм a_d:
  $V_d := \{\varnothing\}$; $X_d := \varnothing$; $X'_d := \big\{x \in \mathbb{X} \colon I(a_d, x) = 1\big\}$; $c_d := 1$; $P_d := \prod\limits_{j=0}^{|X'_d|-1} \big(\frac{k-j}{L-j}\big)$;
  если $n(a_d, \mathbb{X}) \geqslant \varepsilon k$ **то** $Q_\varepsilon := Q_\varepsilon + P_d$;
  для всех $t := 0, \ldots, d-1$
    для всех $v \in V_t$ таких, что $X_{tv} \cap X'_d = \varnothing$
      $\Delta := |X'_d \setminus X'_{tv}|$;
      если $\Delta = 1$ **то**
        скорректировать множество X_{tv} и вероятность P_{tv}:
        $X_{tv} := X_{tv} \sqcup (X'_d \setminus X'_{tv})$; $P'_{tv} := P_{tv}$; $P_{tv} := P_{tv}\ell_{tv}/L_{tv}$;
        если $n(a_t, \mathbb{X}) \geqslant \varepsilon k$ **то** $Q_\varepsilon := Q_\varepsilon + c_{tv}(P_{tv} - P'_{tv})$;

      иначе если $\Delta > 1$ **то**
        добавить в V_t новый индекс w;
        $X_{tw} := X_{tv}$; $X'_{tw} := X'_{tv} \cup X'_d$; $c_{tw} := -c_{tv}$; $P_{tw} := P_{tv} \prod\limits_{j=0}^{\Delta-1} \big(1 - \frac{\ell_{tv}}{L_{tv}-j}\big)$;
        если $n(a_t, \mathbb{X}) \geqslant \varepsilon k$ **то** $Q_\varepsilon := Q_\varepsilon + c_{tw}P_{tw}$;
      иначе
        удалить из V_t индекс v;
        если $n(a_t, \mathbb{X}) \geqslant \varepsilon k$ **то** $Q_\varepsilon := Q_\varepsilon - c_{tv}P_{tv}$;

Некоторые обозначения, использованные в Алгоритме 4.3.1, требуют дополнительных пояснений.

Во-первых, вместо парного индекса ${}_{a_t v}$ для алгоритма a_t всюду используется сокращённое обозначение ${}_{tv}$.

Во-вторых, предполагается $L_{tv} = L - |X_{tv}| - |X'_{tv}|$, $\ell_{tv} = \ell - |X_{tv}|$, и вычисление величин L_{tv} и ℓ_{tv} в явном виде не записывается, чтобы не перегружать псевдокод.

В-третьих, при добавлении алгоритма a_d на шаге 3 создаётся тройка $\langle X_d, X'_d, c_d \rangle$ и вычисляется вероятность P_d получения алгоритма a_d методом μ_d. При этом в индексное множество заносится «пустой элемент», обозначаемый $\varnothing$. Это позволяет в дальнейшем сокращать запись индексов, полагая, что если индекс не двойной, а одинарный, то имеется в виду первый элемент индексного множества V_d, то есть $X_d \equiv X_{d\varnothing}$, $X'_d \equiv X'_{d\varnothing}$, $c_d \equiv c_{d\varnothing}$, $P_d \equiv P_{d\varnothing}$.

Вычисление вероятностей $P_{tw} = \mathsf{P}\big[X_{tw} \subseteq X\big]\big[X'_{tw} \subseteq \bar{X}\big]$ по формуле (4.9) также требует пояснений.

На шаге 3 вычисляется вероятность получения алгоритма a_d методом μ_d:

$$P_d = \frac{C^{\ell}_{L-|X'_d|}}{C^{\ell}_{L}} = \prod_{j=0}^{|X'_d|-1} \left(\frac{k-j}{L-j}\right).$$

Значения P_d одинаковы для всех алгоритмов с равным числом ошибок $n(a_d, \mathbb{X})$. Поэтому их можно вычислить один раз заранее для всех $n = 0, \ldots, L$.

На шаге 8 к порождающему множеству X_{tv} добавляется один объект, что приводит к очень простой корректировке предыдущего значения вероятности P'_{tv}:

$$P_{tv} = \frac{C^{\ell_{tv}-1}_{L_{tv}-1}}{C^{\ell}_{L}} = \frac{C^{\ell_{tv}}_{L_{tv}}}{C^{\ell}_{L}} \frac{\ell_{tv}}{L_{tv}} = P'_{tv} \frac{\ell_{tv}}{L_{tv}}.$$

На шаге 10 из тройки $\langle X_{tv}, X'_{tv}, c_{tv}\rangle$ формируется новая тройка $\langle X_{tw}, X'_{tw}, c_{tw}\rangle$ путём добавления Δ элементов к запрещающему множеству X'_{tw}. Следовательно,

$$P_{tw} = \frac{C^{\ell_{tv}}_{L_{tv}-\Delta}}{C^{\ell}_{L}} = \frac{C^{\ell_{tv}}_{L_{tv}}}{C^{\ell}_{L}} \frac{(L_{tv}-\ell_{tv}-\Delta+1)\cdots(L_{tv}-\ell_{tv})}{(L_{tv}-\Delta+1)\cdots L_{tv}} = P_{tv} \prod_{j=0}^{\Delta-1} \left(1 - \frac{\ell_{tv}}{L_{tv}-j}\right).$$

Алгоритм 4.3.1 может оказаться вычислительно неэффективным, если шаг 10 будет выполняться слишком часто. Каждое его выполнение приводит к добавлению ещё одного слагаемого в сумму (4.10). Следующая теорема позволяет сокращать время вычисления за счёт понижения точности верхней оценки Q_ε.

Теорема 4.21. *Если в Алгоритме 4.3.1 иногда пропускать шаг 10 при $c_{tv} = 1$, то вычисляемое в результате значение Q_ε будет верхней оценкой вероятности переобучения.*

Доказательство. Выполнение шага 10 при $c_{tv} = 1$, $c_{tw} = -1$ приводит к уменьшению текущего вычисленного значения Q_ε на величину $P_{tw} \geqslant 0$. Соответственно, невыполнение шага 10 приводит к исключению из суммы (4.11) отрицательного слагаемого $-P_{tw}$, и, возможно, ещё некоторого числа положительных и отрицательных слагаемых, которые появятся в этой сумме в результате последующих корректировок тройки $\langle X_{tw}, X'_{tw}, c_{tw}\rangle$ при выполнении шагов 10 и 12. Каждая такая корректировка возникает в результате добавления некоторого алгоритма a_d, $d > t$, который «отнимает» часть разбиений у алгоритма a_t, уменьшая слагаемое P_{tw} до величины $\tilde{P}_{tw}$:

$$\tilde{P}_{tw} = \mathsf{P}\big[X_{tw} \subseteq X\big]\big[X'_{tw} \subseteq \bar{X}\big]\big[X'_d \not\subseteq \bar{X}\big] \leqslant P_{tw}.$$

Исключение из суммы (4.11) отрицательного слагаемого $-P_{tw}$ вместе со всеми последующими слагаемыми, корректирующими тройку $\langle X_{tw}, X'_{tw}, c_{tw}\rangle$, может только увеличить значение Q_ε, вычисляемое Алгоритмом 4.3.1. ■

4.3.3 Профили расслоения и связности

Граф расслоения и связности. Для каждого алгоритма $a \in A$ обозначим через E_a множество объектов генеральной выборки $\mathbb{X}$, на которых он допускает ошибку: $E_a = \{x_i \in \mathbb{X}\colon I(a, x_i) = 1\}$. Очевидно, $n(a, \mathbb{X}) = |E_a|$.

Подмножество алгоритмов $A_m = \{a \in A\colon n(a, \mathbb{X}) = m\}$ называется *m-м слоем* множества A.

Разбиение $A = A_0 \sqcup \cdots \sqcup A_L$ называется *расслоением* множества алгоритмов A.

Определение 4.11. *Связностью $q(a)$ алгоритма $a \in A$ будем называть число алгоритмов в следующем слое, допускающих ошибки на тех же объектах, что и a:*

$$q(a) = \#\{a' \in A_{n(a,\mathbb{X})+1}\colon I(a, x) \leqslant I(a', x),\ x \in \mathbb{X}\}.$$

Образно говоря, связность $q(a)$ — это число способов, которыми вектор ошибок алгоритма a может быть «испорчен» ещё на одном каком-то объекте, если рассматривать всевозможные векторы ошибок алгоритмов множества A.

Графом связности или просто графом множества алгоритмов A будем называть направленный граф, вершины которого соответствуют алгоритмам, а рёбрами (a, a') соединяются пары алгоритмов, для которых $E_{a'} \setminus E_a = 1$. Тогда связность $q(a)$ алгоритма a — это число рёбер графа, исходящих из вершины a.

Определение 4.12. *Профилем расслоения и связности множества A называется матрица $\big(\Delta_{mq}\big)_{m=0}^{L}{}_{q=0}^{L}$, где Δ_{mq} — число алгоритмов в m-м слое со связностью q.*

Пример 4.3. На рис. 4.9 слева показана двумерная линейно разделимая выборка длины $L = 10$, состоящая из объектов двух классов, по 5 объектов в каждом классе. Справа построен граф связности множества линейных алгоритмов классификации для данной выборки. По вертикальной оси отложены номера слоёв m. Единственная точка на графе при $m = 0$ соответствует алгоритму, разделяющему объекты на два класса без ошибок. Следующий слой $m = 1$ содержит всевозможные алгоритмы, разделяющие выборку на два класса с одной ошибкой (для данной выборки их оказалось ровно 5). Слой $m = 2$ содержит уже 8 алгоритмов, и т. д.

Упрощённая рекуррентная оценка вероятности переобучения получится, если из Алгоритма 4.3.1 убрать шаги 9–12. В этом случае шаг 10 пропускается всегда, тройки $\langle X_{tw}, X'_{tw}, c_{tw}\rangle$ с отрицательными значениями c_{tw} никогда не создаются, каждому алгоритму a_d соответствует только одна тройка, и все индексные множества V_d, $d = 0, \ldots, D$ одноэлементны. Согласно Теореме 4.21, значение Q_ε, вычисляемое упрощённым рекуррентным Алгоритмом 4.3.1, будет верхней оценкой вероятности переобучения. В следующей теореме эта оценка выписывается в явном виде через профиль расслоения и связности множества A.

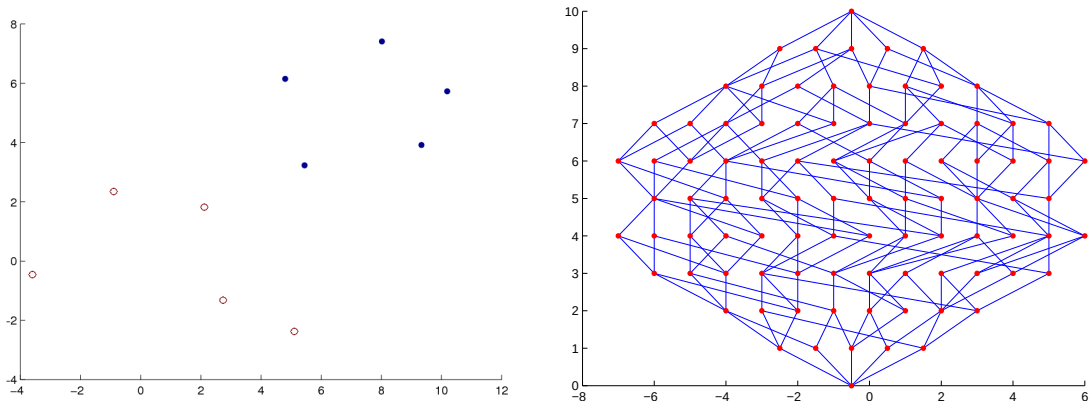


Рис. 4.9. Исходная выборка и граф связности множества линейных алгоритмов классификации.

Теорема 4.22. *Пусть векторы ошибок всех алгоритмов множества A попарно различны, в A есть корректный на $\mathbb{X}$ алгоритм, Δ_{mq} — число алгоритмов в m-м слое со связностью q. Тогда справедлива верхняя оценка вероятности переобучения*

$$Q_\varepsilon \leqslant \sum_{m=\lceil \varepsilon k \rceil}^{L} \sum_{q=0}^{L} \Delta_{mq} \frac{C_{L-m-q}^{\ell-q}}{C_L^\ell}. \tag{4.19}$$

Доказательство. Рассмотрим упрощённый рекуррентный Алгоритм 4.3.1, из которого удалены шаги 9–12. Он даёт верхнюю оценку вероятности переобучения. Для каждого алгоритма $a \in A$ он строит единственную тройку $\langle X_a, X'_a, 1\rangle$, в которой запрещающее множество X'_a совпадает с E_a, а порождающее множество состоит из всех объектов, добавленных на шаге 8. Это те и только те объекты x_i, для которых существует алгоритм $a' \in A$, допускающий на одну ошибку больше, чем a. Очевидно, при этом $X'_{a'} \setminus X'_a = E_{a'} \setminus E_a = \{x_i\}$ — одноэлементное множество. Число таких объектов x_i совпадает со значением связности $q(a)$. Таким образом, $|X'_a| = n(a, \mathbb{X})$, $|X_a| = q(a)$ для произвольного алгоритма $a \in A$. Следовательно, оценка (4.11) принимает вид:

$$\begin{aligned} Q_\varepsilon &\leqslant \sum_{a\in A} \big[n(a,\mathbb{X}) \geqslant \varepsilon k\big] \frac{C_{L_a}^{\ell_a}}{C_L^\ell} = \sum_{a\in A} \big[n(a,\mathbb{X}) \geqslant \varepsilon k\big] \frac{C_{L-n(a,\mathbb{X})-q(a)}^{\ell-q(a)}}{C_L^\ell} = \\ &= \sum_{m=\lceil \varepsilon k \rceil}^{L} \sum_{q=0}^{L} \underbrace{\sum_{a\in A} \big[n(a,\mathbb{X}) = m\big]\big[q(a) = q\big]}_{\Delta_{mq}} \frac{C_{L-m-q}^{\ell-q}}{C_L^\ell}. \end{aligned}$$

■

Согласно оценке (4.19) наибольший вклад в вероятность переобучения вносят алгоритмы с малым числом ошибок, начиная от $m = \lceil \varepsilon k \rceil$. По мере увеличения m комбинаторный множитель $\frac{C_{L-m-q}^{\ell-q}}{C_L^\ell}$ убывает экспоненциально.

Увеличение связности q улучшает оценку. В экспериментах с линейными алгоритмами классификации среднее значение связности q зависело прямо пропорционально от размерности пространства (числа признаков) с коэффициентом пропорциональности, немного большим единицы [59].

В общем случае при увеличении размерности пространства возникают два противонаправленных эффекта: с одной стороны, увеличивается число алгоритмов в каждом слое, что приводит к росту Q_ε; с другой стороны, увеличивается связность q, что приводит к уменьшению Q_ε.

Гипотеза о сепарабельности профиля расслоения–связности. Предварительные эксперименты с линейными алгоритмами классификации и методом ближайших соседей показали, что профиль расслоения и связности Δ_{mq} с высокой точностью является *сепарабельным*:

$$\Delta_{mq} \lessapprox \Delta_m \lambda_q,$$

где Δ_m — коэффициент разнообразия m-го слоя, λ_q — доля алгоритмов m-го слоя, имеющих связность q.

Вектор $(\Delta_m)_{m=0}^L$ предлагается называть *профилем расслоения*, а вектор $(\lambda_q)_{q=0}^L$ — *профилем связности* множества алгоритмов A.

Профиль связности удовлетворяет условию нормировки $\sum\limits_{q=0}^{L} \lambda_q = 1$.

На рис. 4.10 показаны графики зависимости Δ_{mq} от m и q для множества линейных алгоритмов классификации и линейно разделимых двумерных выборок длины $L = 20,\ 50,\ 100$. Хорошо видно, что профиль связности концентрируется в точке $q = 2$, что совпадает с размерностью пространства. С увеличением длины выборки доминирование данной компоненты профиля усиливается[1].

В терминах профилей расслоения и связности слегка ухудшенная оценка (4.19) принимает следующий вид:

$$Q_\varepsilon \leqslant \underbrace{\sum_{m=\lceil \varepsilon k \rceil}^{k} \Delta_m \frac{C_{L-m}^{\ell}}{C_L^{\ell}}}_{\text{VC-оценка}} \underbrace{\sum_{q=0}^{L} \lambda_q \left(\frac{\ell}{L-m} \right)^q}_{\text{поправка на связность}}. \tag{4.20}$$

Первая часть этой оценки представляет собой в VC-оценку (1.62), выраженную через профиль расслоения для частного случая, когда метод минимизации эмпирического риска является корректным (всегда находит алгоритм, не ошибающийся на обучающей выборке). В данном случае это так, поскольку множество A содержит корректный алгоритм, $n(a_0, \mathbb{X}) = 0$.

Вторая часть представляет собой «поправку на связность». Она экспоненциально быстро убывает с ростом q, что делает оценку существенно более точной, чем классическая VC-оценка (1.50) и чем оценка, учитывающая только профиль расслоения (1.62).

При известной $L \times D$-матрице ошибок вычисление по формуле (4.19) занимает $O(D)$ операций, тогда как упрощённый Алгоритм 4.3.1 является более ресурсоёмким и требует $O(D^2)$ операций. Если профиль расслоения и связности Δ_{mq} каким-то

[1] Вычислительные эксперименты, представленные на рис. 4.9 и рис. 4.10, выполнены студентом ВМиК МГУ Ильёй Решетняком.

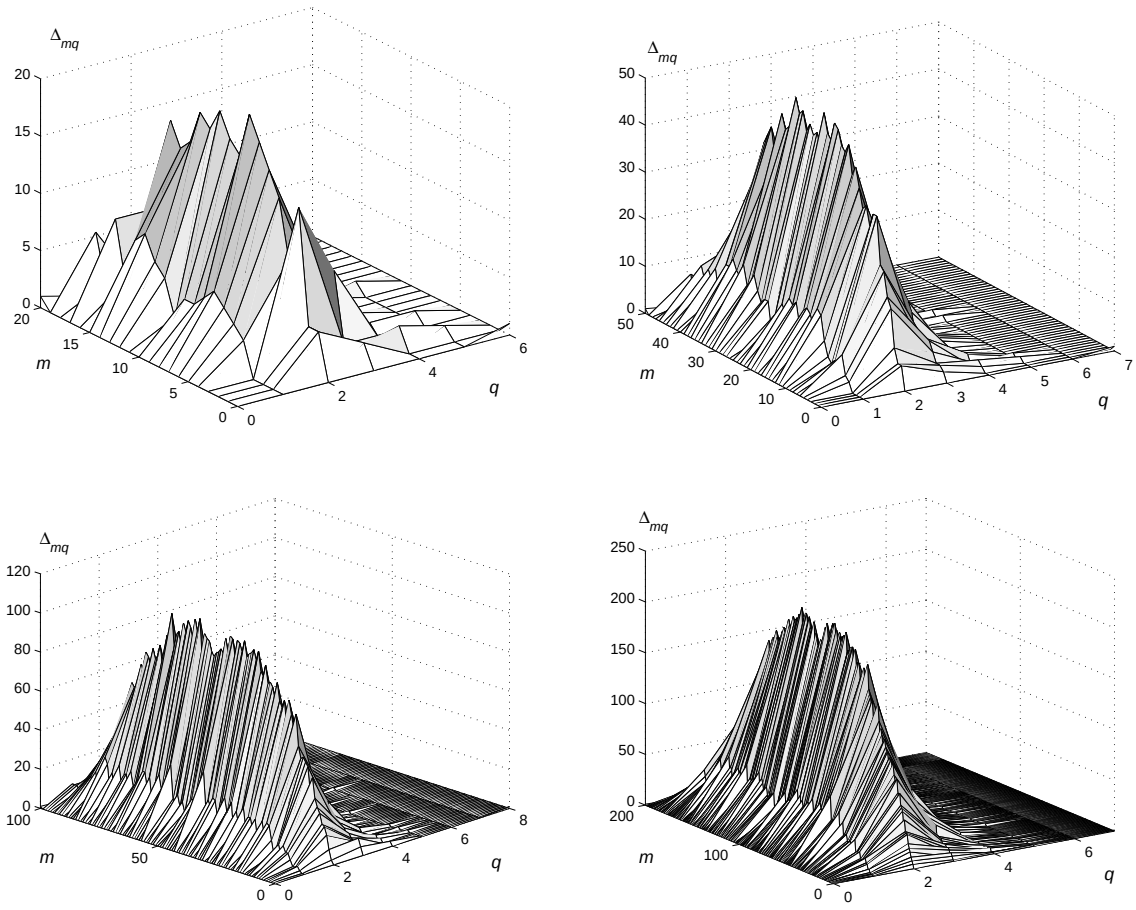


Рис. 4.10. Профили расслоения и связности для двумерных выборок длины $L = 20,\ 50,\ 100,\ 200$. Профиль Δ_{mq} — это количество алгоритмов с числом ошибок m на генеральной выборке и связностью q.

образом удалось оценить заранее, то вычисления займут $O(L^2)$ операций, что в реальных ситуациях существенно меньше, чем $O(D)$. Если же профиль Δ_{mq} представлен в виде разложения $\Delta_m \lambda_q$, то вычисления займут $O(L)$ операций, что уже совершенно приемлемо для практических приложений.

О других методах получения оценок расслоения–связности. Д. Кочедыков в [59] и И. Решетняк показали, что комбинаторные оценки вероятности переобучения, аналогичные (4.19), могут быть получены и без использования понятий порождающих и запрещающих множеств объектов.

Оценки, основанные на неравенствах Бонферрони-Галамбоса [136, 126] оказались лишь немного лучше VC-оценок. Они имеют тот же вид, что и VC-оценки, но функция гипергеометрического распределения $H_L^{\ell,m}(s_m^-(\varepsilon))$ в них заменяется на значение гипергеометрической вероятности $h_L^{\ell,m}(s_m^-(\varepsilon))$.

Затем была применена техника *цепного разложения*, что позволило получить оценки, аналогичные (4.19) и (4.20) [59].

На примере семейства линейных классификаторов в [59] экспериментально показано, что связность q не равна в точности размерности пространства, а несколько превышает её, и что при бóльших размерностях максимум профиля связности не настолько чётко выражен, как на рис. 4.10, а более «размазан». Поправка на связность в оценке (4.20) убывает с ростом размерности h экспоненциально, несколько быстрее, чем 2^{-h}. Уже при небольшой размерности (7–8 признаков) оценка (4.20) на 3 порядка лучше, чем VC-оценка.

О проблемах практического применения оценок расслоения–связности. Оценка вероятности переобучения (4.20) существенно зависит от генеральной выборки $\mathbb{X}$ и метода обучения μ. Из-за этого её практическое применение может натолкнуться на определённые трудности.

Известные оценки обобщающей способности, рассмотренные в главе 2, приводят в основном к двум путям конструктивного улучшения обобщающей способности.

Первый путь связан с модификацией принципа *минимизации эмпирического риска*. Типичные варианты модификации для задач классификации — это либо замена пороговой функции потерь различными *вещественными функциями потерь*, либо введение различных регуляризаторов, *штрафующих сложность* алгоритма. В некоторых случаях удаётся обосновать сразу обе модификации, как в оценке (2.4), где *рademахеровская сложность* $\mathcal{R}(A, X)$ играет роль «заготовки» для получения различных регуляризаторов, зависящих от выбора семейства алгоритмов A. На этом пути модифицируется только оптимизируемый функционал, а численные методы оптимизации вообще не затрагиваются.

Второй путь связан с модификацией процедуры поиска решения. Например, в современных методах *структурной минимизации риска* структура вложенных подсемейств строится в ходе оптимизации и зависит от конкретной обучающей выборки X. Это существенно отличается от исходного варианта VC-теории, где структуру предлагалось фиксировать до того, как станут известны данные X. На этом пути модифицируется численный метод оптимизации. При этом оценки обобщающей способности не обязаны выражаться в аналитическом виде. Они могут вычисляться алгоритмически с учётом выборки данных X и текущего состояния процесса оптимизации.

Второй путь представляется более подходящим для применения точных оценок вероятности переобучения. При этом могут возникать две основные проблемы: проблема эффективного вычисления точных оценок и проблема перехода от ненаблюдаемых оценок к наблюдаемым. Обе вполне преодолимы и являются в значительной мере техническими. Способы их решения зависят от выбранного семейства алгоритмов A и метода численной оптимизации.

Вычисление оценок расслоения–связности предполагает детальный анализ нижних слоёв графа связности, которые вносят максимальный вклад в значение Q_ε. Это означает, что оценка Q_ε зависит от окрестности оптимального алгоритма. Многие методы оптимизации, используемые для решения задач обучения, представляют собой сходящиеся итерационные процессы, в которых наибольшее число итераций приходится как раз на окрестность оптимума. Основная идея заключается в том, что-

бы использовать информацию, получаемую в ходе итераций, для оценивания обобщающей способности. Если этой информации окажется недостаточно, то возможно организовать дополнительное обследование окрестности оптимума, например, путём покоординатной «раскачки» вектора параметров. Оценка обобщающей способности позволит принять решение, следует ли продолжать искать лучший оптимум, или качество найденного решения вполне достаточно.

Практическая реализация этих идей требует конкретизации семейства алгоритмов A и метода оптимизации. Рассмотрение частных случаев выходит за рамки данной работы и является предметом дальнейших исследований.

4.4 Основные выводы

1. Разработано несколько общих подходов, позволяющих получать *точные оценки вероятности переобучения* в рамках слабой вероятностной аксиоматики.
2. Первый подход основан на выделении *порождающих и запрещающих множеств* объектов для каждого алгоритма в семействе. Доказывается, что порождающие и запрещающие множества можно указать всегда, а коль скоро они указаны, то можно выписать точные формулы как для вероятности переобучения, так и для вероятности получить каждый из алгоритмов. С помощью данного подхода в настоящей работе получены точные оценки для *модельных семейств* алгоритмов — монотонных и унимодальных цепочек и единичной окрестности наилучшего алгоритма. Другими авторами получены точные оценки для семейств более общего вида — монотонных и унимодальных многомерных сеток, пучков монотонных цепочек, хэммингових шаров заданного радиуса.
3. Второй подход основан на разбиении генеральной выборки на блоки. Получаемые *блочные оценки* вычислительно эффективны при малом числе алгоритмов в семействе. В частности, они позволяют выписать точные оценки для пары и тройки алгоритмов.
4. Третий подход применён при получении точных оценок вероятности переобучения для модельных семейств — *слоёв* и *интервалов булева куба*. Эти оценки, в частности, показывают, что вероятность переобучения очень быстро возрастает при увеличении числа «пограничных» объектов, для которых алгоритмы семейства реализуют все возможные дихотомии.
5. Четвёртый подход основан на *рекуррентном вычислении вероятности переобучения* при добавлении в семейство ещё одного алгоритма. Доказывается, что если в рекуррентной процедуре пропускать определённые шаги, то она будет выполняться гораздо быстрее и с гарантией даст либо верхнюю, либо нижнюю оценку вероятности переобучения. Более того, имеется возможность обменивать точность оценки на время вычисления. При максимальном упрощении рекуррентной процедуры верхняя оценка вероятности переобучения выписывается в явном аналитическом виде. Она похожа на VC-оценку, но зависит от *профиля расслоения и связности* семейства алгоритмов. Для связных семейств полу-

ченная оценка оказывается экспоненциально лучшей (относительно размерности пространства) по сравнению с классическими VC-оценками.

6. Точные оценки вероятности переобучения, как правило, являются ненаблюдаемыми, то есть зависят от статистических характеристик полной выборки. При их практическом применении возникает техническая задача перехода к наблюдаемым оценкам. В данной работе эта задача детально не рассматривается. Мы ограничиваемся описанием общей методологии таких переходов и некоторыми примерами, см. параграфы 1.1.3, 1.2.3, 2.3.3.

Глава 5

Комбинаторные оценки полного скользящего контроля

В данной главе выводятся формулы для эффективного вычисления функционала *полного скользящего контроля* (CCV), определяемого как средняя по всем разбиениям частота ошибок на контрольной выборке. Рассматриваются два практически важных частных случая — метод ближайшего соседа и монотонные классификаторы.

В параграфе 5.2 вводится понятие *профиля компактности* выборки, с его помощью выписывается точная формула CCV для метода ближайшего соседа. Предлагается метод отбора эталонных объектов, оптимизирующий CCV. Эксперименты показывают, что данный метод не склонен к переобучению.

В параграфе 5.3 вводятся понятия *верхнего и нижнего клина* объекта, понятие *профиля монотонности* выборки и выписывается не сильно завышенная верхняя оценка функционала CCV. Предлагается метод построения монотонных корректирующих операций, оптимизирующий CCV.

5.1 Функционал полного скользящего контроля

До сих пор для оценивания обобщающей способности метода обучения μ по генеральной выборке $\mathbb{X}$ мы вводили либо вероятность переобучения

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon\big],$$

либо вероятность большой частоты ошибок на контрольной выборке

$$R_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\nu(\mu X, \bar{X}) \geqslant \varepsilon\big].$$

Оба функционала зависят от параметра точности ε. В слабой аксиоматике они интерпретируются как функции распределения, соответственно, величины переобученности $\delta_\mu(X, \bar{X}) = \nu(\mu X, \bar{X}) - \nu(\mu X, X)$ и частоты ошибок на контроле $\nu(\mu X, \bar{X})$.

В данной главе рассматривается другой функционал, называемый в литературе *полным скользящем контролем* (complete cross-validation, CCV) [182] — это средняя

частота ошибок на контроле, если усреднять по всевозможным разбиениям генеральной выборки на обучение и контроль:

$$\mathrm{CCV}(\mu, \mathbb{X}) = \frac{1}{C_L^\ell} \sum_{(X,\bar{X})} \nu(\mu X, \bar{X}).$$

В терминах слабой аксиоматики CCV есть математическое ожидание частоты ошибок на контроле, $\mathrm{CCV} = \mathsf{E}\nu(\mu X, \bar{X})$. Недостаток CCV в том, что он характеризует среднее значение, но ничего не говорит о возможном разбросе (дисперсии) частоты ошибок $\nu(\mu X, \bar{X})$. Поэтому его нельзя непосредственно использовать для получения гарантированных верхних оценок.

Двусторонние соотношения между функционалами позволяют получать верхние оценки R_ε и Q_ε через CCV, и, наоборот, CCV через R_ε или Q_ε.

Теорема 5.1. *Для произвольных μ, $\mathbb{X}$ и $\varepsilon \in [0,1]$*

$$\varepsilon R_\varepsilon \leqslant \mathrm{CCV} \leqslant \varepsilon + (1-\varepsilon) R_\varepsilon;$$
$$\varepsilon Q_\varepsilon \leqslant \mathrm{CCV} \leqslant \varepsilon + (1-\varepsilon) Q_\varepsilon + \mathsf{E}\nu(\mu X, X);$$

Доказательство. Обозначим $\nu = \nu(\mu X, X)$, $\bar{\nu} = \nu(\mu X, \bar{X})$.

Первая оценка следует из того, что для любых $\bar{\nu}, \varepsilon \in [0,1]$ справедлива цепочка неравенств $\varepsilon[\bar{\nu} \geqslant \varepsilon] \leqslant \bar{\nu} \leqslant \varepsilon + (1-\varepsilon)[\bar{\nu} \geqslant \varepsilon]$.

Вторая оценка следует из того, что для любых $\nu, \bar{\nu}, \varepsilon \in [0,1]$ справедлива цепочка неравенств $\varepsilon[\bar{\nu} - \nu \geqslant \varepsilon] \leqslant \varepsilon[\bar{\nu} \geqslant \varepsilon] \leqslant \bar{\nu} \leqslant \varepsilon + (1-\varepsilon)[\bar{\nu} - \nu \geqslant \varepsilon] + \nu$. ■

Эти оценки являются достаточно грубыми и в данной работе не используются. Заметим также, что оценка $\varepsilon R_\varepsilon \leqslant \mathrm{CCV}$ представляет собой классическое неравенство Маркова $\mathsf{P}[\xi \geqslant \varepsilon] \leqslant \frac{1}{\varepsilon}\mathsf{E}\xi$, записанное для случайной величины $\xi = \nu(\mu X, \bar{X})$ [172].

5.2 Априорные ограничения компактности

Рассмотрим задачу классификации с конечным множеством классов $\mathbb{Y}$. Индикатор ошибки имеет вид $I(a,x) = \big[y(x) \neq a(x)\big]$, где $y\colon \mathbb{X} \to \mathbb{Y}$ — целевая зависимость, $a\colon \mathbb{X} \to \mathbb{Y}$ — алгоритм классификации.

При решении прикладных задач классификации и распознавания образов часто делается эмпирическое предположение, что классы образуют локализованные «компактные» подмножества объектов, следовательно, схожие объекты, как правило, лежат в одном классе. Это предположение называют *гипотезой компактности*. Простейшим методом обучения, построенным на его основе, является метод ближайших соседей.

5.2.1 Профиль компактности выборки

Пусть на множестве $\mathbb{X}$ определена функция расстояния $\rho(x, x')$.

Метод ближайшего соседа (nearest neighbor) — это метод обучения μ, который запоминает обучающую выборку $X \subset \mathbb{X}$ и строит алгоритм $a = \mu X$, работающий следующим образом:

$$a(x; X) = y\big(\arg\min_{x' \in X} \rho(x, x')\big) \text{ для всех } x \in \mathbb{X}.$$

Точное выражение функционала скользящего контроля CCV для метода ближайшего соседа и некоторых его модификаций найдено в [182] для случая двух классов, $|\mathbb{Y}| = 2$. Авторы этой работы ставили целью вывод эффективных вычислительных формул. Воспользуемся этим результатом, чтобы ввести характеристику выборки, являющуюся строгой формализацией гипотезы компактности.

Для каждого объекта x_i, $i = 1, \ldots, L$ выборки $\mathbb{X}$ расположим остальные $L - 1$ объектов в порядке возрастания расстояния до x_i, пронумеровав их двойными индексами: $x_i = x_{i0}$, $x_{i1}, x_{i2}, \ldots, x_{i,L-1}$. Таким образом,

$$0 = \rho(x_i, x_{i0}) \leqslant \rho(x_i, x_{i1}) \leqslant \cdots \leqslant \rho(x_i, x_{i,L-1}).$$

Обозначим через $r_m(x_i)$ ошибку, возникающую, если правильный ответ $y(x_i)$ на объекте x_i заменить ответом на его m-ом соседе:

$$r_m(x_i) = \big[y(x_i) \neq y(x_{im})\big]; \quad i = 1, \ldots, L; \quad m = 1, \ldots, L - 1.$$

Определение 5.1. *Профилем компактности выборки* $\mathbb{X}$ *называется функция* $K(m, \mathbb{X})$, *выражающая долю объектов выборки, для которых правильный ответ не совпадает с правильным ответом на* m*-ом соседе:*

$$K(m, \mathbb{X}) = \frac{1}{L} \sum_{i=1}^{L} r_m(x_i); \quad m = 1, \ldots, L - 1.$$

Профиль компактности является формальным выражением гипотезы компактности. Чем проще задача, то есть чем чаще близкие объекты оказываются в одном классе, тем сильнее «прижимается к нулю» начальный участок профиля. И, наоборот, в задачах, трудных для метода ближайшего соседа, где ближайшие объекты практически не несут информации о классе, профиль вырождается в константу, близкую к 0.5, см. средний ряд графиков на рис. 5.1.

5.2.2 Точная оценка полного скользящего контроля

Интуитивно очевидная связь профиля компактности с качеством классификации подтверждается следующей теоремой. Идея доказательства взята из [182].

Теорема 5.2. *Для метода ближайшего соседа* μ *справедливо следующее выражение функционала полного скользящего контроля* CCV*:*

$$\mathrm{CCV}(\mu, \mathbb{X}) = \sum_{m=1}^{k} K(m, \mathbb{X}) \frac{C_{L-1-m}^{\ell-1}}{C_{L-1}^{\ell}}. \tag{5.1}$$

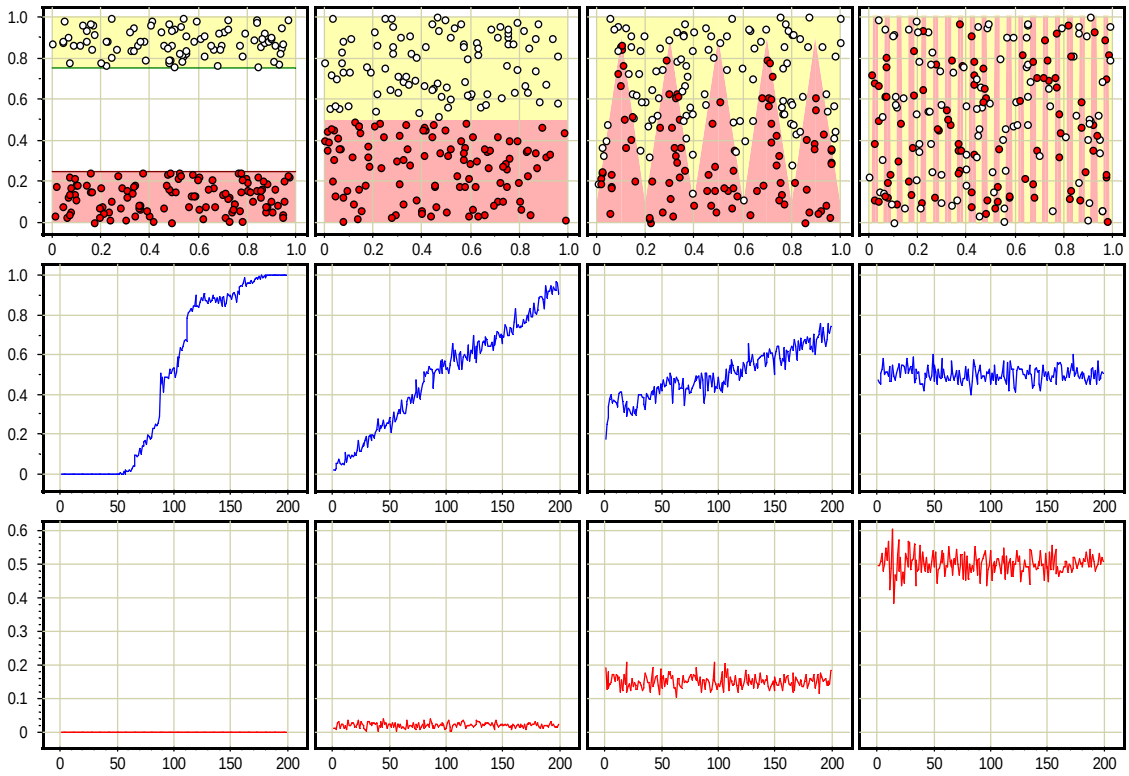


Рис. 5.1. Верхний ряд: 4 модельные задачи классификации, в порядке возрастания трудности, $L = 200$. Средний ряд: профили компактности этих задач. Чем ниже проходит начальный участок профиля, тем «проще» задача для алгоритма 1NN, и тем выше обобщающая способность. Нижний ряд: зависимость CCV от длины контрольной выборки k при фиксированной длине обучения $\ell = 200$.

Доказательство. Запишем в функционале скользящего контроля частоту ошибок через сумму индикаторов ошибки и переставим знаки суммирования:

$$\mathrm{CCV} = \frac{1}{C_L^\ell} \sum_{X,\bar{X}} \frac{1}{k} \sum_{x \in \bar{X}} I(\mu X, x) = \frac{1}{k} \sum_{i=1}^{L} \frac{1}{C_L^\ell} \underbrace{\sum_{X,\bar{X}} [x_i \in \bar{X}]\, I(\mu X, x_i)}_{N_i}. \tag{5.2}$$

Внутренняя сумма, обозначенная через N_i, выражает число разбиений выборки $\mathbb{X}$, при которых объект x_i оказывается в контрольной подвыборке и алгоритм μX допускает на нём ошибку. Данная ситуация реализуется для таких разбиений, при которых m первых объектов из последовательности $x_{i0}, x_{i1}, \ldots, x_{i,L-1}$ попадают в контрольную подвыборку, m-ый сосед x_{im} находится в обучающей подвыборке и принадлежит другому классу, то есть $r_m(x_i) = 1$. Поскольку m принимает значения от 1 до k, число таких разбиений в точности равно

$$N_i = \sum_{m=1}^{k} r_m(x_i)\, C_{L-1-m}^{\ell-1},$$

поскольку $C_{L-1-m}^{\ell-1}$ есть число способов выбрать $(\ell-1)$ обучающих объектов из оставшихся $(L-1-m)$. Подставляя N_i в (5.2), и используя определение профиля компактности, получаем требуемое выражение функционала CCV. ■

Комбинаторный множитель $\gamma_m = C_{L-1-m}^{\ell-1}/C_{L-1}^{\ell}$ убывает с ростом m быстрее геометрической прогрессии, поскольку

$$\gamma_{m+1} = q(m)\gamma_m, \qquad q(m) = 1 - \frac{\ell-1}{L-1-m} < \frac{k}{L-1},$$

и $q(m)$ — убывающая функция от m. Поэтому для обеспечения малого значения функционала CCV достаточно потребовать, чтобы профиль $K(m)$ принимал малые значения при малых m, то есть чтобы близкие объекты лежали преимущественно в одном классе. При больших m рост $K(m)$ компенсируется убыванием комбинаторного множителя, поэтому классификации далёких друг от друга объектов не влияют на значение функционала CCV. Таким образом, форма профиля компактности может рассматриваться как характеристика выборки, показывающая, насколько метод ближайшего соседа при выбранной метрике ρ подходит для решения данной задачи.

При $k = 1$ профиль компактности вырождается в точку и совпадает с самим функционалом, который в этом случае называют скользящим контролем с *исключением объектов по одному* (leave-one-out, LOO). В общем случае профиль состоит из k точек, причём относительный вклад $K(m)$ быстро уменьшается с ростом m. Например, при $k = 1, 2, 3$:

$$\begin{aligned}
k=1: &\quad \text{CCV} = K(1);\\
k=2: &\quad \text{CCV} = K(1)\tfrac{\ell}{\ell+1} + K(2)\tfrac{1}{\ell+1};\\
k=3: &\quad \text{CCV} = K(1)\tfrac{\ell}{\ell+2} + K(2)\tfrac{2\ell}{(\ell+1)(\ell+2)} + K(3)\tfrac{2}{(\ell+1)(\ell+2)}.
\end{aligned}$$

Чем больше длина k контрольной выборки, тем меньше вклад начальных элементов профиля в CCV. Учитывая, что профиль, как правило, в целом возрастает, функционал CCV должен был бы возрастать с ростом длины контрольной выборки. С другой стороны, с ростом k повышается плотность объектов в выборке, и объекты начинают более надёжно классифицироваться по своим ближайшим соседям, что приводит к уменьшению начальных элементов профиля. Таким образом, функционал CCV должен был бы уменьшаться с ростом длины контрольной выборки. Нижний ряд графиков на рис. 5.1 показывает, что в данной модельной задаче эти два процесса полностью компенсируют друг друга, в результате длина контроля k не влияет на значение CCV. Для практики это означает, что можно обходиться малыми значениями k, для которых функционал CCV вычисляется проще. Пока остаётся открытым вопрос, всегда ли происходит такая компенсация, или это свойство данной серии двумерных модельных задач.

Вычисление профиля компактности требует $O(\ell^2)$ операций, упорядочивание объектов по близости — $O(\ell^2 \log \ell)$ операций. После этого вычисление CCV производится за $O(k)$ операций. Это гораздо быстрее, чем производить суммирование по всем C_L^ℓ разбиениям, что становится практически нереально уже при $k > 2$.

Быстрое вычисление CCV позволяет оптимизировать параметры метода, однако в классическом варианте метод ближайшего соседа не имеет параметров. В следующем параграфе рассмотрена задача оптимизации подмножества эталонных объектов. Можно было бы оптимизировать по CCV веса объектов или саму метрику ρ, в частности, веса признаков во взвешенной евклидовой метрике. Пока эти задачи также остаются открытыми.

Заметим, что *ёмкость* семейства алгоритмов, индуцируемого методом ближайшего соседа, бесконечна, поэтому классическая VC-теория не даёт никаких оценок для данного случая.

5.2.3 Задача отбора эталонных объектов

В метрических методах классификации, таких, как метод ближайших соседей, метод потенциальных функций, метод парзеновского окна, имеет смысл запоминать не всю обучающую выборку, а только подмножество наиболее типичных объектов, называемых *эталонами*. Отбор эталонов обычно преследует несколько целей: сокращение объёма хранимых данных, повышение скорости классификации, повышение качества классификации за счёт удаления нетипичных (шумовые) объектов.

Известные эвристические методы последовательного отбора эталонов Stolp, λ-Stolp [51] и FRiS-Stolp [6], основаны, фактически, на оценивании локальных плотностей классов в каждом объекте и вычислении отношений этих оценок. Эти методы неплохо зарекомендовали себя на практике, однако остаются открытыми теоретические вопросы: какой функционал они минимизируют, почему они обладают хорошей обобщающей способностью, почему в них использованы именно такие эвристики, и какие из многочисленных возможных вариантов этих эвристик могли бы работать ещё лучше.

Другой известный пример метода, в котором происходит отбор объектов — метод опорных векторов SVM [123, 118]. В этом методе эталонными (опорными) становятся объекты, лежащие на границе классов. Недостаток SVM в том, что в числе опорных оказываются также шумовые выбросы, расположенные внутри чужих классов. Похожий на SVM метод релевантных векторов RVM [211] отбирает в качестве опорных объекты, отстоящие от границы классов «на разумном расстоянии», и игнорирует шумовые выбросы. Как правило, RVM выбирает меньшее число опорных объектов по сравнению с SVM, и обобщающая способность RVM также лучше, чем у SVM, особенно на задачах с высоким уровнем шума.

Аналогичное свойство «релевантности» хотелось бы обеспечить и при отборе эталонов в метрических алгоритмах классификации. Однако желательно, чтобы это свойство не закладывалось в виде эвристик, а было бы следствием оптимизации функционала обобщающей способности.

О сложности задачи отбора эталонных объектов. Доказано, что задача отбора множества эталонных объектов в методе ближайшего соседа путём минимизации функционала CCV является NP-полной [52]. Тем самым, оправдано применение эвристических переборных алгоритмов, в том числе жадных алгоритмов.

Алгоритм отбора эталонов CCV-2005. В совместной работе с А. Колосковым [32] предложен метод отбора эталонов для алгоритма ближайшего соседа, основанный на минимизации функционала CCV.

Обозначим через $\Omega \subseteq \mathbb{X}$ искомое множество эталонов, через $r_m^\Omega(x_i)$ — ошибку, возникающую, если правильный ответ $y(x_i)$ на объекте x_i заменить ответом на его m-ом соседе, при условии, что соседи берутся только из множества Ω.

Определение 5.2. *Профилем компактности выборки $\mathbb{X}$ относительно множества эталонов Ω называется функция $K^\Omega(m, \mathbb{X})$, выражающая долю объектов выборки, для которых правильный ответ не совпадает с правильным ответом на m-ом соседе из множества Ω:*

$$K^\Omega(m, \mathbb{X}) = \frac{1}{L} \sum_{i=1}^{L} r_m^\Omega(x_i); \quad m = 1, \ldots, L-1.$$

Теорема 5.3. *Для метода ближайшего соседа μ^Ω, использующего в качестве эталонов только объекты множества $\Omega \subseteq \mathbb{X}$, справедливо следующее выражение функционала* CCV*:*

$$\mathrm{CCV}(\mu^\Omega, \mathbb{X}) = \sum_{m=1}^{k} K^\Omega(m, \mathbb{X}) \frac{C_{L-1-m}^{\ell-1}}{C_{L-1}^{\ell}}. \tag{5.3}$$

Доказательство. Полностью аналогично доказательству Теоремы 5.2. ■

Жадный алгоритм минимизации $\mathrm{CCV}(\mu^\Omega, \mathbb{X})$, предложенный в [32], начинает с $\Omega = \mathbb{X}$ и затем по одному удаляет из множества Ω неэталонные объекты. На каждом шаге находится тот объект, удаление которого минимизирует $\mathrm{CCV}(\mu^\Omega, \mathbb{X})$.

Оказалось, что в этом процессе сначала удаляются все шумовые выбросы, при этом функционал CCV убывает. Затем удаляются неинформативные «внутренние» объекты классов, которые не нужны для хорошей классификации окружающих объектов и сами хорошо классифицируются по своему окружению. При их удалении функционал либо не изменяется, либо увеличивается на ничтожно малую величину. Только в заключительной стадии функционал начинает заметно возрастать. Тогда процесс удаления объектов из Ω останавливается, и все оставшиеся объекты принимаются в качестве эталонных.

В модельных экспериментах оказалось, что данный метод вообще не подвержен переобучению. Частота ошибок классификации на отложенной тестовой выборке изменялась синхронно со значением CCV (аналогичный результат для следующего алгоритма показан на рис. 5.3).

Интересным для приложений побочным результатом является разделение всех объектов на три категории: шумовые, неинформативные и эталонные.

Недостатком данного метода является низкая эффективность по времени. Методы типа Stolp работают быстрее, так как в них множество эталонов формируется путём последовательного добавления, а не удаления объектов.

Алгоритм отбора эталонов CCV-2009. В совместной работе с М. Ивановым [53] предыдущий алгоритм был обобщён и улучшен сразу по нескольким направлениям[1].

1. Рассмотрен метод q ближайших соседей (а не только $q = 1$) и произвольное число классов (а не только $|\mathbb{Y}| = 2$).

2. Реализованы обе стратегии отбора эталонов — последовательное удаление неэталонных объектов и последовательное добавление эталонов.

3. Предложены эффективные алгоритмы итерационного пересчёта вкладов каждого объекта в функционал CCV при удалении и добавлении объектов в Ω. Пересчитываются только вклады объектов, расположенных поблизости от удаляемого или добавляемого объекта. Для быстрого построения прямых и обратных окрестностей объектов используются метрические деревья. В результате этих технических усовершенствований существенно повышена эффективность алгоритма, который теперь может применяться к выборкам из десятков и сотен тысяч объектов.

Вычислительный эксперимент был проведён на модельной двумерной выборке из $L = 1000$ объектов двух классов, порождаемых сферическими гауссовскими распределениями с дисперсиями 1 и 2 и расстоянием между центрами 5. На рис. 5.2 показан результат применения алгоритма жадного удаления объектов для метода одного ближайшего соседа. Этот алгоритм, как и CCV-2005, производит разбиение выборки на шумовые, внутренние и эталонные объекты. Эталоны каждого класса выстраиваются вдоль границы, но на некотором удалении от неё. Таким образом, реализуемая данным методом стратегия отбора эталонных объектов, с точки зрения результата, аналогична методу RVM.

После отбора эталонов форма границы становится более гладкой. Заметим также, что в области наиболее плотного пересечения классов граница неплохо описывается параболой, которая является оптимальным байесовским классификатором в данной модельной задаче; а в более разреженной области форма границы упрощается и становится похожей на линейную.

На рис. 5.3 показана зависимость функционала CCV от количества удалённых объектов $L - |\Omega|$. Левый участок $[0, 60]$ соответствует начальной стадии удаления шумов (около 40 объектов). Правый участок $[980, 1000]$ показывает, что число критически важных эталонов равно 6, но увеличение числа эталонов до 16 позволяет уменьшить частоту ошибок на тестовой выборке с 3,6% до 2,0%.

На рис. 5.2 выделено 42 эталона, которые обеспечивают минимум $\mathrm{CCV}(\Omega)$ и одновременно минимальную частоту ошибок на тестовой выборке 1,8%. Длинный средний участок $[60, 980]$, вырезанный из графика на рис. 5.3, соответствует стадии удаления неинформативных объектов, на этом участке значение $\mathrm{CCV}(\Omega)$ почти постоянно. Тонкой линией показана частота ошибок алгоритма $a(x, \Omega)$ на независимой тестовой выборке из 2000 объектов, сгенерированной с помощью того же распределения. Хорошо видно, что она изменяется синхронно с функционалом $\mathrm{CCV}(\Omega)$. Это означает, что отбор эталонных объектов не подвержен переобучению.

---

[1]Все математические результаты принадлежат М. Иванову, поэтому здесь приводятся только основные идеи, результаты экспериментов и выводы.

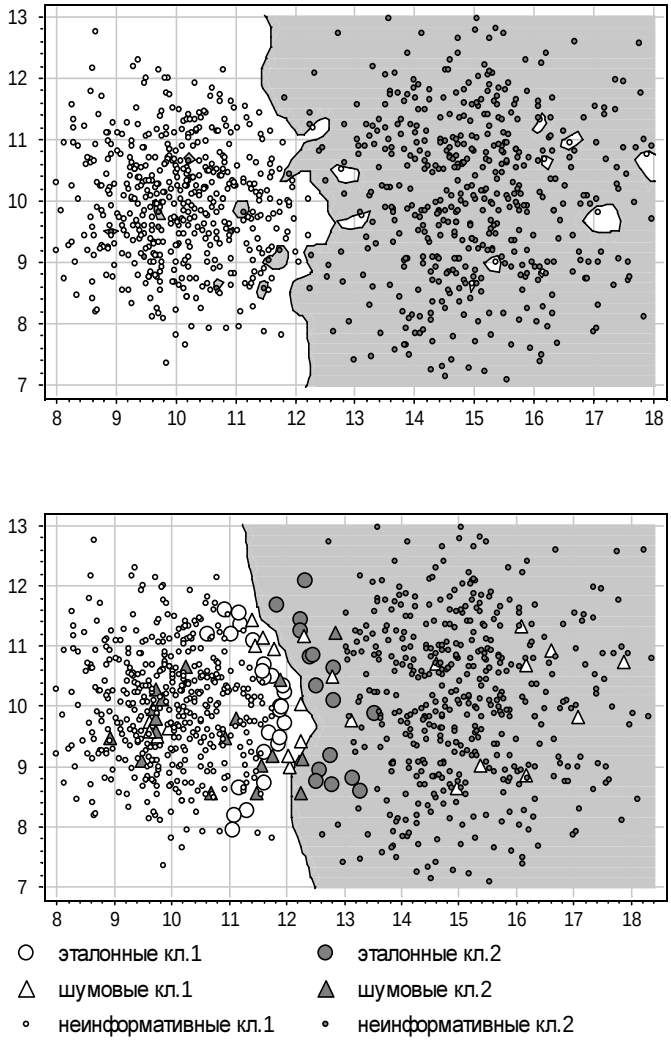


Рис. 5.2. Сверху: модельная задача классификации: 1000 объектов, алгоритм 1NN. Снизу: результат отбора эталонов путём последовательного удаления объектов, отобрано 26 эталонов класса 1 и 16 — класса 2.

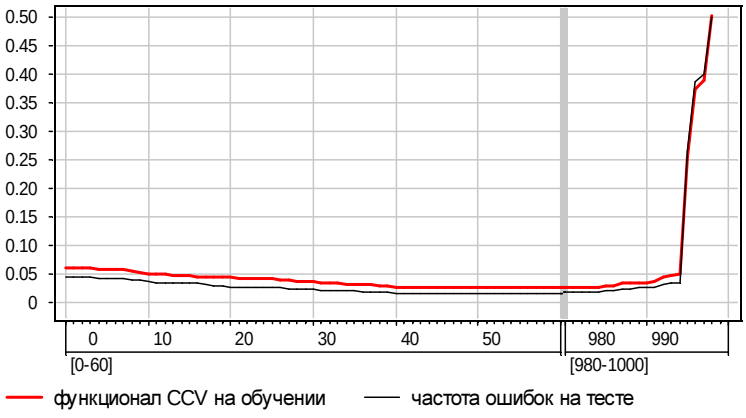


Рис. 5.3. Зависимость функционала CCV $Q(\Omega)$ от количества удаленных объектов $L - |\Omega|$ для алгоритма жадного удаления неэталонных объектов.

5.3 Априорные ограничения монотонности

Рассмотрим задачу классификации, в которой множество X частично упорядочено, $\mathbb{Y} = \{0, 1\}$, индикатор ошибки имеет вид $I(a, x) = \big[y(x) \neq a(x)\big]$.

Допустим, что имеется априорная информация о монотонности или почти-монотонности целевой зависимости $y(x)$. Данный тип ограничений может возникать на практике в нескольких различных случаях.

Во-первых, ограничения монотонности могут быть результатом формализации экспертных знаний вида «чем больше значение признака $f(x)$ тем больше значение $y(x)$» или, наоборот, «чем меньше $f(x)$, тем больше $y(x)$». Например, в медицинских задачах возможны суждения типа «чем старше пациент, тем выше риск осложнений» или «чем выше скорость кровотока в вене, тем выше риск рестеноза шунта». В задачах кредитного скоринга возможны суждения типа «чем больше заработная плата заёмщика и чем дольше он проживает на одном месте, тем ниже риск дефолта». Поскольку такие суждения часто носят не обязательный, а рекомендательный характер, целесообразно рассматривать не только строго монотонные, но и «почти-монотонные» зависимости.

Во-вторых, ограничения монотонности могут возникать при построении композиций алгоритмов. *Композицией алгоритмов* классификации $a_t(x) = \big[b_t(x) \geqslant 0\big]$, $t = 1, \ldots, T$, называется алгоритм $a(x) = \big[F(b_1(x), \ldots, b_T(x)) \geqslant 0\big]$, где отображение $F\colon \mathbb{R}^T \to \mathbb{R}$ называется *корректирующей операцией*, алгоритмы a_t называются *базовыми*. Наиболее распространённым типом композиций является *голосование!взвешенное*, когда корректирующая операция определяется как линейная выпуклая комбинация базовых алгоритмов [48, 197, 116, 161]:

$$F\big(b_1(x), \ldots, b_T(x)\big) = \sum_{t=1}^{T} \alpha_t b_t(x), \quad \alpha_t \geqslant 0.$$

Для взвешенного голосования важно свойство неотрицательности коэффициентов α_t, то есть что линейная функция F не убывает по всем её T аргументам. Отрицательность коэффициента α_t означала бы, что базовый алгоритм a_t слишком ненадёжен, и его ответы надо заменять на противоположные. Такие алгоритмы целесообразно вовсе исключать из композиции. Исходя из этих соображений в [76, 17] было предложено естественное обобщение линейных корректирующих операций — *монотонные корректирующие операции*, в которых F — произвольная монотонно неубывающая функция T вещественных аргументов. Задача построения монотонной корректирующей операции при фиксированных базовых алгоритмах сводится к построению монотонной функции, как можно точнее проходящей через заданные точки. При этом в роли объектов выступают T-мерные векторы $\big(b_1(x), \ldots, b_T(x)\big)$. В [17] описаны эффективные численные методы построения монотонных корректирующих операций в задачах классификации и восстановления регрессии.

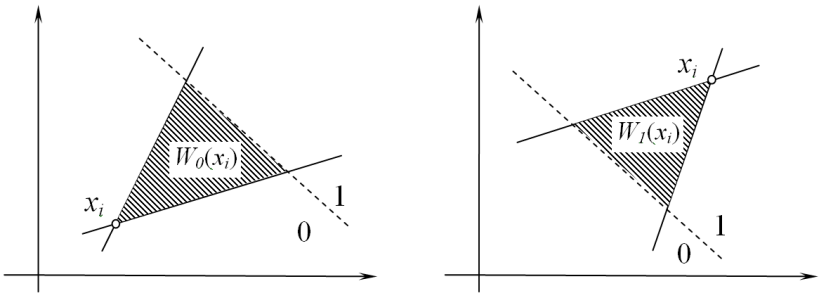


Рис. 5.4. Верхний (слева) и нижний (справа) клинья объекта x_i в условной двумерной задаче классификации с естественным отношением порядка на множестве $X = \mathbb{R}^2$. Пунктирной линией показана граница между классами.

5.3.1 Профиль монотонности выборки

Допустим, что метод обучения μ выбирает алгоритмы из множества A всех монотонных отображений $F\colon \mathbb{X} \to \mathbb{Y}$.

Определение 5.3. *Степенью немонотонности выборки $\mathbb{X}$ называется наименьшая частота ошибок, допускаемых на ней монотонными алгоритмами:*

$$\theta(\mathbb{X}) = \min_{a \in A} \nu(a, \mathbb{X}).$$

Выборка $\mathbb{X}$ называется монотонной, если из $x_i \leqslant x_j$ следует $y(x_i) \leqslant y(x_j)$ для всех $i, j = 1, \ldots, L$. Выборка монотонна тогда и только тогда, когда $\theta(\mathbb{X}) = 0$. Если метод μ минимизирует эмпирический риск, то есть строит алгоритмы с минимальной частотой ошибок на обучающей выборке в классе всех монотонных функций A, то метод μ будет корректным на любой монотонной выборке [17].

Определение 5.4. *Верхним и нижним клином объекта $x_i \in \mathbb{X}$ называются, соответственно, множества (см. рис. 5.4)*

$$W_0(x_i) = \{x \in \mathbb{X}\colon x_i < x \text{ и } y(x) = 0\};$$
$$W_1(x_i) = \{x \in \mathbb{X}\colon x < x_i \text{ и } y(x) = 1\}.$$

Введём сокращённое обозначение $W_i = W_{y(x_i)}(x_i)$.

Мощность клина $w_i = |W_i|$ характеризует глубину погружения объекта x_i в тот класс, которому он принадлежит. Чем меньше w_i, тем ближе объект к границе класса. Объекты, не имеющие своего клина ($w_i = 0$) будем называть *граничными*.

Если монотонный алгоритм допускает ошибку на объекте x_i, то он допускает ошибку и на всех объектах из клина W_i. Данный факт существенно используется при выводе верхней оценки CCV.

Определение 5.5. *Профилем монотонности выборки $\mathbb{X}$ называется функция $M(m, \mathbb{X})$, выражающая долю объектов выборки с клином мощности m:*

$$M(m, \mathbb{X}) = \frac{1}{L} \sum_{i=1}^{L} [w_i = m]; \quad m = 0, \ldots, L-1.$$

5.3.2 Верхняя оценка полного скользящего контроля

Теорема 5.4. *Если метод μ минимизирует эмпирический риск в классе всех монотонных функций и степень немонотонности выборки $\mathbb{X}$ равна θ, то*

$$\mathrm{CCV}(\mu,\mathbb{X}) \leqslant \sum_{m=0}^{\theta L+k-1} M(m,\mathbb{X}) \sum_{s=\max\{0,m-k+1\}}^{\min\{\theta L,\ell,m\}} \frac{C_m^s C_{L-1-m}^{\ell-s}}{C_{L-1}^{\ell}}. \tag{5.4}$$

Доказательство. Запишем в функционале скользящего контроля частоту ошибок через сумму индикаторов ошибки и поменяем местами знаки суммирования:

$$\mathrm{CCV} = \frac{1}{C_L^\ell} \sum_{X,\bar{X}} \frac{1}{k} \sum_{x\in\bar{X}} I(\mu X, x) = \frac{1}{k} \sum_{i=1}^{L} \frac{1}{C_L^\ell} \underbrace{\sum_{X,\bar{X}} [x_i \in \bar{X}]\, I(\mu X, x_i)}_{N_i}. \tag{5.5}$$

Внутренняя сумма, обозначенная через N_i, выражает число разбиений выборки $\mathbb{X}$, при которых объект x_i оказывается в контрольной подвыборке, и построенный по обучающей подвыборке алгоритм допускает на нём ошибку.

Оценим N_i, воспользовавшись тем, что если алгоритм a монотонный и допускает ошибку на объекте x_i, то он допускает ошибку и на всех объектах из клина W_i.

В зависимости от соотношения мощности клина w_i и степени немонотонности выборки возможны два случая.

Если $w_i \geqslant \theta L + k$, то ни при каком разбиении монотонный алгоритм не будет ошибаться на x_i, поскольку $\theta L + k$ есть максимальное число ошибок, которое может допустить монотонная функция на всей выборке $\mathbb{X}$. Это вытекает из допущения, что метод μ строит алгоритм с минимальным числом ошибок на обучающей выборке в классе всех монотонных функций. Минимальное число ошибок на любой подвыборке X не превосходит минимального числа ошибок на всей выборке $\mathbb{X}$. Следовательно число ошибок на обучении не превышает θL. Таким образом, в этом случае $N_i = 0$.

Рассмотрим второй случай, когда $w_i < \theta L + k$. Пусть s — число объектов из W_i, находящихся в обучающей подвыборке,

$$\max\{0, w_i - k + 1\} \leqslant s \leqslant \min\{\theta L, \ell, w_i\}.$$

Имеется $C_{w_i}^s$ способов выбрать s обучающих объектов из клина W_i. Для каждого из этих способов имеется $C_{L-1-w_i}^{\ell-s}$ вариантов выбрать $\ell - s$ обучающих объектов из множества $\mathbb{X} \setminus (W_i \cup \{x_i\})$. В итоге получаем оценку числа разбиений:

$$N_i \leqslant \sum_{s=\max\{0,w_i-k+1\}}^{\min\{\theta L,\ell,w_i\}} C_{w_i}^s C_{L-1-w_i}^{\ell-s}. \tag{5.6}$$

Представим N в виде $N = C_L^\ell = \frac{L}{k} C_{L-1}^\ell$ и подставим оценку (5.6) в (5.5), учитывая, что $N_i = 0$ при $w_i \geqslant \theta L + k$:

$$\mathrm{CCV} \leqslant \frac{1}{k} \sum_{\substack{i=1 \\ w_i<\theta L+k}}^{L} \frac{k}{L} \sum_{s=\max\{0,w_i-k+1\}}^{\min\{\theta L,\ell,w_i\}} \frac{C_{w_i}^s C_{L-1-w_i}^{\ell-s}}{C_{L-1}^{\ell}}.$$

Применяя определение профиля монотонности, получаем оценку (5.4). ■

Следствие 5.4.1. *Оценка (5.4) монотонно не убывает по θ, достигая наименьшего значения при $\theta = 0$, когда выборка монотонна и метод μ является корректным на генеральной выборке $\mathbb{X}$ (то есть $\nu(\mu X, X) = 0$ для всех $X \in [\mathbb{X}]^\ell$):*

$$\mathrm{CCV}(\mu, \mathbb{X}) \leqslant \sum_{m=0}^{k-1} M(m, \mathbb{X}) \frac{C_{L-1-m}^{\ell}}{C_{L-1}^{\ell}}. \tag{5.7}$$

Оценка (5.4), в отличие от завышенных сложностных оценок, никогда не превышает 1. Наибольшее значение 1 достигается, если $w_i = 0$ для всех $i = 1, \ldots, L$. Это тот случай, когда оба класса состоят из попарно несравнимых объектов, и вся выборка распадается на две антицепи. Наименьшее значение достигается, когда выборка монотонна и линейно упорядочена. В этом случае число клиньев мощности w не превышает 2 для всех $w = 1, \ldots, k$, откуда вытекает $\mathrm{CCV} \leqslant 2/\ell$.

Комбинаторный множитель в (5.4) убывает с ростом m быстрее геометрической прогрессии. Чтобы обеспечить малое значение функционала CCV, достаточно потребовать, чтобы функция $M(m, \mathbb{X})$ принимала малые значения при малых m. При больших m её рост компенсируется комбинаторным множителем. Таким образом, качество монотонного классификатора тем выше, чем меньше объектов имеют клинья небольшой мощности. Для этого отношение порядка на множестве объектов X должно быть близко к линейному вблизи границы классов. Форма профиля монотонности может рассматриваться как формальное выражение априорной информации о плотности отношения порядка [78, 79] вблизи границы классов.

Ёмкость класса монотонных классификаторов бесконечна, поскольку на выборке длины L, состоящей из попарно несравнимых элементов, реализуется ровно 2^L дихотомий. Таким образом, классическая теория Вапника-Червоненкиса вообще не даёт оценок качества для данного случая. Известно [204], что *эффективная ёмкость* класса монотонных функций не превосходит длины максимальной антицепи в выборке $\mathbb{X}$. Оценка (5.4) существенно более точная, особенно при малых выборках.

Интересно отметить большое структурное сходство оценок (5.1) и (5.7), полученных для таких различных, на первый взгляд, априорных ограничений, как компактность и монотонность.

5.3.3 Монотонные композиции алгоритмов классификации

В данном параграфе рассматриваются монотонные композиции алгоритмов классификации, предложенные автором в [16] и метод улучшения их обобщающей способности, разработанный И. Гузом [36, 37] на основе полученных выше оценок CCV. Краткое изложение этих результатов приводится здесь с целью продемонстрировать практическое применение оценок CCV.

Рассмотрим задачу классификации на два класса $\mathbb{Y} = \{0, 1\}$ и *алгоритмическую композицию* с *монотонной корректирующей операцией* [76, 16, 17]:

$$a(x) = F\big(b_1(x), \ldots, b_T(x)\big), \tag{5.8}$$

где отображения $b_t \colon \mathbb{X} \to \mathbb{R},\ t = 1, \ldots, T$ называются базовыми *алгоритмическими операторами*, $F \colon \mathbb{R}^T \to \mathbb{Y}$ является монотонно неубывающей функцией всех своих T аргументов и называется *корректирующей операцией*. Монотонную корректирующую операцию можно рассматривать как обобщение взвешенного голосования: любая линейная выпуклая корректирующая операция является монотонной; обратное в общем случае неверно, см. также стр. 89.

Процесс построения композиции (5.8) в общих чертах напоминает бустинг. Базовые алгоритмические операторы строятся последовательно. На каждой итерации t фиксируются предыдущие операторы $b_1, \ldots, b_{t-1}$, добавляется очередной оператор b_t и перестраивается корректирующая операция $F(b_1, \ldots, b_T)$. Основное отличие от бустинга заключается в том, что критерием настройки оператора b_t по обучающей выборке X является минимизация числа *дефектных пар*.

Дефектной парой набора операторов $B_t = \{b_1, \ldots, b_t\}$ относительно обучающей выборки X называется пара объектов (x_i, x_j) из X такая, что $y_i < y_j$ и $b(x_i) \geqslant b(x_j)$ для всех $b \in B_t$. Очевидно, любая монотонная композиция $a(x)$ допускает ошибку хотя бы на одном из двух объектов дефектной пары.

Для устранения дефектной пары (x_i, x_j) набора операторов B_{t-1} оператор b_t должен удовлетворять условию $b_t(x_i) < b_t(x_j)$. Для этого достаточно, чтобы оператор b_t не ошибался на объектах x_i и x_j. Аналогично бустингу, минимизация общего числа ошибок композиции сводится к пересчёту весов w_i всех обучающих объектов $x_i \in X$ перед построением оператора b_t. Доказано [16], что вес w_i должен быть пропорционален числу дефектных пар набора B_{t-1}, в которые входит объект x_i.

Эксперименты [36] на реальных задачах классификации показали, что монотонная композиция повышает обобщающую способность, обходясь лишь несколькими базовыми операторами, тогда как алгоритму AdaBoost обычно требуются десятки и сотни базовых алгоритмов, чтобы достичь сопоставимого качества классификации. В то же время оказалось, что монотонная композиция гораздо быстрее переобучается: после добавления первых трёх базовых операторов качество классификации на контрольных данных в большинстве случаев начинало ухудшаться.

Это явление легко объясняется с точки зрения оценки CCV (5.4). Критерий настройки очередного оператора b_t направлен на разрушение дефектных пар предыдущих операторов B_{t-1}, но он не стремится сохранять *правильные пары* набора B_{t-1}, из которых как раз и образуются клинья.

Правильной парой набора операторов $B_t = \{b_1, \ldots, b_t\}$ относительно обучающей выборки X называется пара объектов (x_i, x_j) из X такая, что $y_i < y_j$ и $b(x_i) < b(x_j)$ для всех $b \in B_t$.

В модифицированном методе обучения монотонной композиции, предложенном И. Гузом [37], вес w_i полагается равным суммарному числу дефектных и правильных пар набора B_{t-1}, в которых участвует объект x_i. Предварительные эксперименты показывают, что эта модификация действительно снижает переобучение и позволяет увеличивать число базовых алгоритмических операторов до 8–10.

5.4 Основные выводы

1. Определяется функционал *полного скользящего контроля* (CCV) как средняя по всем разбиениям частота ошибок на контрольной выборке.
2. Вводится понятие *профиля компактности* выборки, которое можно рассматривать как строгую формализацию эвристической «гипотезы компактности». Доказывается точная формула CCV для метода ближайшего соседа, зависящая от профиля компактности. Предлагается метод отбора эталонных объектов, оптимизирующий CCV. Эксперименты показывают, что данный метод не переобучается. Интересным для приложений побочным результатом является разделение всех объектов на три категории: шумовые, неинформативные и эталонные.
3. Для задач классификации с априорными ограничениями монотонности (или почти-монотонности) целевой зависимости вводится понятие *профиля монотонности* выборки. Доказываются слабо завышенные верхние оценки функционала CCV. Описывается метод построения монотонных корректирующих операций, оптимизирующий полученную верхнюю оценку.

Заключение

Результаты, выносимые на защиту.

1. Слабая вероятностная аксиоматика, основанная на единственном вероятностном предположении — о независимости наблюдений в конечной выборке. Общая постановка задач эмпирического предсказания.
2. VC-оценки вероятности переобучения, учитывающие степень некорректности метода обучения.
3. Методика эмпирического измерения факторов завышенности VC-оценок вероятности переобучения.
4. Метод получения точных оценок вероятности переобучения, основанный на выделении множеств порождающих и запрещающих объектов для каждого алгоритма в семействе.
5. Рекуррентный алгоритм вычисления точных, верхних и нижних оценок вероятности переобучения.
6. Блочный метод вычисления точных оценок вероятности переобучения.
7. Точные оценки вероятности переобучения для ряда модельных семейств алгоритмов: слоя и интервала булева куба, монотонных и унимодальных цепочек, единичной окрестности.
8. Оценка вероятности переобучения, учитывающая профиль расслоения и связности семейства алгоритмов.
9. Точные оценки полного скользящего контроля для метода ближайшего соседа, выражающиеся через профиль компактности выборки.
10. Верхние оценки полного скользящего контроля для семейства монотонных алгоритмов, выражающиеся через профиль монотонности выборки.

Направления дальнейших исследований.

- Получение точных оценок вероятности переобучения для более широкого класса модельных семейств.

- Получение точных или слабо завышенных оценок вероятности переобучения для реальных семейств в случаях «наихудших» и «типичных» выборок.
- Разработка оптимизационных методов обучения, позволяющих управлять качеством результирующего алгоритма в ходе итерационного процесса обучения.
- Уточнение границ применимости слабой вероятностной аксиоматики.

Список обозначений

$U \sqcup V$ — дизъюнктное объединение, $U \cup V$ при условии $U \cap V = \varnothing$;

$[U]^\ell$ — множество всех ℓ-элементных подмножеств множества U;

$[x] = \begin{cases} 0, & x=\text{ложь}; \\ 1, & x=\text{истина}; \end{cases}$ — индикаторная (характеристическая) функция предиката x;

$x_+ = \begin{cases} 0, & x \leqslant 0; \\ x, & x > 0. \end{cases}$ — операция положительной срезки;

$\lceil x \rceil$ — функция «потолок» — минимальное целое, не меньшее x;

$\lfloor x \rfloor$ — функция «пол» — максимальное целое, не большее x;

$C_n^k = \frac{n!}{k!\,(n-k)!}$ — биномиальные коэффициенты;

$\operatorname{Arg}\min\limits_{u \in U} f(u) = \big\{u' \in U\colon f(u') \leqslant f(u),\ \forall u \in U\big\}$ — множество точек минимума;

$\arg\min\limits_{u \in U} f(u)$ — произвольный элемент из множества точек минимума;

Предметный указатель

Д

Е

З

И

К

Л

М

Н

О

П

Р

С

Т

Я

Список иллюстраций

Список таблиц

Список литературы

[1] *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989.

[2] *Айзерман М. А., Браверман Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. — 320 pp.

[3] *Алимов Ю. И.* Альтернатива методу математической статистики. — Знание, 1980.

[4] *Беляев Ю. К.* Вероятностные методы выборочного контроля. — М.: Наука, 1975.

[5] *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. — М.: Наука, 1983.

[6] *Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А.* Сходство и компактность // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 89–92.

[7] *Ботов П. В.* Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 7–10.

[8] *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.

[9] *Вапник В. Н., Червоненкис А. Я.* О равномерной сходимости частот появления событий к их вероятностям // *ДАН СССР*. — 1968. — Т. 181, № 4. — С. 781–784.

[10] *Вапник В. Н., Червоненкис А. Я.* О равномерной сходимости частот появления событий к их вероятностям // *Теория вероятностей и ее применения*. — 1971. — Т. 16, № 2. — С. 264–280.

[11] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1974.

[12] *Венжега А. В., Ументаев С. А., Орлов А. А., Воронцов К. В.* Проблема переобучения при отборе признаков в линейной регрессии с фиксированными коэффициентами // Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 90–93.

[13] Вероятность и математическая статистика: Энциклопедия / Под ред. Ю. В. Прохорова. — М.: Большая российская энциклопедия, 2003.

[14] *Вовк В. Г., Шейфер Г. Р.* Вклад А. Н. Колмогорова в основания теории вероятностей // *Проблемы передачи информации*. — 2003. — Т. 39, № 1. — С. 24–35.

[15] *Воронцов К. В.* Качество восстановления зависимостей по эмпирическим данным // Математические методы распознавания образов: 7-ая Всерос. конф. Тезисы докл. — Пущино, 1995. — С. 24–26.

[16] *Воронцов К. В.* О проблемно-ориентированной оптимизации базисов задач распозна-

вания // *ЖВМ и МФ.* — 1998. — Т. 38, № 5. — С. 870–880.

[17] *Воронцов К. В.* Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания // *ЖВМ и МФ.* — 2000. — Т. 40, № 1. — С. 166–176.

[18] *Воронцов К. В.* Оценка качества монотонного решающего правила вне обучающей выборки // Интеллектуализация обработки информации: Тезисы докл. — Симферополь, 2002. — С. 24–26.

[19] *Воронцов К. В.* О комбинаторном подходе к оценке качества обучения алгоритмов // Математические методы распознавания образов: 11-ая Всерос. конф. Тезисы докл. — Пущино, 2003. — С. 47–49.

[20] *Воронцов К. В.* Комбинаторные обоснования обучаемых алгоритмов // *ЖВМиМФ.* — 2004. — Т. 44, № 11. — С. 2099–2112.

[21] *Воронцов К. В.* Комбинаторные оценки качества обучения по прецедентам // *Докл. РАН.* — 2004. — Т. 394, № 2. — С. 175–178.

[22] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / Под ред. О. Б. Лупанов. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

[23] *Воронцов К. В.* Комбинаторный подход к повышению качества логических классификаторов // Интеллектуализация обработки информации: Тезисы докл. — Симферополь, 2004. — С. 44.

[24] *Воронцов К. В.* Обзор современных исследований по проблеме качества обучения алгоритмов // *Таврический вестник информатики и математики.* — 2004. — № 1. — С. 5–24.

[25] *Воронцов К. В.* Слабая вероятностная аксиоматика и надёжность эмпирических предсказаний // Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 21–25.

[26] *Воронцов К. В.* Комбинаторный подход к проблеме переобучения // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 18–21.

[27] *Воронцов К. В.* Методы машинного обучения, основанные на индукции правил // Труды семинара «Знания и онтологии ELSEWHERE 2009», ассоциированного с 17-й международной конференцией по понятийным структурам ICCS-17, Москва, 21–26 июля. — Высшая школа экономики, 2009. — С. 57–71.

[28] *Воронцов К. В.* Точные оценки вероятности переобучения // *Доклады РАН.* — 2009. — Т. 429, № 1. — С. 15–18.

[29] *Воронцов К. В., Ивахненко А. А.* Эмпирические оценки локальной функции роста в задачах поиска логических закономерностей // *Искусственный Интеллект.* — 2006. — С. 281–284.

[30] *Воронцов К. В., Ивахненко А. А., Инякин А. С., Лисица А. В., Минаев П. Ю.* «Полигон» — распределённая система для эмпирического анализа задач и алгоритмов классификации // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 503–506.

[31] *Воронцов К. В., Инякин А. С., Лисица А. В.* Система эмпирического измерения каче-

ства алгоритмов классификации // Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 577–580.

[32] *Воронцов К. В., Колосков А. О.* Профили компактности и выделение опорных объектов в метрических алгоритмах классификации // *Искусственный Интеллект.* — 2006. — С. 30–33.

[33] *Гаек Я., Шидак З.* Теория ранговых критериев. — М.: Наука, 1971.

[34] *Головко В. А.* Нейронные сети: обучение, организация и применение. — М.: ИПРЖР, 2001.

[35] *Гоппа В. Д.* Введение в алгебраическую теорию информации. — М.: Наука, 1995.

[36] *Гуз И. С.* Нелинейные монотонные композиции классификаторов // Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 111–114.

[37] *Гуз И. С.* Исследование обобщающей способности семейства монотонных функций // *Сборник трудов МФТИ. Моделирование и обработка информации.* — 2008.

[38] *Гуров С. И.* Точечная оценка вероятности 0-события // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 22–25.

[39] *Донской В. И., Башта А. И.* Дискретные модели принятия решений при неполной информации. — Симферополь: Таврия, 1992. — 166 с.

[40] *Дуда Р., Харт П.* Распознавание образов и анализ сцен. — М.: Мир, 1976.

[41] *Дэйвид Г.* Порядковые статистики. — М.: Наука, 1979. — 336 с.

[42] *Дюличева Ю. Ю.* Оценка VCD r-редуцированного эмпирического леса // *Таврический вестник информатики и математики.* — 2003. — № 1. — С. 31–42.

[43] *Журавлёв Ю. И.* Непараметрические задачи распознавания образов // *Кибернетика.* — 1976. — № 6.

[44] *Журавлёв Ю. И.* Экстремальные алгоритмы в математических моделях для задач распознавания и классификации // *Доклады АН СССР. Математика.* — 1976. — Т. 231, № 3.

[45] *Журавлёв Ю. И.* Корректные алгебры над множествами некорректных (эвристических) алгоритмов. Часть I // *Кибернетика.* — 1977. — № 4. — С. 5–17.

[46] *Журавлёв Ю. И.* Корректные алгебры над множествами некорректных (эвристических) алгоритмов. Часть II // *Кибернетика.* — 1977. — № 6. — С. 21–27.

[47] *Журавлёв Ю. И.* Корректные алгебры над множествами некорректных (эвристических) алгоритмов. Часть III // *Кибернетика.* — 1978. — № 2. — С. 35–43.

[48] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // *Проблемы кибернетики.* — 1978. — Т. 33. — С. 5–68.

[49] *Журавлёв Ю. И., Рудаков К. В.* Об алгебраической коррекции процедур обработки (преобразования) информации // *Проблемы прикладной математики и информатики.* — 1987. — С. 187–198.

[50] *Журавлёв Ю. И., Рязанов В. В., Сенько О. В.* «Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006.

[51] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.

[52] *Зухба А. В.* Np-полнота задачи оптимального отбора эталонных объектов в методе ближайшего соседа // Труды 52-й научной конференции МФТИ «Современные про-

блемы фундаментальных и прикладных наук». Часть VII. Управление и прикладная математика. — Т. 2. — М.: МФТИ, 2009. — С. 61–63.

[53] *Иванов М. Н., Воронцов К. В.* Отбор эталонов, основанный на минимизации функционала полного скользящего контроля // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 119–122.

[54] *Ивахненко А. А., Воронцов К. В.* Верхние оценки переобученности и профили разнообразия логических закономерностей // Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 33–37.

[55] *Катериночкина Н. Н.* Методы выделения максимальной совместной подсистемы системы линейных неравенств. Сообщение по прикладной математике. — Москва: Вычислительный центр РАН, 1997.

[56] *Кобзарь А. И.* Прикладная математическая статистика. — М.: Физматлит, 2006.

[57] *Колмогоров А. Н.* Комбинаторные основания теории информации и исчисления вероятностей // *Успехи математических наук*. — 1983. — Т. 38, № 4. — С. 27–36.

[58] *Колмогоров А. Н.* Теория информации и теория алгоритмов / Под ред. Ю. В. Прохоров. — М.: Наука, 1987. — 304 с.

[59] *Кочедыков Д. А.* Структуры сходства в семействах алгоритмов классификации и оценки обобщающей способности // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 45–48.

[60] *Кочедыков Д. А., Ивахненко А. А., Воронцов К. В.* Система кредитного скоринга на основе логических алгоритмов классификации // Математические методы распознавания образов-12. — М.: МАКС Пресс, 2005. — С. 349–353.

[61] *Кочедыков Д. А., Ивахненко А. А., Воронцов К. В.* Применение логических алгоритмов классификации в задачах кредитного скоринга и управления риском кредитного портфеля банка // Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 484–488.

[62] *Лбов Г. С.* Методы обработки разнотипных экспериментальных данных. — Новосибирск: Наука, 1981.

[63] *Матросов В. Л.* Корректные алгебры ограниченной ёмкости над множествами некорректных алгоритмов // *ДАН СССР*. — 1980. — Т. 253, № 1. — С. 25–30.

[64] *Матросов В. Л.* О критериях полноты модели алгоритмов вычисления оценок и её алгебраических замыканий // *ДАН СССР*. — 1981. — Т. 258, № 4. — С. 791–796.

[65] *Матросов В. Л.* Оптимальные алгоритмы в алгебраических замыканиях операторов вычисления оценок // *ДАН СССР*. — 1982. — Т. 262, № 4. — С. 818–822.

[66] *Матросов В. Л.* Ёмкость алгебраических расширений модели алгоритмов вычисления оценок // *ЖВМиМФ*. — 1984. — Т. 24, № 11. — С. 1719–1730.

[67] *Матросов В. Л.* Нижние границы ёмкости многомерных алгебр алгоритмов вычисления оценок // *ЖВМиМФ*. — 1984. — Т. 24, № 12. — С. 1881–1892.

[68] *Матросов В. Л.* Ёмкость алгоритмических многочленов над множеством алгоритмов вычисления оценок // *ЖВМиМФ*. — 1985. — Т. 25, № 1. — С. 122–133.

[69] *Минский М., Пайперт С.* Персепроны. — М.: Мир, 1971.

[70] Нейроинформатика / А. Н. Горбань, В. Л. Дунин-Барковский, А. Н. Кирдин, Е. М. Миркес, А. Ю. Новоходько, Д. А. Россиев, С. А. Терехов и др. — Новосибирск: Наука,

1998. — 296 с.

[71] *Норушис А.* Построение логических (древообразных) классификаторов методами нисходящего поиска (обзор) // Статистические проблемы управления. Вып. 93 / Под ред. Ш. Раудис. — Вильнюс, 1990. — С. 131–158.

[72] *Орлов А. И.* Эконометрика: Учебник для вузов. — М.: Экзамен, 2003. — 576 с.

[73] *Орлов А. И.* Нечисловая статистика. — М.: МЗ-Пресс, 2004.

[74] *Райгородский А. М.* Экстремальные задачи теории графов и анализ данных. — М.: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2008. — 118 с.

[75] *Растригин Л. А., Эренштейн Р. Х.* Коллективные правила распознавания. — М.: Энергия, 1981. — 244 рр.

[76] *Рудаков К. В., Воронцов К. В.* О методах оптимизации и монотонной коррекции в алгебраическом подходе к проблеме распознавания // *Докл. РАН.* — 1999. — Т. 367, № 3. — С. 314–317.

[77] *Рязанов В. В., Сенько О. В.* О некоторых моделях голосования и методах их оптимизации // *Распознавание, классификация, прогноз.* — 1990. — Т. 3. — С. 106–145.

[78] *Сёмочкин А. Н.* Линейные достроения частичного порядка на конечных множествах // *Деп. в ВИНИТИ.* — 1998. — № 2964–В98. — С. 19.

[79] *Сёмочкин А. Н.* Оценки функционала качества для класса алгоритмов с универсальными ограничениями монотонности // *Деп. в ВИНИТИ.* — 1998. — № 2965–В98. — С. 20.

[80] *Смирнов Н. В.* Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках // *Бюлл. Московского ун-та, серия А.* — 1939. — № 2. — С. 3–14.

[81] *Ульянов Ф. М., Воронцов К. В.* Проблема переобучения функций близости при построении алгоритмов вычисления оценок // Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 105–108.

[82] *Успенский В. А.* Четыре алгоримтических лица случайности. — М.: Изд-во МЦНМО, 2009. — 48 с.

[83] *Фрей А. И.* Точные оценки вероятности переобучения для симметричных семейств алгоритмов // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 66–69.

[84] *Хайкин С.* Нейронные сети: полный курс, 2-е издание. — М.: Издательский дом «Вильямс», 2006.

[85] *Цюрмасто П. А., Воронцов К. В.* Анализ сходства алгоритмов классификации в оценках обобщающей способности // Интеллектуализация обработки информации (ИОИ-2008): Тезисы докл. — Симферополь: КНЦ НАН Украины, 2008. — С. 232–234.

[86] *Шень А. Х.* Частотный подход к определению понятия случайной последовательности // Семиотика и информатика. — М.: ВИНИТИ, 1982. — Т. 18. — С. 14–42.

[87] *Эфрон Б.* Нетрадиционные методы многомерного статистического анализа. — М: Финансы и статистика, 1988.

[88] *Abu-Mostafa Y. S.* Hints // *Neural Computation.* — 1995. — Vol. 7, no. 4. — Pp. 639–671.

[89] *Ambroladze A., Parrado-Hernández E., Shawe-Taylor J.* Tighter PAC-Bayes bounds //

Advances in Neural Information Processing Systems 19 / Ed. by B. Schölkopf, J. Platt, T. Hoffman. — Cambridge, MA: MIT Press, 2007. — Pp. 9–16.

[90] *Anthony M.* Uniform glivenko-cantelli theorems and concentration of measure in the mathematical modelling of learning: Tech. Rep. LSE-CDAM-2002-07: 2002.

[91] *Anthony M., Bartlett P. L.* Neural Network Learning: Theoretical Foundations. — Cambridge University Press, Cambridge, 1999.

[92] *Anthony M., Shawe-Taylor J.* A result of Vapnik with applications // *Discrete Applied Mathematics*. — 1993. — Vol. 47, no. 2. — Pp. 207–217.

[93] *Antos A., Kegl B., Linder T., Lugosi G.* Data-dependent margin-based generalization bounds for classification // *Journal of Machine Learning Research*. — 2002. — Pp. 73–98.

[94] *Asuncion A., Newman D.* UCI machine learning repository: Tech. rep.: University of California, Irvine, School of Information and Computer Sciences, 2007.

[95] *Audibert J.-Y.* PAC-Bayesian Statistical Learning Theory: Ph.D. thesis. — 2004.

[96] *Bartlett P.* Lower bounds on the Vapnik-Chervonenkis dimension of multi-layer threshold networks // Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory. — ACM Press, New York, NY, 1993. — Pp. 144–150.

[97] *Bartlett P.* The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network // *IEEE Transactions on Information Theory*. — 1998. — Vol. 44, no. 2. — Pp. 525–536.

[98] *Bartlett P., Bousquet O., Mendelson S.* Localized rademacher complexities // COLT: 15th Annual Conference on Computational Learning Theory. — Springer, Berlin, 2002. — Pp. 44–58.

[99] *Bartlett P., Bousquet O., Mendelson S.* Local rademacher complexities. — Vol. 33. — Institute of Mathematical Statistics, 2005. — P. 1497–1537.

[100] *Bartlett P., Shawe-Taylor J.* Generalization performance of support vector machines and other pattern classifiers // Advances in Kernel Methods. — MIT Press, Cambridge, USA, 1999. — Pp. 43–54.

[101] *Bartlett P. L.* For valid generalization the size of the weights is more important than the size of the network // Advances in Neural Information Processing Systems / Ed. by M. C. Mozer, M. I. Jordan, T. Petsche. — Vol. 9. — The MIT Press, 1997. — P. 134.

[102] *Bartlett P. L., Long P. M., Williamson R. C.* Fat-shattering and the learnability of real-valued functions // *Journal of Computer and System Sciences*. — 1996. — Vol. 52, no. 3. — Pp. 434–452.

[103] *Bartlett P. L., Mendelson S., Philips P.* Local complexities for empirical risk minimization // COLT: 17th Annual Conference on Learning Theory / Ed. by J. Shawe-Taylor, Y. Singer. — Springer-Verlag, 2004. — Pp. 270–284.

[104] *Bauer M., Godreche C., Luck J. M.* Statistics of persistent events in the binomial random walk: Will the drunken sailor hit the sober man? // *J.STAT.PHYS.* — 1999. — Vol. 96. — P. 963.

[105] *Bax E. T.* Similar classifiers and VC error bounds: Tech. Rep. CalTech-CS-TR97-14: 1997.

[106] *Bontempi G., Birattari M.* A bound on the cross-validation estimate for algorithm assessment // Eleventh Belgium/Netherlands Conference on Artificial Intelligence (BNAIC). — 1999. — Pp. 115–122.

[107] *Bottou L., Cortes C., Vapnik V.* On the effective VC dimension. — 1994.

[108] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances // *ESAIM: Probability and Statistics.* — 2005. — no. 9. — Pp. 323–375.

[109] *Boucheron S., Lugosi G., Massart P.* A sharp concentration inequality with applications // *Random Structures and Algorithms.* — 2000. — Vol. 16, no. 3. — Pp. 277–292.

[110] *Boucheron S., Lugosi G., Massart P.* Concentration inequalities using the entropy method // *The Annals of Probability.* — 2003. — Vol. 31, no. 3.

[111] *Bousquet O.* Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms: Ph.D. thesis / Ecole Polytechnique, France. — 2002.

[112] *Bousquet O., Elisseeff A.* Algorithmic stability and generalization performance // Advances in Neural Information Processing Systems 13. — 2001. — Pp. 196–202.

[113] *Bousquet O., Elisseeff A.* Stability and generalization // *Journal of Machine Learning Research.* — 2002. — no. 2. — Pp. 499–526.

[114] *Breiman L.* Bagging predictors // *Machine Learning.* — 1996. — Vol. 24, no. 2. — Pp. 123–140.

[115] *Breiman L.* Bias, variance, and arcing classifiers: Tech. Rep. 460: Statistics Department, University of California, 1996.

[116] *Breiman L.* Arcing classifiers // *The Annals of Statistics.* — 1998. — Vol. 26, no. 3. — Pp. 801–849.

[117] *Breiman L., Friedman J., Stone C. J., Olshen R. A.* Classification and Regression Trees. — Belmont, California, U.S.A.: Wadsworth Publishing Company, 1984.

[118] *Burges C. J. C.* A tutorial on support vector machines for pattern recognition // *Data Mining and Knowledge Discovery.* — 1998. — Vol. 2, no. 2. — Pp. 121–167.

[119] *Chernoff H.* A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations // *Annals of Math. Stat.* — 1952. — Vol. 23. — Pp. 493–509.

[120] *Chvátal V.* The tail of the hypergeometric distribution // *Discrete Mathematics.* — 1979. — Vol. 25, no. 3. — Pp. 285–287.

[121] *Cohen W. W.* Fast effective rule induction // Proc. of the 12th International Conference on Machine Learning, Tahoe City, CA. — Morgan Kaufmann, 1995. — Pp. 115–123.

[122] *Cohen W. W., Singer Y.* A simple, fast and effective rule learner // Proc. of the 16 National Conference on Artificial Intelligence. — 1999. — Pp. 335–342.

[123] *Cortes C., Vapnik V.* Support-vector networks // *Machine Learning.* — 1995. — Vol. 20, no. 3. — Pp. 273–297.

[124] *Devroye L. P., Wagner T. J.* Distribution-free inequalities for the deleted and holdout error estimates // *IEEE Transactions on Information Theory.* — 1979. — Vol. 25, no. 2. — Pp. 202–207.

[125] *Devroye L. P., Wagner T. J.* Distribution-free performance bounds for potential function rules // *IEEE Transactions on Information Theory.* — 1979. — Vol. 25, no. 5. — Pp. 601–604.

[126] *Dohmen K., Tittmann P.* Bonferroni-type inequalities and binomially bounded functions // *Electronic Notes in Discrete Mathematics. 6th Czech-Slovak International Symposium on Combinatorics, Graph Theory, Algorithms and Applications.* — 2007. — Vol. 28. — Pp. 91–93.

[127] *Efron B.* The Jackknife, the Bootstrap, and Other Resampling Plans. — SIAM, Philadelphia, 1982.

[128] *Elisseeff A., Evgeniou T., Pontil M.* Stability of randomized learning algorithms // *Journal of Machine Learning Research.* — 2005. — no. 6. — Pp. 55–79.

[129] *Evgeniou T., Pontil M., Elisseeff A.* Leave one out error, stability, and generalization of voting combinations of classifiers: Tech. Rep. 2001-21-TM: INSEAD, 2001.

[130] *Freund Y.* Boosting a weak learning algorithm by majority // COLT: Proceedings of the Workshop on Computational Learning Theory. — Morgan Kaufmann Publishers, 1990.

[131] *Freund Y.* Self bounding learning algorithms // COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers. — 1998.

[132] *Freund Y., Schapire R. E.* A decision-theoretic generalization of on-line learning and an application to boosting // European Conference on Computational Learning Theory. — 1995. — Pp. 23–37.

[133] *Freund Y., Schapire R. E.* Experiments with a new boosting algorithm // International Conference on Machine Learning. — 1996. — Pp. 148–156.

[134] *Freund Y., Schapire R. E.* Discussion of the paper "Arcing classifiers" by Leo Breiman // *The Annals of Statistics.* — 1998. — Vol. 26, no. 3. — Pp. 824–832.

[135] *Fürnkranz J., Flach P. A.* Roc 'n' rule learning-towards a better understanding of covering algorithms // *Machine Learning.* — 2005. — Vol. 58, no. 1. — Pp. 39–77.

[136] *Galambos J., Simonelli I.* Bonferroni-type Inequalities with Applications. — Springer, 1996.

[137] *Germain P., Lacasse A., Laviolette F., Marchand M.* A PAC-Bayes risk bound for general loss functions // *Advances in neural information processing systems.* — 2007. — no. 19. — Pp. 449–456.

[138] *Germain P., Lacasse A., Laviolette F., Marchand M.* PAC-Bayesian learning of linear classifiers // ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning. — ACM, 2009. — Pp. 353–360.

[139] *Golea M., Bartlett P., Lee W. S., Mason L.* Generalization in decision trees and DNF: Does size matter? // Advances in Neural Information Processing Systems / Ed. by M. I. Jordan, M. J. Kearns, S. A. Solla. — Vol. 10. — The MIT Press, 1998.

[140] *Grove A. J., Schuurmans D.* Boosting in the limit: Maximizing the margin of learned ensembles // AAAI/IAAI. — 1998. — Pp. 692–699.

[141] *Hebb D.* The organization of behavior. — New York: Wiley, 1949.

[142] *Herbrich R., C.Williamson R.* Algorithmic luckiness // *Journal of Machine Learning Research.* — 2002. — no. 3. — Pp. 175–212.

[143] *Herbrich R., Williamson R. C.* Learning and generalization: theoretical bounds. — Cambridge, MA, USA: MIT Press, 2002. — Pp. 619–623.

[144] *Holden S. B.* Cross-validation and the pac learning model: Tech. Rep. RN/96/64: Dept. of CS, Univ. College, London, 1996.

[145] *Hosmer D. W., Lemeshow S.* Applied Logistic Regression, second ed. — New York: Wiley, 2000.

[146] *Jackson J.* On the efficiency of noise-tolerant pac algorithms derived from statistical queries // *Annals of Mathematics and Artificial Intelligence.* — 2003. — Vol. 39, no. 3. — Pp. 291–313.

[147] *Jackson J., Shamir E., Shwartzman C.* Learning with queries corrupted by classification noise // *Discrete Applied Mathematics.* — 1999. — Vol. 92, no. 2-3. — Pp. 157–175.

[148] *Jacobs R. A., Jordan M. I., Nowlan S. J., Hinton G. E.* Adaptive mixtures of local experts // *Neural Computation.* — 1991. — no. 3. — Pp. 79–87.

[149] *Karpinski M., Macintyre A.* Polynomial bounds for VC dimension of sigmoidal neural networks // 27th ACM Symposium on Theory of Computing, Las Vegas, Nevada, US. — 1995. — Pp. 200–208.

[150] *Kearns M.* Efficient noise-tolerant learning from statistical queries // Proceedings of the 25-th annual ACM symposium on Theory of computing, May 16-18, 1993, San Diego, California, United States. — ACM, 1993. — Pp. 392–401.

[151] *Kearns M.* A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split // Advances in Neural Information Processing Systems / Ed. by D. S. Touretzky, M. C. Mozer, M. E. Hasselmo. — Vol. 8. — The MIT Press, 1996. — Pp. 183–189.

[152] *Kearns M.* Efficient noise-tolerant learning from statistical queries // *Journal of the ACM.* — 1998. — Vol. 45, no. 6. — Pp. 983–1006.

[153] *Kearns M., Valiant L. G.* Cryptographic limitations on learning Boolean formulae and finite automata // Proc. of the 21st Annual ACM Symposium on Theory of Computing. — 1989. — Pp. 433–444.

[154] *Kearns M. J., Mansour Y., Ng A. Y., Ron D.* An experimental and theoretical comparison of model selection methods // 8th Conf. on Computational Learning Theory, Santa Cruz, California, US. — 1995. — Pp. 21–30.

[155] *Kearns M. J., Ron D.* Algorithmic stability and sanity-check bounds for leave-one-out cross-validation // Computational Learning Theory. — 1997. — Pp. 152–162.

[156] *Kearns M. J., Schapire R. E.* Efficient distribution-free learning of probabilistic concepts // Computational Learning Theory and Natural Learning Systems, Volume I: Constraints and Prospect, edited by Stephen Jose Hanson, George A. Drastal, and Ronald L. Rivest, Bradford/MIT Press. — 1994. — Vol. 1.

[157] *Kohavi R.* A study of cross-validation and bootstrap for accuracy estimation and model selection // 14th International Joint Conference on Artificial Intelligence, Palais de Congres Montreal, Quebec, Canada. — 1995. — Pp. 1137–1145.

[158] *Koltchinskii V.* Rademacher penalties and structural risk minimization // *IEEE Transactions on Information Theory.* — 2001. — Vol. 47, no. 5. — Pp. 1902–1914.

[159] *Koltchinskii V., Panchenko D.* Rademacher processes and bounding the risk of function learning // High Dimensional Probability, II / Ed. by D. E. Gine, J.Wellner. — Birkhauser, 1999. — Pp. 443–457.

[160] *Koltchinskii V., Panchenko D.* Empirical margin distributions and bounding the generalization error of combined classifiers // *The Annals of Statistics.* — 2002. — Vol. 30, no. 1. — Pp. 1–50.

[161] *Kuncheva L.* Combining pattern classifiers. — John Wiley & Sons, Inc., 2004.

[162] *Kutin S., Niyogi P.* The interaction of stability and weakness in AdaBoost: Tech. Rep. TR-2001-30: University of Chicago, Computer Science Department, 2001.

[163] *Kutin S., Niyogi P.* Almost-everywhere algorithmic stability and generalization error: Tech.

Rep. TR-2002-03: University of Chicago, Computer Science Department, 2002.

[164] *Langford J.* Quantitatively Tight Sample Complexity Bounds: Ph.D. thesis / Carnegie Mellon Thesis. — 2002.

[165] *Langford J.* Tutorial on practical prediction theory for classification // *Journal of Machine Learning Research.* — 2005. — Vol. 6. — Pp. 273–306.

[166] *Langford J., Blum A.* Microchoice bounds and self bounding learning algorithms // Computational Learing Theory. — 1999. — Pp. 209–214.

[167] *Langford J., McAllester D.* Computable shell decomposition bounds // Proc. 13th Annu. Conference on Comput. Learning Theory. — Morgan Kaufmann, San Francisco, 2000. — Pp. 25–34.

[168] *Langford J., Seeger M.* Bounds for averaging classifiers: Tech. Rep. CMU-CS-01-102: Carnegie Mellon University, January 2001.

[169] *Langford J., Shawe-Taylor J.* PAC-Bayes and margins // Advances in Neural Information Processing Systems 15. — MIT Press, 2002. — Pp. 439–446.

[170] *Laviolette F., Marchand M.* PAC-Bayes risk bounds for stochastic averages and majority votes of sample-compressed classifiers // *Journal of Machine Learning Research.* — 2007. — Vol. 8. — Pp. 1461–1487.

[171] *Laviolette F., Marchand M., Shah M.* PAC-Bayes approach to the set covering machine // *Journal of Machine Learning Research.* — 2006. — no. 18. — Pp. 731–738.

[172] *Lugosi G.* On concentration-of-measure inequalities. — Machine Learning Summer School, Australian National University, Canberra. — 2003.

[173] *Mann H. B., Whitney D. R.* On a test of whether one of two random variables is stochastically larger than the other // *The Annals of Mathematical Statistics.* — 1947. — Vol. 18, no. 1. — Pp. 50–60.

[174] *Marchand M., Shawe-Taylor J.* Learning with the set covering machine // Proc. 18th International Conf. on Machine Learning. — Morgan Kaufmann, San Francisco, CA, 2001. — Pp. 345–352.

[175] *Martin J. K.* An exact probability metric for decision tree splitting and stopping // *Machine Learning.* — 1997. — Vol. 28, no. 2-3. — Pp. 257–291.

[176] *Mason L., Bartlett P., Baxter J.* Direct optimization of margins improves generalization in combined classifiers // Proceedings of the 1998 conference on Advances in Neural Information Processing Systems II. — MIT Press, 1999. — Pp. 288–294.

[177] *Mason L., Bartlett P., Golea M.* Generalization error of combined classifiers: Tech. rep.: Department of Systems Engineering, Australian National University, 1997.

[178] *Mazurov V., Khachai M., Rybin A.* Committee constructions for solving problems of selection, diagnostics and prediction // *Proceedings of the Steklov Institute of mathematics.* — 2002. — Vol. 1. — Pp. 67–101.

[179] *McAllester D.* PAC-Bayesian model averaging // COLT: Proceedings of the Workshop on Computational Learning Theory. — Morgan Kaufmann Publishers, 1999.

[180] *McDiarmid C.* On the method of bounded differences // *In Surveys in Combinatorics, London Math. Soc. Lecture Notes Series.* — 1989. — Vol. 141. — Pp. 148–188.

[181] *Mertens S., Engel A.* Vapnik-Chervonenkis dimension of neural networks with binary weights // *Phys. Rev. E.* — 1997. — Vol. 55, no. 4. — Pp. 4478–4488.

[182] *Mullin M., Sukthankar R.* Complete cross-validation for nearest neighbor classifiers // Proceedings of International Conference on Machine Learning. — 2000. — Pp. 639–646.

[183] *Ng A. Y.* Preventing overfitting of cross-validation data // Proc. 14th International Conference on Machine Learning. — Morgan Kaufmann, 1997. — Pp. 245–253.

[184] *Osborne M. L.* The seniority logic: A logic for a committee machine // *IEEE Trans. on Comp.* — 1977. — Vol. C-26, no. 12. — Pp. 1302–1306.

[185] *Philips P.* Data-Dependent Analysis of Learning Algorithms: Ph.D. thesis / The Australian National University, Canberra. — 2005.

[186] *Quinlan J.* Induction of decision trees // *Machine Learning.* — 1986. — Vol. 1, no. 1. — Pp. 81–106.

[187] *Quinlan J. R.* C4.5: Programs for machine learning. — Morgan Kaufmann, San Francisco, CA, 1993.

[188] *Ratsch G., Onoda T., Muller K. R.* An improvement of adaboost to avoid overfitting // Advances in Neutral Information Processing Systems, Kitakyushu, Japan. — 1998. — Pp. 506–509.

[189] *Ratsch G., Onoda T., Muller K.-R.* Soft margins for AdaBoost // *Machine Learning.* — 2001. — Vol. 42, no. 3. — Pp. 287–320.

[190] *Rivest R. L.* Learning decision lists // *Machine Learning.* — 1987. — Vol. 2, no. 3. — Pp. 229–246.

[191] *Rogers W., Wagner T.* A finite sample distribution-free performance bound for local discrimination rules // *Annals of Statistics.* — 1978. — Vol. 6, no. 3. — Pp. 506–514.

[192] *Rosenblatt R.* Principles of neuro dynamics. — New York: Spartan books, 1959.

[193] *Rückert U., Kramer S.* Towards tight bounds for rule learning // Proc. 21th International Conference on Machine Learning, Banff, Canada. — 2004. — P. 90.

[194] *Schapire R.* The boosting approach to machine learning: An overview // MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA. — 2001.

[195] *Schapire R. E.* The strength of weak learnability // *Machine Learning.* — 1990. — Vol. 5. — Pp. 197–227.

[196] *Schapire R. E., Freund Y., Lee W. S., Bartlett P.* Boosting the margin: a new explanation for the effectiveness of voting methods // *Annals of Statistics.* — 1998. — Vol. 26, no. 5. — Pp. 1651–1686.

[197] *Schapire R. E., Singer Y.* Improved boosting using confidence-rated predictions // *Machine Learning.* — 1999. — Vol. 37, no. 3. — Pp. 297–336.

[198] *Seeger M.* PAC-Bayesian generalization error bounds for Gaussian process classification // *Journal of Machine Learning Research.* — 2002. — Vol. 3. — Pp. 233–269.

[199] *Shawe-Taylor J., Bartlett P. L.* Structural risk minimization over data-dependent hierarchies // *IEEE Trans. on Information Theory.* — 1998. — Vol. 44, no. 5. — Pp. 1926–1940.

[200] *Shawe-Taylor J., Cristianini N.* Robust bounds on generalization from the margin distribution: Tech. Rep. NC2–TR–1998–029: Royal Holloway, University of London, 1998.

[201] *Shawe-Taylor J., Cristianini N.* Margin distribution bounds on generalization // EuroCOLT '99: Proceedings of the 4th European Conference on Computational Learning Theory. — Springer-Verlag, 1999. — Pp. 263–273.

[202] *Shawe-Taylor J., Cristianini N.* On the generalization of soft margin algorithms // *IEEE Trans. on Information Theory.* — 2002. — Vol. 48, no. 10. — Pp. 2721–2735.

[203] *Sill J.* Generalization bounds for connected function classes. — citeseer.ist.psu.edu/127284.html.

[204] *Sill J.* The capacity of monotonic functions // *Discrete Applied Mathematics (special issue on VC dimension).* — 1998. — Vol. 86. — Pp. 95–107.

[205] *Sill J.* Monotonicity and connectedness in learning systems: Ph.D. thesis / California Institute of Technology. — 1998.

[206] *Sill J., Abu-Mostafa Y. S.* Monotonicity hints // Advances in Neural Information Processing Systems / Ed. by M. C. Mozer, M. I. Jordan, T. Petsche. — Vol. 9. — The MIT Press, 1997. — P. 634.

[207] *Skurichina M., Kuncheva L., Duin R.* Bagging and boosting for the nearest mean classifier: Effects of sample size on diversity and accuracy // Multiple Classifier Systems (Proc. Third International Workshop MCS, Cagliari, Italy) / Ed. by F. Roli, J. Kittler. — Vol. 2364. — Springer, Berlin, 2002. — Pp. 62–71.

[208] *Smola A., Bartlett P., Scholkopf B., Schuurmans D.* Advances in large margin classifiers. — MIT Press, Cambridge, MA. — 2000.

[209] *Talagrand. M.* Sharper bounds for gaussian and empirical processes // *Annals of Probability.* — 1994. — no. 22. — Pp. 28–76.

[210] *Talagrand M.* Concentration of measure and isoperimetric inequalities in product space // *Publ. Math. I.H.E.S.* — 1995. — no. 81. — Pp. 73–205.

[211] *Tipping M.* The relevance vector machine // Advances in Neural Information Processing Systems, San Mateo, CA. — Morgan Kaufmann, 2000.

[212] *Valiant L. G.* A theory of the learnable // *Communications of the ACM.* — 1984. — Vol. 27. — Pp. 1134–1142.

[213] *Vapnik V.* Estimation of Dependencies Based on Empirical Data. — Springer-Verlag, New York, 1982.

[214] *Vapnik V.* The nature of statistical learning theory. — Springer-Verlag, New York, 1995.

[215] *Vapnik V.* Statistical Learning Theory. — Wiley, New York, 1998.

[216] *Vapnik V., Levin E., Cun Y. L.* Measuring the VC-dimension of a learning machine // *Neural Computation.* — 1994. — Vol. 6, no. 5. — Pp. 851–876.

[217] *Vayatis N., Azencott R.* Distribution-dependent Vapnik-Chervonenkis bounds // *Lecture Notes in Computer Science.* — 1999. — Vol. 1572. — Pp. 230–240.

[218] *von Mises R.* Mathematical Theory of Probability and Statistics. — New York, Academic Press, 1964.

[219] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // *Pattern Recognition and Image Analysis.* — 2008. — Vol. 18, no. 2. — Pp. 243–259.

[220] *Vorontsov K. V.* On the influence of similarity of classifiers on the probability of overfitting // Pattern Recognition and Image Analysis: new information technologies (PRIA-9). — Vol. 2. — Nizhni Novgorod, Russian Federation, 2008. — Pp. 303–306.

[221] *Vorontsov K. V.* Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // *Pattern Recognition and Image Analysis.* — 2009. — Vol. 19, no. 3. — Pp. 412–420.

[222] *Wilcoxon F.* Individual comparisons by ranking methods // *Biometrics Bulletin.* — 1945. — Vol. 1, no. 6. — Pp. 80–83.

[223] *Williamson R., Shawe-Taylor J., Scholkopf B., Smola A.* Sample based generalization bounds: Tech. Rep. NeuroCOLT Technical Report NC-TR-99-055: 1999.

[224] *Wolpert D. H.* Stacked generalization // *Neural Networks.* — 1992. — no. 5. — Pp. 241–259.