Проблема узнавания. Развитие идей М. М. Бонгарда

Воронцов Константин Вячеславович
k.vorontsov@iai.msu.ru

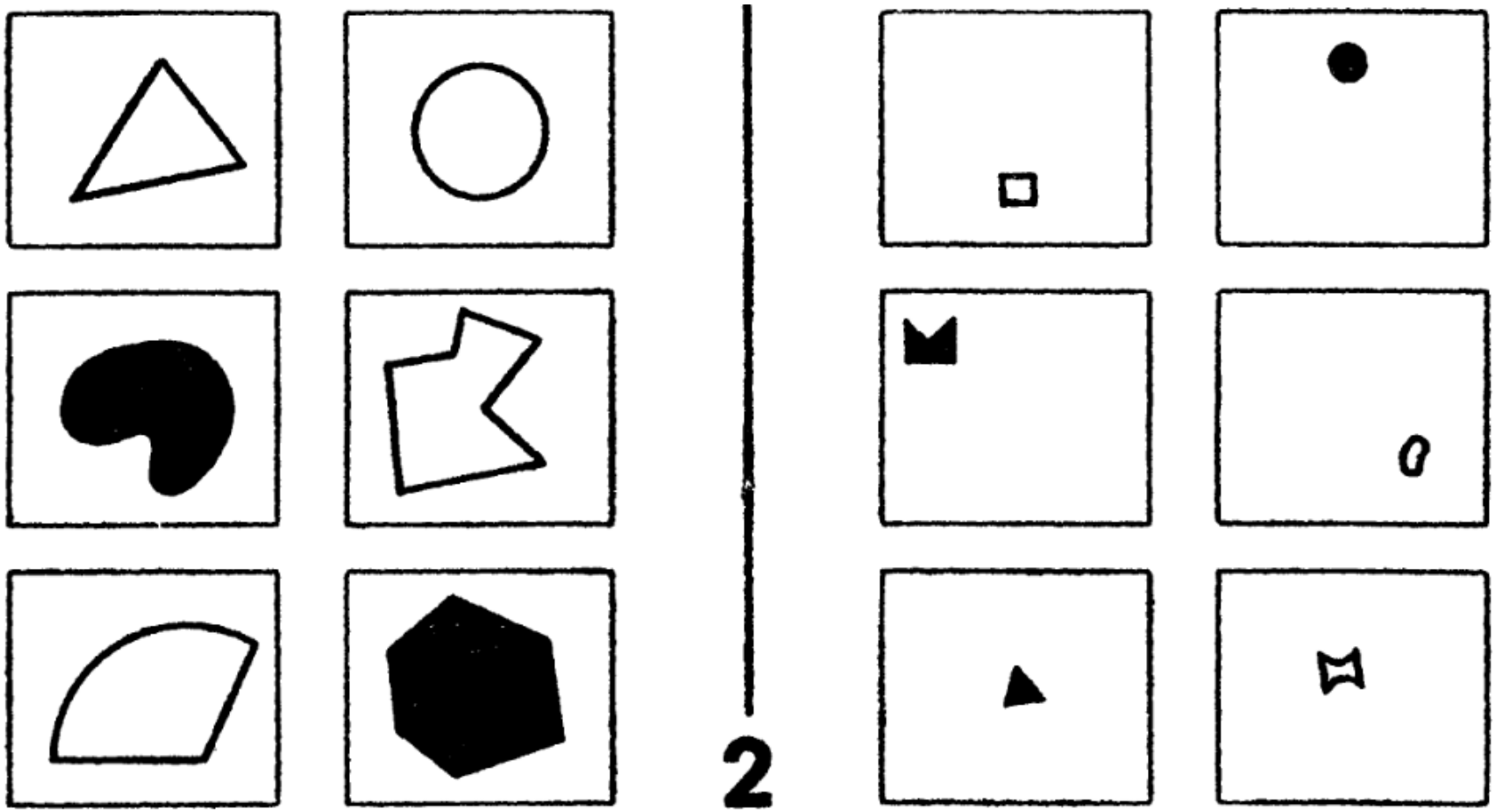
д.ф.-м.н., профессор РАН,
зав. лаб. машинного обучения и семантического анализа
Института искусственного интеллекта МГУ,
зав. каф. математических методов прогнозирования ВМК МГУ,
зав. каф. интеллектуальных систем МФТИ,
г.н.с. ФИЦ «Информатика и управление» РАН

Научная конференция к 100-летию М. М. Бонгарда
26 ноября 2024

Содержание

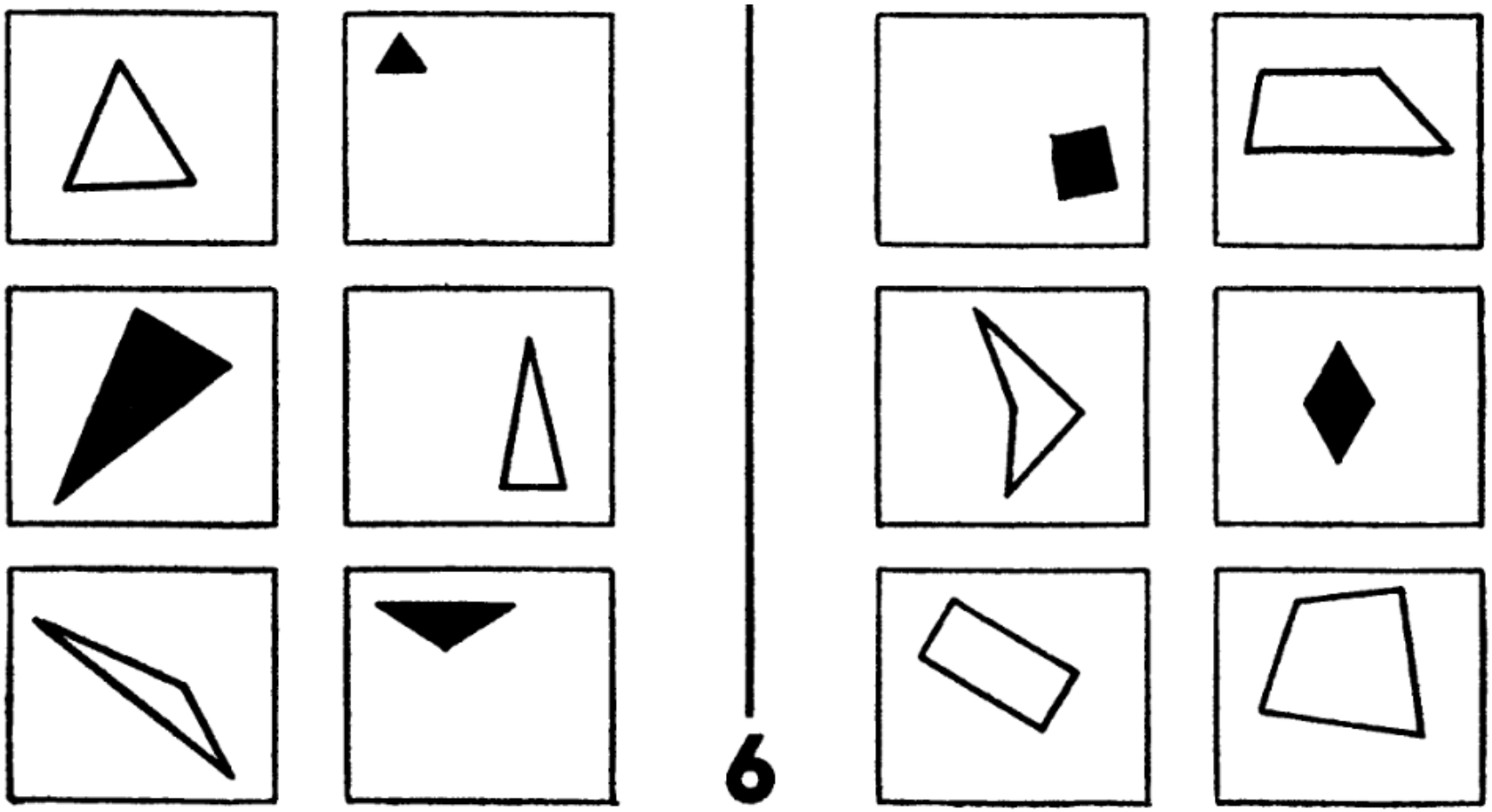
Тесты Бонгарда [Проблема узнавания, 1967]

Обучающая выборка: по 6 объектов каждого из двух классов.
Требуется найти правило классификации.



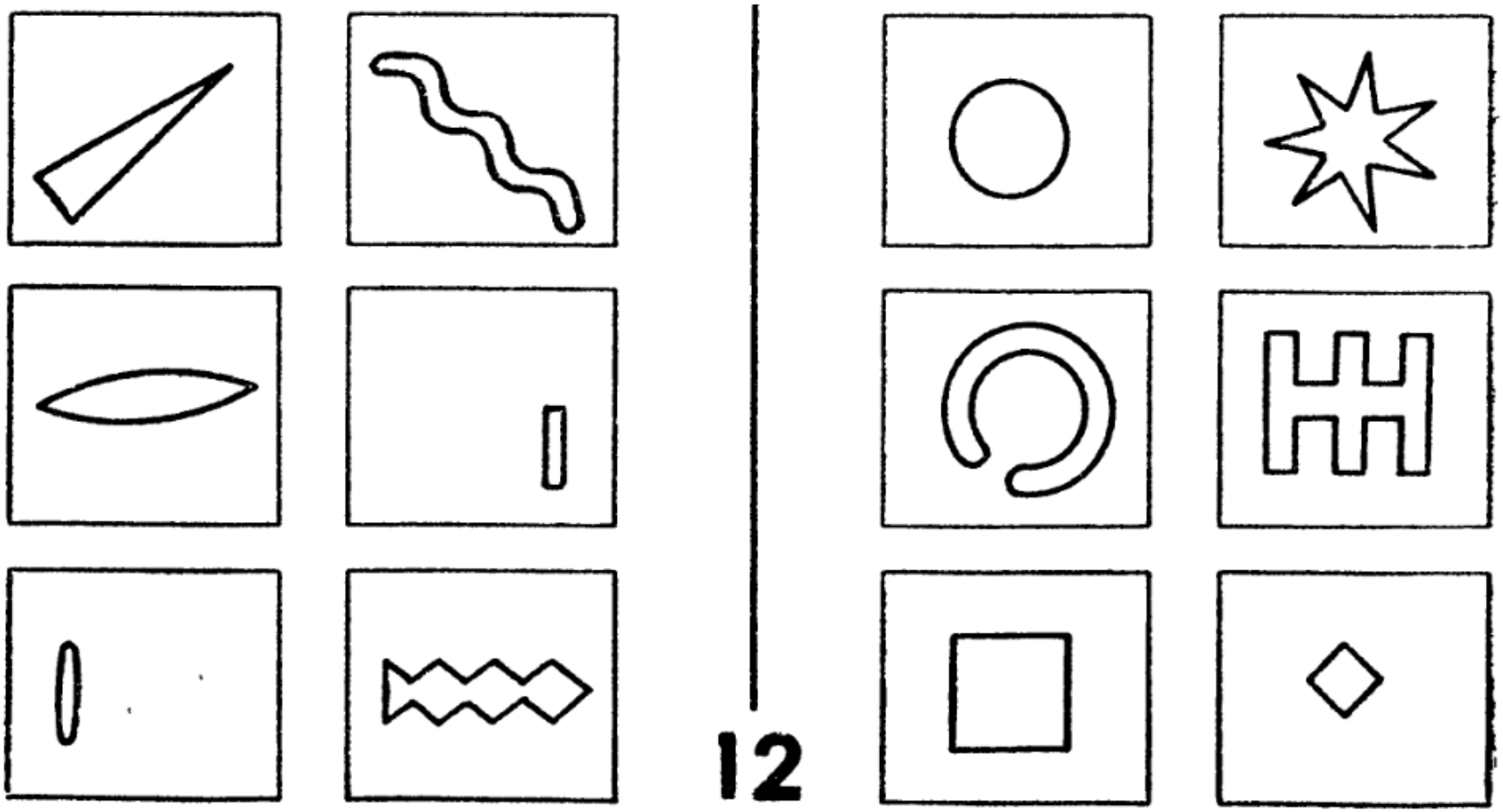
Тесты Бонгарда [Проблема узнавания, 1967]

Что даёт нам уверенность, что мы нашли верное правило?
Безошибочная классификация примеров обучающей выборки.



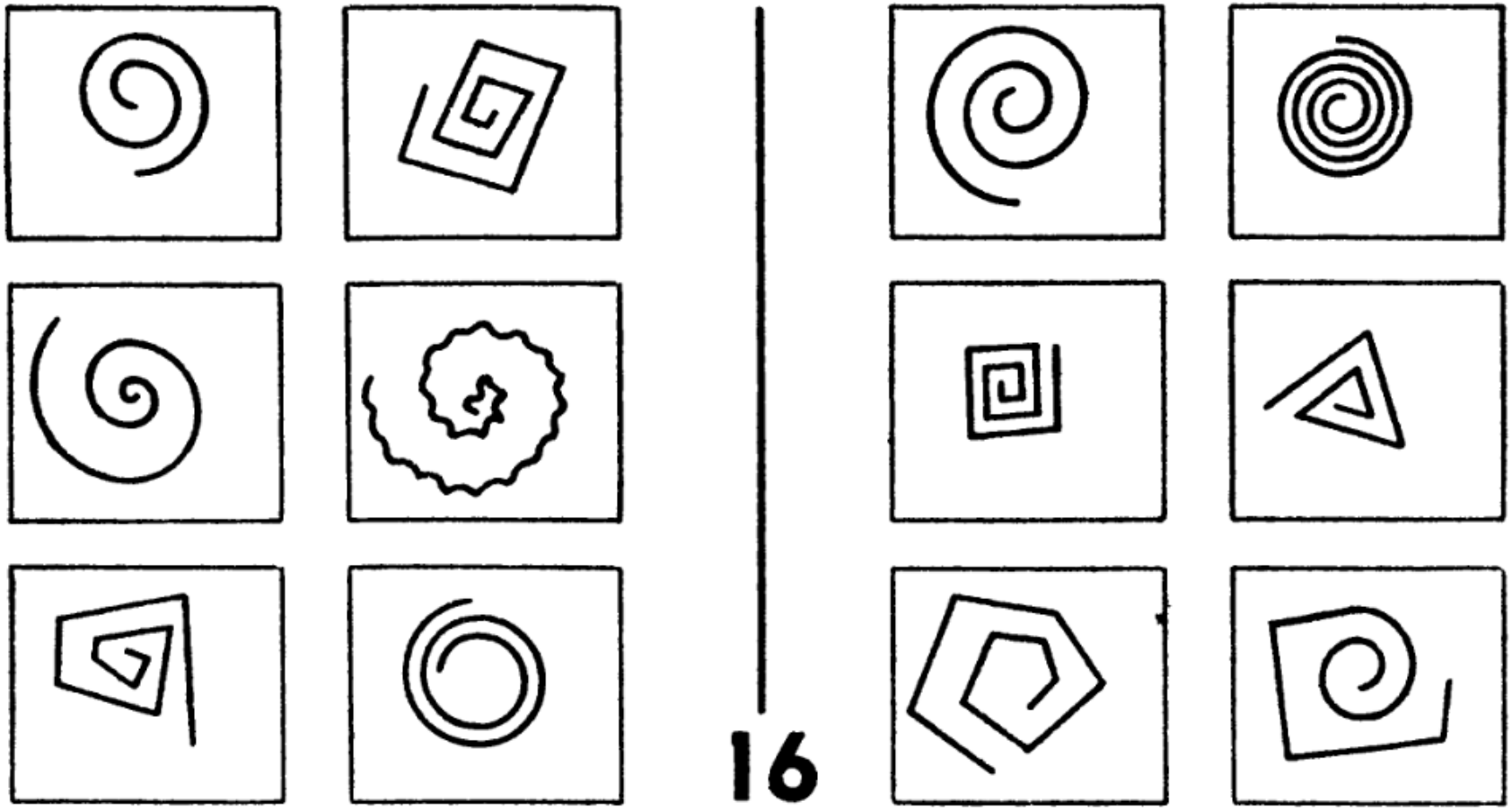
Тесты Бонгарда [Проблема узнавания, 1967]

Что ещё даёт нам уверенность, что мы нашли верное правило?
Простота, общность, изящество найденного правила.



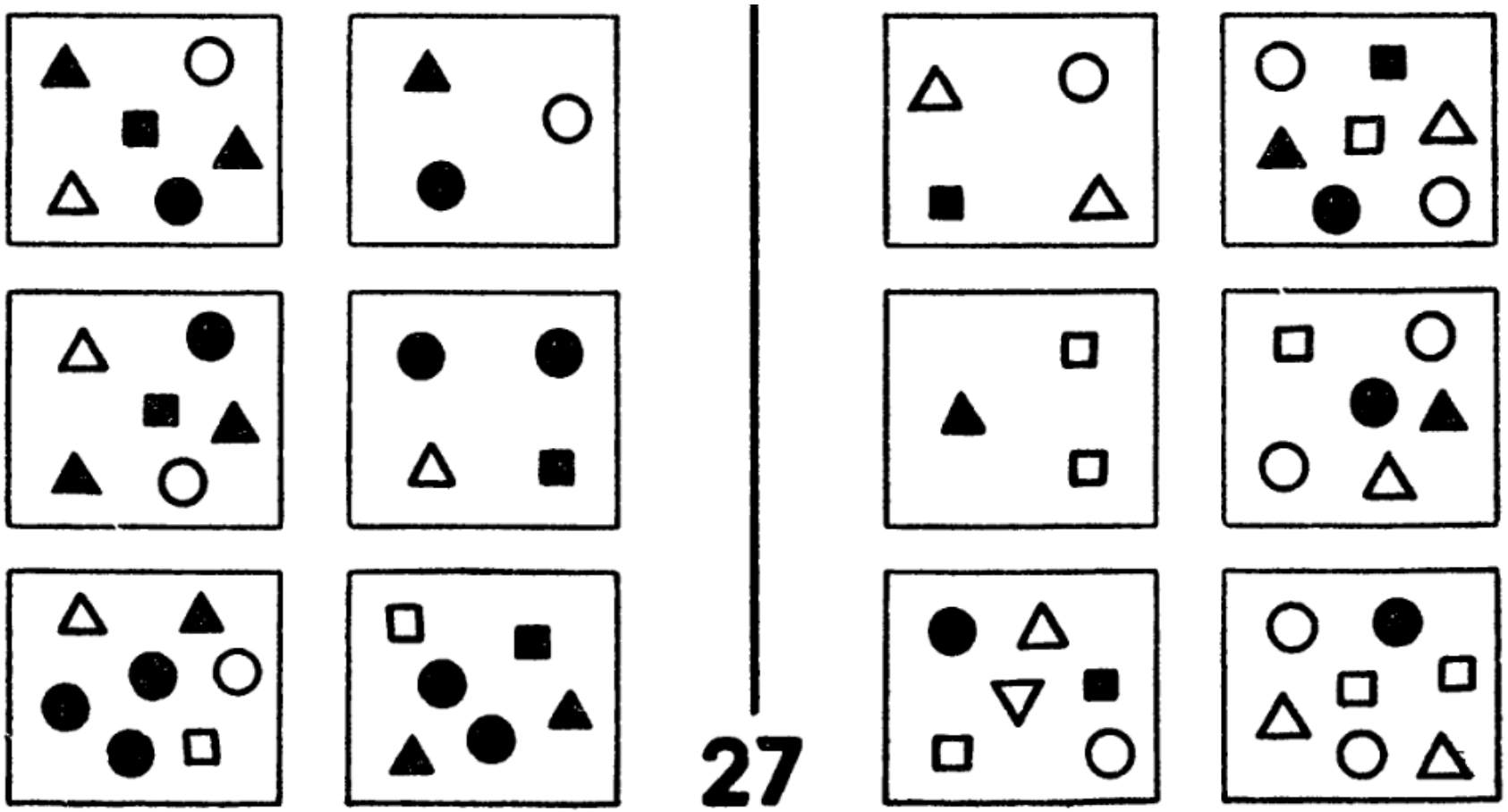
Тесты Бонгарда [Проблема узнавания, 1967]

Мы решаем эти задачи почти мгновенно. Чем мы пользуемся?
Однако для компьютера они сложны. Чего ему не хватает?



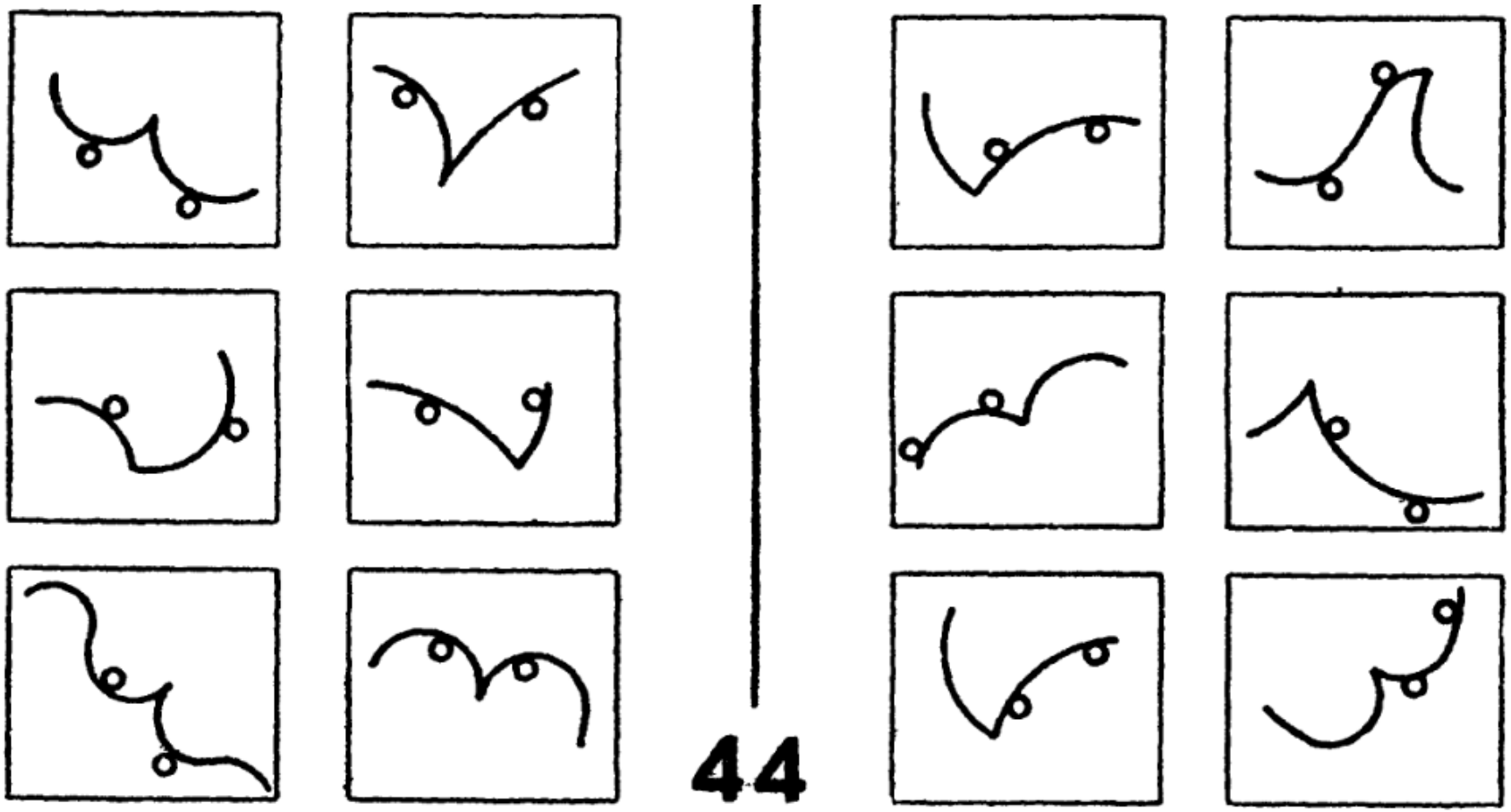
Тесты Бонгарда [Проблема узнавания, 1967]

Нужно ли закладывать знания геометрии в явном виде?
Или возможно выработать необходимые понятия на примерах?



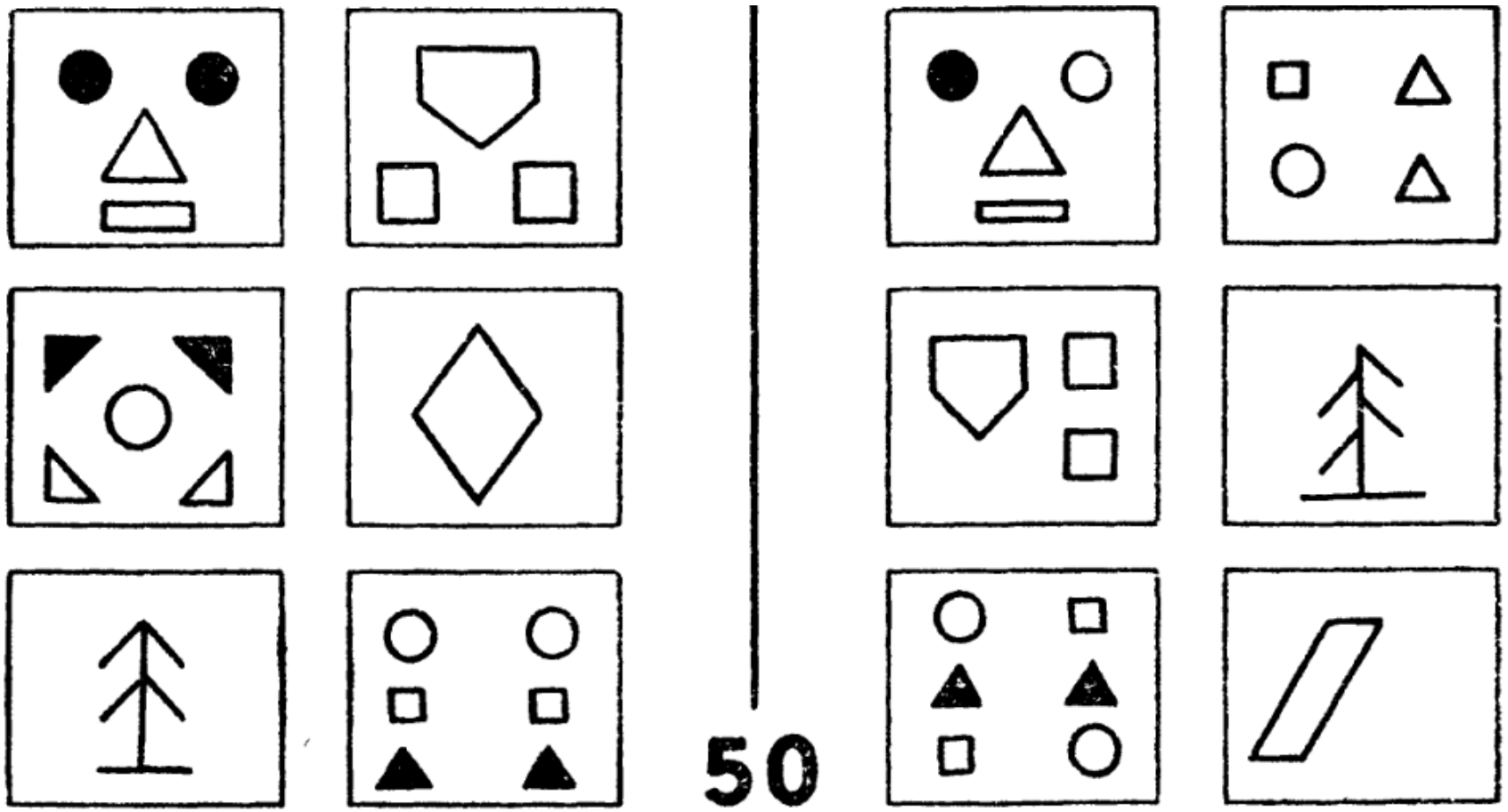
Тесты Бонгарда [Проблема узнавания, 1967]

Как вычислять полезные признаки по «сырым» данным?
Возможно ли поручить перебор признаков и моделей машине?



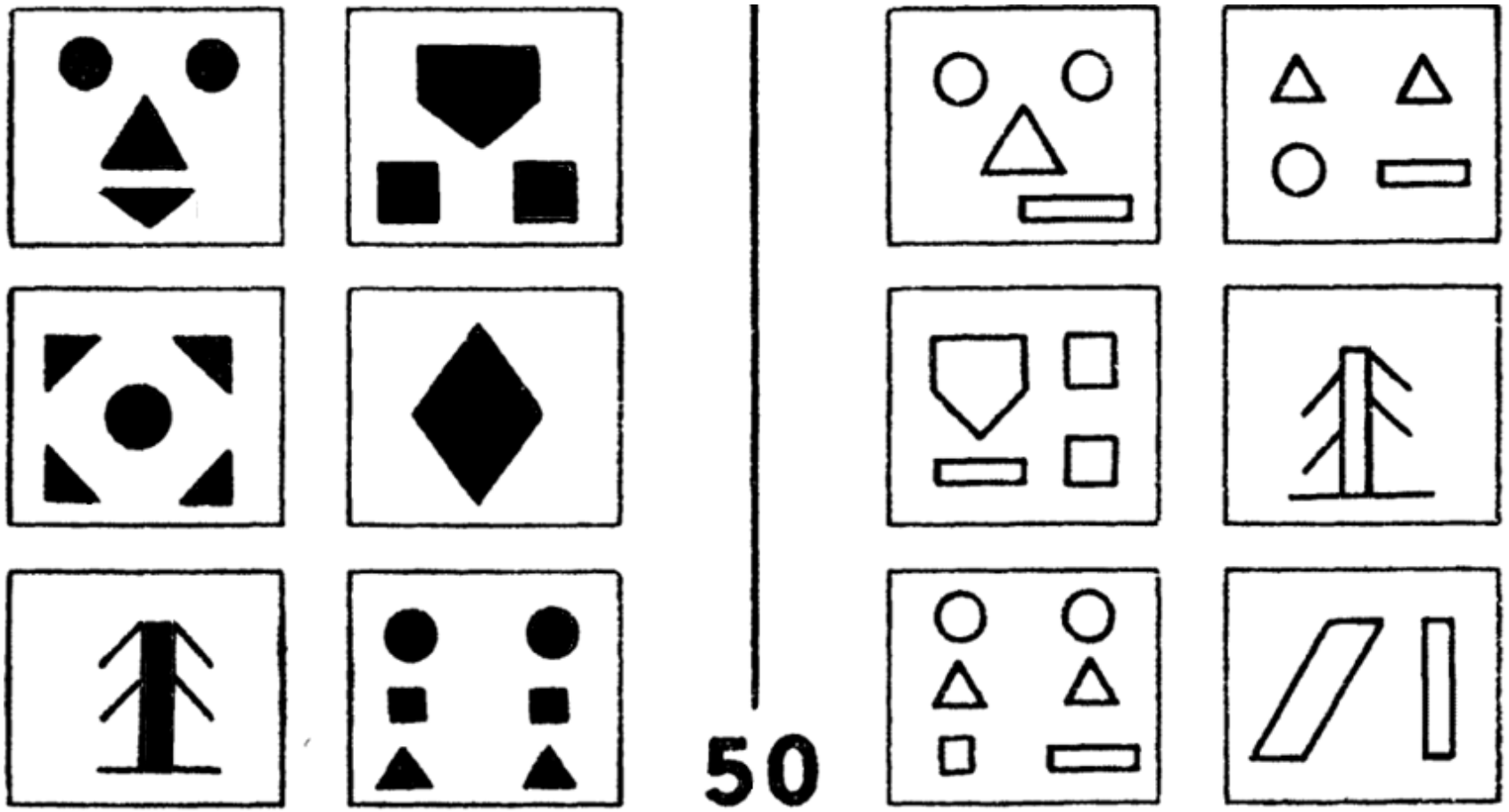
Тесты М. М. Бонгарда [Проблема узнавания, 1967]

Каков риск вывести из данных ложное правило, *предрассудок*?
Как этот риск зависит от числа примеров и сложности правил?



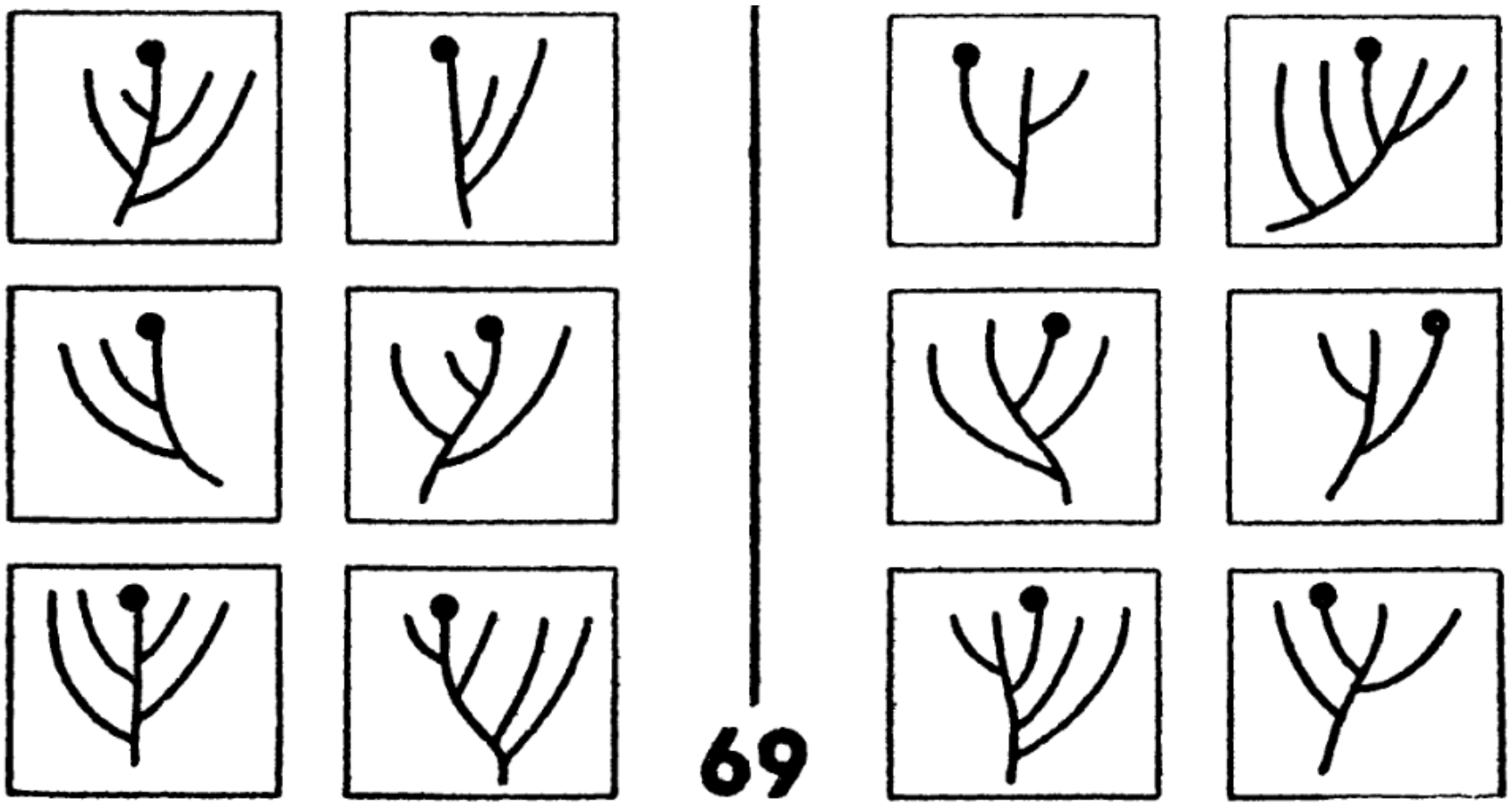
Тесты Бонгарда [Проблема узнавания, 1967]

Какого числа примеров достаточно для выработки правила?
Что делать, если к выборке подходит много разных правил?



Тесты Бонгарда [Проблема узнавания, 1967]

Эти вопросы составляют основу машинного обучения сегодня.
М.М.Бонгард поставил все эти проблемы в середине 60-х!



Принцип эмпирической индукции

«Не следует полагаться на сформулированные аксиомы и формальные базовые понятия, какими бы привлекательными и справедливыми они не казались. Законы природы нужно «расшифровывать» из фактов опыта. Следует искать правильный метод анализа и обобщения опытных данных; здесь логика Аристотеля не подходит в силу её абстрактности, оторванности от реальных процессов и явлений.»

Фрэнсис Бэкон (1561–1626)

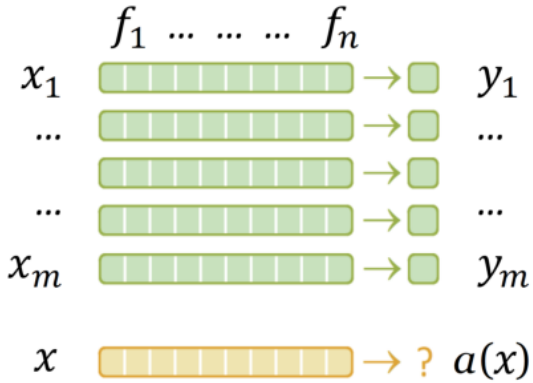
Таблица открытия: множество объектов $\{x_i : i = 1, \ldots, \ell\}$

- $f_j(x)$ — измеряемые *признаки* объектов, $j = 1, \ldots, n$
- $y_i \in \mathbb{R}$ — измеряемое значение *целевого свойства* x_i, либо $y_i \in \{0, 1\}$ — отсутствие или наличие *целевого свойства*

Фрэнсис Бэкон. Новый органон. 1620.

Восстановление зависимостей по эмпирическим данным

Дано: обучающая выборка
объектов $x_i = \big(f_1(x_i), \ldots, f_n(x_i)\big) \in X$
с ответами $y_i = y(x_i) \in Y$, $i = 1, \ldots, \ell$



Найти: параметры w модели $a(x, w)$,
приближающей зависимость $y : X \to Y$

Критерий: минимум эмпирического риска

$$\sum_{i=1}^{\ell} \mathscr{L}(a(x_i, w), y_i) \to \min_w,$$

$\mathscr{L}(a, y)$ — функция потерь модели a при правильном ответе y.

Основные типы задач машинного обучения с учителем:

- $\mathscr{L}(a, y) = (a - y)^2$ в задачах регрессии, $y_i \in \mathbb{R}$
- $\mathscr{L}(a, y) = [a \neq y]$ в задачах классификации, $y_i \in \{0, 1\}$

Научная школа М. М. Бонгарда

- 1958: Программа «Открой закон» восстанавливала зависимость полным *комбинаторным перебором* формул
- 1959: Программа «Арифметика» для сокращения перебора использовала оценки *информативности*
- 1961: Программа «КоРа» перебирала *информативные тройки* признаков

Михаил Моисеевич Бонгард (1924–1971)

«КоРа-3»: первое применение распознавания незрительных образов для распознавания границы нефть-вода в скважине.

Введены принципы *голосования*, *скользящего контроля*, понятия *информативности* и *предрассудка* (переобучения).

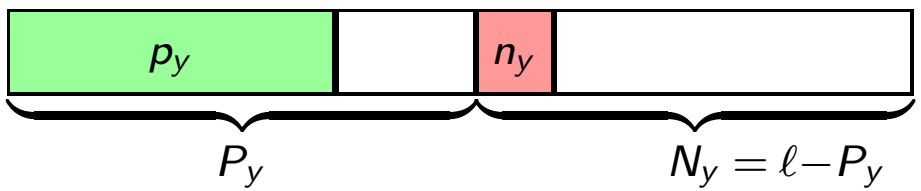
Бонгард М. М., Вайнцвайг М. Н., Губерман Ш. А. Извекова М. Л., Смирнов М. С. Использование обучающейся программы для выявления нефтеносных пластов. 1966.

Понятие информативной логической закономерности

Логическая закономерность (правило, rule) — это предикат $R\colon X \to \{0,1\}$, удовлетворяющий двум требованиям:

1. *интерпретируемость*:
   1) R записывается на естественном языке
   2) R зависит от небольшого числа признаков (1–7)
2. *информативность* относительно одного из классов $y \in Y$:
$$\begin{cases} p_y(R) = \#\{x_i\colon R(x_i)=1 \text{ и } y_i=y\} \to \max \\ n_y(R) = \#\{x_i\colon R(x_i)=1 \text{ и } y_i\neq y\} \to \min \end{cases}$$
3. *непротиворечивость* (необязательное): $n_y(R) = 0$

$$\frac{p_y(R)}{P_y} \gg \frac{n_y(R)}{N_y}$$



Если $R(x) = 1$, то говорят «R выделяет x» (R covers x).

Требование интерпретируемости

1) $R_y(x)$ записывается на естественном языке
2) $R_y(x)$ зависит от небольшого числа признаков (не более 7)

Пример (из области медицины)

Если «возраст $>$ 60» **и** «пациент ранее перенёс инфаркт»,
то операцию не делать, риск отрицательного исхода 60%

Пример (из области кредитного скоринга)

Если «в анкете указан домашний телефон»
и «зарплата $>$ \$2000» **и** «сумма кредита $<$ \$5000»
то кредит можно выдать, риск дефолта 5%

Замечание. *Риск* — частотная оценка вероятности класса, вычисляемая, как правило, по отложенной контрольной выборке

Обучение логических классификаторов

Алгоритмов *индукции правил* (rule induction) очень много!

Основные шаги их построения — надо выбрать:

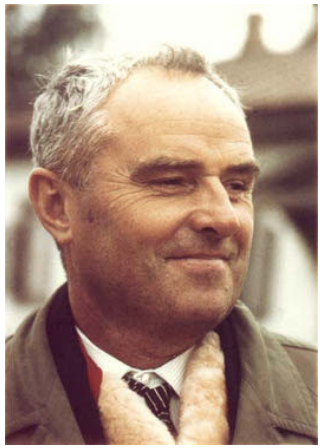
1. семейство правил, в котором будем искать закономерности (низкой размерности конъюнкции, синдромы, шары,...)
2. способ порождения правил (rule generation) (поиск в ширину, в глубину, эволюционные алгоритмы,...)
3. критерий отбора информативных правил (rule selection) (энтропийный, статистический, свёртки критериев,...)
4. модель классификации с правилами в роли признаков, например, линейный классификатор (weighted voting):

$$a(x, w) = \arg\max_{y \in Y} \sum_{j=1}^{n_y} w_{yj} R_{yj}(x)$$

Научная школа А. Г. Ива́хненко

Метод группового учёта аргументов (МГУА) на принципах *самоорганизации моделей* — переборной оптимизации структуры модели

- Качество моделей оценивается в процессе перебора по многим *внешним критериям:*
  — скользящий контроль
  — помехоустойчивость моделирования
  — баланс / согласованность прогнозов и др.
- Первая 8-слойная глубокая нейросеть (1965)
- Сотни применений, около 300 диссертаций



Алексей Григорьевич Ива́хненко (1913–2007)

Ивахненко А. Г., Лапа В. Г. Кибернетические предсказывающие устройства. 1965.
Ивахненко А. Г., Зайченко Ю. П., Димитров В. Д. Принятие решений на основе самоорганизации. 1976.
Ивахненко А. Г. Индуктивный метод самоорганизации моделей сложных систем. 1982.

Научная школа М. А. Айзермана

- *Гипотеза компактности*: схожие объекты, как правило, находятся в одном классе
- Идея *метода потенциальных функций* заимствуется из физики
- Линейная модель классификации: взвешенное голосование функций сходства $f_i(x) = K(x, x_i)$ между x и x_i:

$$a(x) = \arg\max_{y \in Y} \sum_{i:\, y_i = y} \alpha_{yi} K(x, x_i)$$

Марк Аронович Айзерман (1913–1992)

Айзерман М. А., Браверман Э. М., Розоноэр Л. И. Теоретические основы метода потенциальных функций в задаче об обучении автоматов разделению входных ситуаций на классы. 1964.
Айзерман М. А., Браверман Э. М., Розоноэр Л. И. Метод потенциальных функций в теории обучения машин. 1970.
Аркадьев А. Г., Браверман Э. М. Обучение машин распознаванию образов. 1964.

Научная школа Ю. И. Журавлёва

Объединение принципов отбора признаков, информативности, голосования и сходства

- *алгоритмы вычисления оценок* (АВО)

$$a(x) = \arg\max_{y \in Y} \sum_{i:\, y_i = y} \sum_{\omega \in \Omega} \alpha_{\omega i} B_\omega(x, x_i)$$

B_ω — бинарные функции сходства по информативным наборам признаков ω:

$$B_\omega(x, x_i) = \bigwedge_{j \in \omega} \big[\, |f_j(x) - f_j(x_i)| < \varepsilon \big]$$

Юрий Иванович Журавлёв (1935–2022)

- принципы простоты, информативности, непротиворечивости $\rightarrow$ *тупиковые тесты* / *тупиковые представительные наборы*
- принцип голосования $\rightarrow$ *алгебраический подход* к построению корректных композиций алгоритмов

Принципы информативности, непротиворечивости, тупиковости

- *информативность* предиката $R(x)$ класса $y \in Y$:
$$\begin{cases} p_y(R) = \#\{x_i\colon R(x_i)=1 \text{ и } y_i=y\} \to \max \\ n_y(R) = \#\{x_i\colon R(x_i)=1 \text{ и } y_i\neq y\} \to \min \end{cases}$$
- *информативность* функции сходства $B(x, x')$:
$$\begin{cases} p(B) = \#\{(x_i, x_j)\colon B(x_i, x_j)=1 \text{ и } y_i=y_j\} \to \max \\ n(B) = \#\{(x_i, x_j)\colon B(x_i, x_j)=1 \text{ и } y_i\neq y_j\} \to \min \end{cases}$$
- *непротиворечивость*: $n_y(B) = 0$
  — *тест* ω: $B_\omega(x_i, x_j) = 0$, $\forall i, j$: $y_i \neq y_j$
  — *представительный набор* (ω, i): $B_\omega(x_i, x_j) = 0$, $\forall j$: $y_i \neq y_j$
- *тупиковость*: никакое подмножество признаков $\omega' \subset \omega$ не является тестом (или представительным набором)

Дмитриев А. Н., Журавлев Ю. И., Кренделев Ф. П. Об одном принципе классификации и прогноза геологических объектов и явлений. 1968.
Журавлёв Ю. И., Никифоров В. В. Алгоритмы распознавания, основанные на вычислении оценок, 1971.

Ансамблирование предсказательных моделей

$X^\ell = (x_i, y_i)_{i=1}^\ell \subset X \times Y$ — обучающая выборка, $y_i = y(x_i)$

$a_t \colon X \to Y, \quad t = 1, \ldots, T$ — обучаемые *базовые алгоритмы*

Идея ансамблирования (Ю.И.Журавлёв): как из множества по отдельности плохих алгоритмов a_t построить один хороший?

Декомпозиция базовых алгоритмов $a_t(x) = C(b_t(x))$

$a_t \colon X \xrightarrow{b_t} R \xrightarrow{C} Y$, где R — удобное *пространство оценок*,

b_t — базовые *алгоритмические операторы*,

C — *решающее правило* простого вида.

Ансамбль (композиция) базовых алгоритмов $a_1, \ldots, a_T$,

$F \colon R^T \to R$ — *корректирующая* (агрегирующая) *операция*

$$a(x) = C\big(F(b_1(x), \ldots, b_T(x))\big)$$

Ю.И.Журавлёв. Об алгебраическом подходе к решению задач распознавания или классификации. Проблемы кибернетики, 1978.

Агрегирующие (корректирующие) функции

Общие требования к агрегирующей функции:

- $F(b_1, \ldots, b_T, x) \in \left[\min_t b_t, \max_t b_t\right]$ — среднее по Коши $\forall x$
- $F(b_1, \ldots, b_T, x)$ монотонно не убывает по всем b_t

Примеры агрегирующих функций:

- простое голосование (simple voting):
$$F(b_1, \ldots, b_T) = \frac{1}{T} \sum_{t=1}^{T} b_t$$
- взвешенное голосование (weighted voting):
$$F(b_1, \ldots, b_T) = \sum_{t=1}^{T} \alpha_t b_t, \quad \sum_{t=1}^{T} \alpha_t = 1, \quad \alpha_t \geqslant 0$$
- смесь алгоритмов (mixture of experts)
с функциями компетентности (gating function) $g_t : X \to \mathbb{R}$
$$F(b_1, \ldots, b_T, x) = \sum_{t=1}^{T} g_t(x) b_t(x)$$

Обучение предсказательных моделей и их ансамблей

$\mathscr{L}(b, x_i)$ — функция потерь модели $b(x_i, w)$ при ответе y_i
Минимизация эмпирического риска для базовых алгоритмов:

$$\sum_{i=1}^{\ell} \mathscr{L}\big(b_t(x_i, w), y_i\big) \;\to\; \min_{w}$$

Минимизация эмпирического риска для добавления базового алгоритма b_T в ансамбль при фиксации предыдущих:

$$\sum_{i=1}^{\ell} \mathscr{L}\left(\sum_{t=1}^{T-1} \alpha_t b_t(x_i, w_t) + \alpha_T b_T(x_i, w_T), y_i\right) \;\to\; \min_{\alpha_T, w_T}$$

Ю.И.Журавлёв. Корректные алгебры над множествами некорректных (эвристических) алгоритмов (I, II, III). Кибернетика, Киев, 1977–1978.
M.Kearns, L.G.Valiant. Cryptographic limitations on learning Boolean formulae and finite automata. 1989.
Y.Freund, R.E.Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. 1995.
К.В.Рудаков, К.В.Воронцов. О методах оптимизации и монотонной коррекции в алгебраическом подходе к проблеме распознавания. Доклады РАН, 1999.

Композиции обучаемых моделей

- Простое и взвешенное голосование
  Мазуров В. Д. Комитеты системы неравенств и задача распознавания. 1971.
  Журавлёв Ю. И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. 1977.
  Freund Y., Schapire R. E. A decision-theoretic generalization of on-line learning and an application to boosting. 1995.
  Friedman G. Greedy Function Approximation: A Gradient Boosting Machine. 1999.
- Случайный лес
  Breiman L. Random Forests. 2001.
- Восстановление смесей распределений, EM-алгоритм
  Шлезингер М. И. О самопроизвольном различении образов. 1965.
  Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM-algorithm. 1977.
- Смеси классификаторов с областями компетентности
  Растригин Л. А., Эренштейн Р. Х. Коллективные правила распознавания. 1981.
  Jacobs R. A., Jordan M. I., Nowlan S. J., Hinton G. E. Adaptive mixtures of local experts. 1991.

Градиентный бустинг и *случайный лес* — универсальные и наиболее успешные методы классификации.

MatrixNet и *CatBoost* — эффективные реализации от Яндекса.

Философия ансамблирования

Ансамблировать можно только нечто гомогенное.

1. **Декомпозиция** — разделение модели алгоритма a_t на алгоритмический оператор b_t и решающее правило C:
$$a_t = C \circ b_t$$
2. **Гомогенизация** — разнородные модели имеют общее пространство оценок R и общую структуру алгоритмического оператора b_t как отображения
$$b_t \colon X \to R$$
3. **Ансамблирование** — совместное обучение базовых алгоритмических операторов для решения общей задачи:
$$a = C \circ F(b_1, \ldots, b_T)$$

Оценка вероятности выбрать конъюнкцию-предрассудок

$x_i \in \{0,1\}^n$ — объекты описываются n бинарными признаками;
$P(x) = 2^{-n}$ — равномерное распределение на $\{0,1\}^n$;
$K_r(x)$ — конъюнкция ранга r, из r признаков или их отрицаний;
$|\mathscr{K}_n^r| = C_n^r 2^r$ — число различных таких конъюнкций.

$P\{K_r(x)=0\} = 1 - 2^{-r}$ — вероятность, что K_r ложна случайно;
$P\{K_r(X^\ell)=0\} = \prod_i P\{K_r(x_i)=0\} = \left(1-2^{-r}\right)^\ell$ — вероятность, что K случайно ложна на всех объектах из $X^\ell = (x_1, \ldots, x_\ell)$.

Верхняя оценка вероятности, что в результате поиска $K_r \in \mathscr{K}_n^r$ найденная K_r окажется случайно ложной на выборке X^ℓ:

$$\mathsf{P}\{\exists K_r\colon K_r(X^\ell)=0\} = \mathsf{P}\left(\bigcup\nolimits_{K_r}\{K_r(X^\ell)=0\}\right) \leqslant \quad \text{неравенство Буля (union bound)}$$

$$\leqslant \sum\nolimits_{K_r} \mathsf{P}\{K_r(X^\ell)=0\} = C_n^r 2^r \left(1-2^{-r}\right)^\ell.$$

Закревский Аркадий Дмитриевич. Логика распознавания. Минск, 1988.

Оценка вероятности выбрать конъюнкцию-предрассудок

Верхняя оценка вероятности $\mathsf{P}(r)$ найти случайно ложную конъюнкцию заданного ранга r, при $\ell = 200$, $n = 100$:

r	P	r	P
1	$1.24 \cdot 10^{-58}$	4	$1.56 \cdot 10^{2}$
2	$2.04 \cdot 10^{-21}$	5	$4.21 \cdot 10^{6}$
3	$3.26 \cdot 10^{-6}$	6	$3.27 \cdot 10^{9}$

Зависимость граничного значения r, после которого резко увеличивается верхняя оценка вероятности $\mathsf{P}(r)$:

n	$\ell =$ 20	50	100	200	500	1000
10	1	2	3	4	5	6
30	1	2	2	3	4	5
100	1	1	2	3	4	5

Закревский Аркадий Дмитриевич. Логика распознавания. Минск, 1988.

Научная школа В. Н. Вапника и А. Я. Червоненкиса

Семейство классификаторов A *обучаемо*:

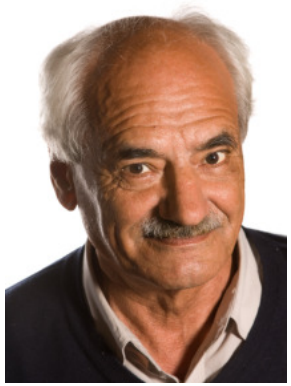
$$\mathsf{P}\left\{\sup_{a\in A}\left|P(a)-\nu(a,X^\ell)\right|>\varepsilon\right\}\leqslant\eta,$$

$P(a)$ — вероятность ошибки классификатора,
$\nu(a,X^\ell)$ — эмпирический риск — частота ошибок классификатора a на выборке.

Основные результаты VC-теории:

- Обосновано ограничение сложности A
- Понятие *ёмкости* семейства, VCdim
- Метод структурной минимизации риска
- Метод опорных векторов, SVM

Владимир Наумович Вапник



Алексей Яковлевич Червоненкис (1938–2014)

Вапник В. Н., Червоненкис А. Я.
Теория распознавания образов. М.: Наука, 1974.

Проблема оценивания обобщающей способности

Дано:

$X^L = \{x_1, \ldots, x_L\}$ — *генеральное множество* объектов

$X^L = X^\ell \sqcup X^k$ — разбиение на *обучающую* и *контрольную* части

$A = \{a \colon X \times W \to Y\}$ — модель, *семейство алгоритмов*

$\mu \colon (X \times Y)^\ell \to A$ — *метод обучения*

$\mathscr{L}(a, x)$ — *функция потерь* алгоритма a на объекте x

$Q(a, U) = \frac{1}{|U|} \sum\limits_{x \in U} \mathscr{L}(a, x)$ — *средняя потеря* на выборке U

Найти:

способ оценивать и минимизировать $Q(\mu(X^\ell), X^k)$, не зная X^k,
для широкого класса задач $(Y,\ A,\ W)$ и методов (μ);

при упрощающих предположениях:

— функция потерь бинарная;

— все разбиения $X^\ell \sqcup X^k$ случайны и равновероятны.

Бинарная функция потерь. Матрица ошибок

$X^L = \{x_1, \ldots, x_L\}$ — конечное *генеральное множество* объектов

$A = \{a_1, \ldots, a_D\}$ — конечное множество (семейство) *алгоритмов*

$\mathscr{L}(a, x) \equiv I(a, x) = [\,a$ ошибается на $x\,]$ — *индикатор ошибки*

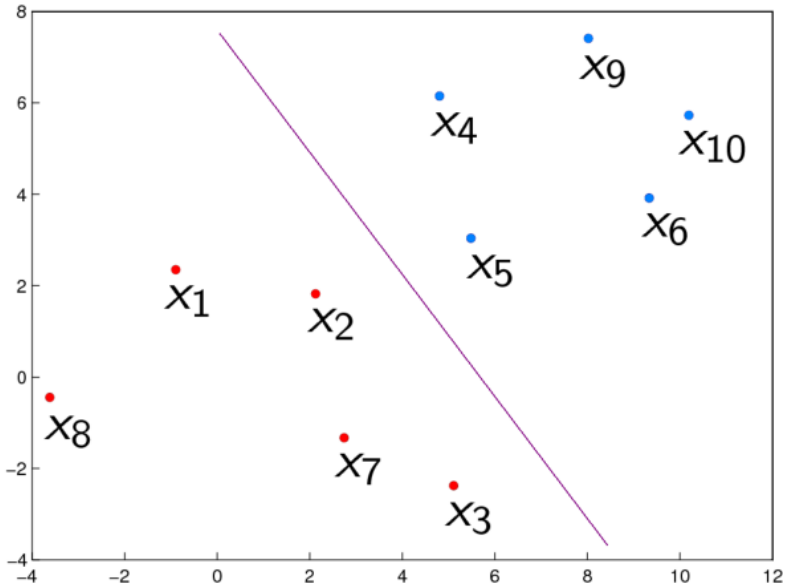
$L \times D$-*матрица ошибок* с попарно различными столбцами:

	a_1	a_2	a_3	a_4	a_5	a_6	$\cdots$	a_D	
x_1	1	1	0	**0**	0	1	$\cdots$	1	X^ℓ — наблюдаемая
$\ldots$	0	0	0	**0**	1	1	$\cdots$	1	(обучающая) выборка
x_ℓ	0	0	1	**0**	0	0	$\cdots$	0	длины ℓ
$x_{\ell+1}$	0	0	0	**1**	1	1	$\cdots$	0	X^k — скрытая
$\ldots$	0	0	0	**1**	0	0	$\cdots$	1	(контрольная) выборка
x_L	0	1	1	**1**	1	1	$\cdots$	0	длины $k = L - \ell$

$n(a, X) = \sum\limits_{x \in X} I(a, x)$ — *число ошибок* $a \in A$ на выборке $X \subset X^L$

$\nu(a, X) = n(a, X)/|X|$ — *частота ошибок* a на выборке X

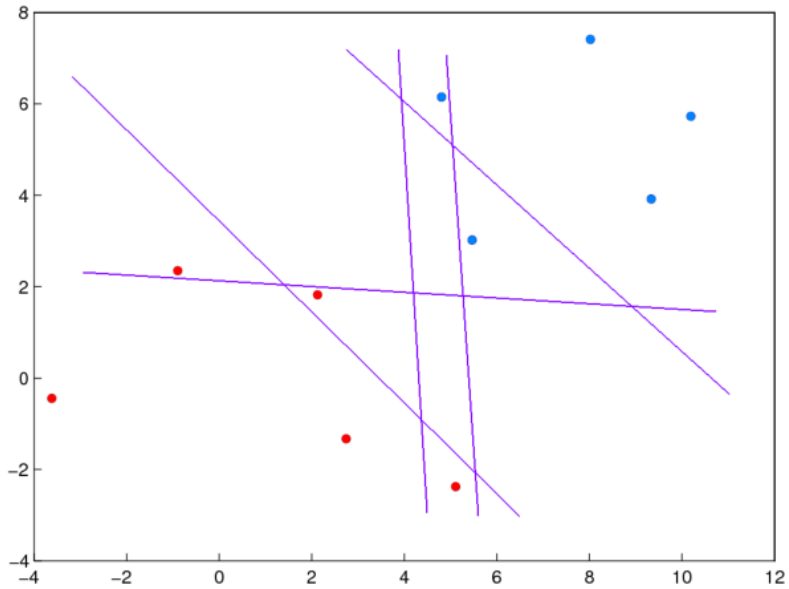
Пример. Матрица ошибок линейных классификаторов



1 вектор с 0 ошибками

x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0

Пример. Матрица ошибок линейных классификаторов

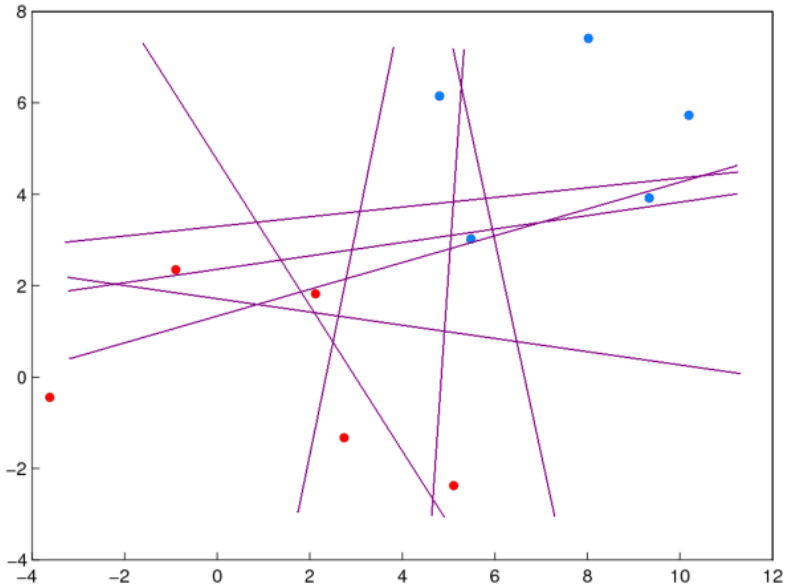


1 вектор с 0 ошибками

5 векторов с 1 ошибкой

x_1	0	1	0	0	0	0
x_2	0	0	1	0	0	0
x_3	0	0	0	1	0	0
x_4	0	0	0	0	1	0
x_5	0	0	0	0	0	1
x_6	0	0	0	0	0	0
x_7	0	0	0	0	0	0
x_8	0	0	0	0	0	0
x_9	0	0	0	0	0	0
x_{10}	0	0	0	0	0	0

Пример. Матрица ошибок линейных классификаторов



1 вектор с 0 ошибками
5 векторов с 1 ошибкой
8 векторов с 2 ошибками
и т. д...

x_1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	...
x_2	0	0	1	0	0	0	1	1	0	0	0	0	0	0	...
x_3	0	0	0	1	0	0	0	1	1	0	0	0	0	1	...
x_4	0	0	0	0	1	0	0	0	1	1	0	0	0	0	...
x_5	0	0	0	0	0	1	0	0	0	1	1	1	0	0	...
x_6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	...
x_7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	...
x_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_{10}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...

Задача оценивания вероятности переобучения

Переобученность — разность частот ошибок на X^k и на X^ℓ:

$$\delta(\mu, X^\ell, X^k) = \nu\big(\mu(X^\ell), X^k\big) - \nu\big(\mu(X^\ell), X^\ell\big).$$

Переобучение — это событие $\delta(\mu, X^\ell, X^k) \geqslant \varepsilon$.

Основное вероятностное предположение:

$\mathsf{P} \equiv \mathsf{E} \equiv \frac{1}{C_L^\ell} \sum_{X^\ell \subset X^L}$ — все разбиения $X^\ell \sqcup X^k = X^L$ равновероятны

Интерпретация 1: это CCV, полный скользящий контроль.
Интерпретация 2: это гипотеза независимости выборки X^L.

Основная задача — оценить *вероятность переобучения*:

$$R_\varepsilon(\mu, X^L) = \mathsf{P}\big[\delta(\mu, X^\ell, X^k) \geqslant \varepsilon\big].$$

$\widehat{\mathsf{P}} \equiv \widehat{\mathsf{E}} \equiv \frac{1}{|N|} \sum_{X^\ell \in N}$ — эмпирическая оценка методом Монте-Карло по случайному подмножеству разбиений N

Простейший, но важный частный случай

Пусть $A = \{a\}$ — одноэлементное множество, $m = n(a, X^L)$.

Тогда вероятность переобучения есть вероятность большого отклонения частот ошибок алгоритма a в двух подвыборках:

$$R_\varepsilon(a, X^L) = \mathsf{P}\big[\delta(a, X^\ell, X^k) \geqslant \varepsilon\big] = \mathsf{P}\big[\nu\big(a, X^k\big) - \nu\big(a, X^\ell\big) \geqslant \varepsilon\big].$$

Теорема

Для любого X^L, любого $\varepsilon \in [0, 1]$ вероятность переобучения описывается функцией гипергеометрического распределения:

$$R_\varepsilon(a, X^L) = \mathcal{H}_L^{\ell,\, m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right),$$

где $\mathcal{H}_L^{\ell,\, m}(z) = \sum\limits_{s=0}^{\lfloor z \rfloor} \dfrac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$.

Доказательство

1. Обозначим $s = n(a, X^\ell)$.

2. «Школьная» задача по теории вероятностей:
в урне L шаров, m из них чёрные; извлекаем ℓ шаров наугад.
Какова вероятность того, что s из них чёрные?

$$\mathsf{P}\big[n(a, X^\ell) = s\big] = C_m^s C_{L-m}^{\ell-s} / C_L^\ell.$$

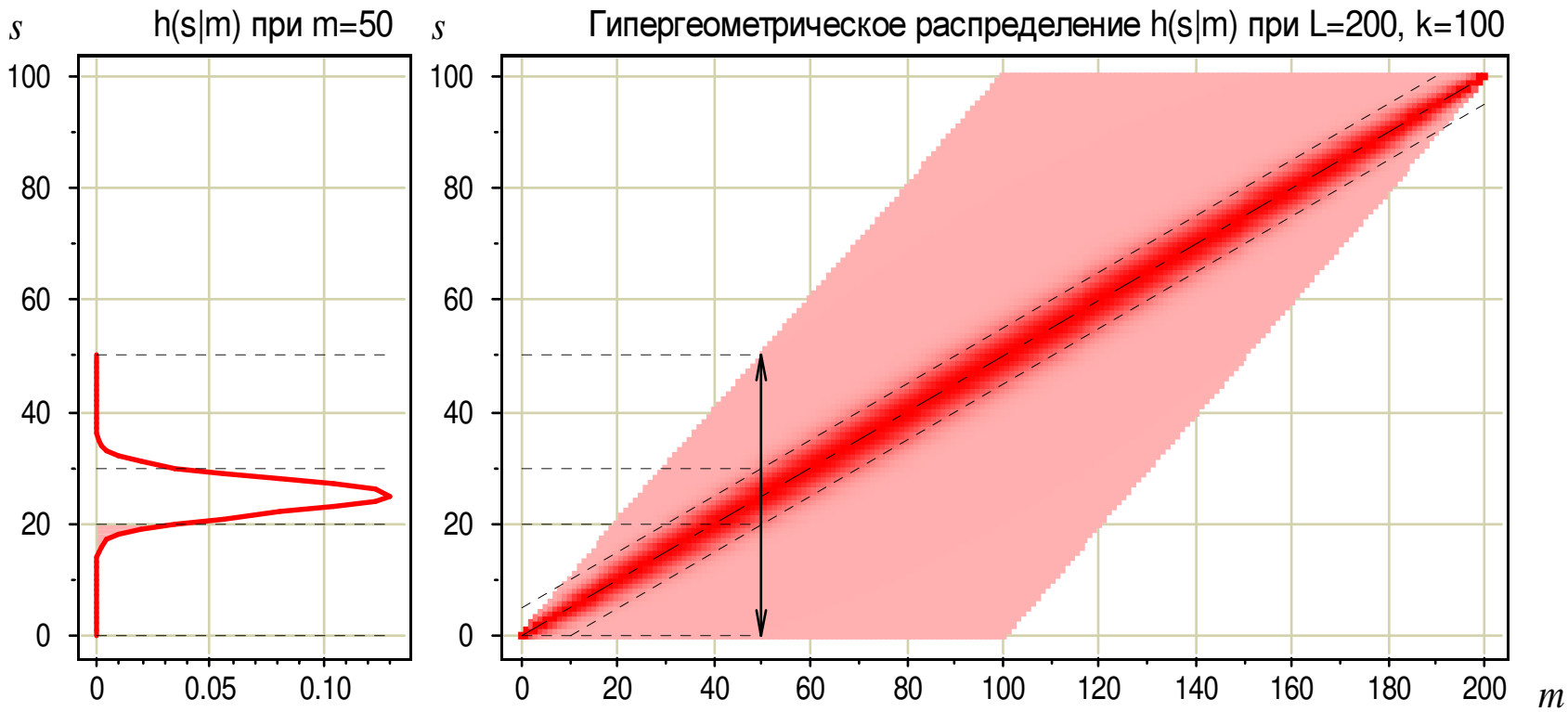
3. Распишем R_ε, подставив $\nu(a, X^k) = \frac{m-s}{k}$, $\nu(a, X^\ell) = \frac{s}{\ell}$:

$$R_\varepsilon(a, X^L) = \mathsf{P}\big[\nu\big(a, X^k\big) - \nu\big(a, X^\ell\big) \geqslant \varepsilon\big] =$$

$$= \sum_{s=0}^{\ell} \Big[\underbrace{\tfrac{m-s}{k} - \tfrac{s}{\ell} \geqslant \varepsilon}_{s \leqslant \frac{\ell}{L}(m - \varepsilon k)}\Big] \underbrace{\mathsf{P}\big[n(a, X^\ell) = s\big]}_{C_m^s C_{L-m}^{\ell-s} / C_L^\ell} =$$

$$= \mathcal{H}_L^{\ell, m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right). \qquad \blacksquare$$

Гипергеометрическое распределение $h(s|m) = C_m^s C_{L-m}^{\ell-s} / C_L^\ell$



Концентрация вероятностной меры, закон больших чисел:
предсказание числа $m = n(a, X^L)$ по числу $s = n(a, X^\ell)$
возможно благодаря узости гипергеометрического пика,
причём при $\ell, k \to \infty$ он сужается, и $\nu(a, X^\ell) \to \nu(a, X^k)$

Принцип равномерной сходимости частот

Рассмотрим случай, когда A произвольное, конечное.

1. Вероятность переобучения оценим сверху вероятностью большого *равномерного отклонения* частот: для любых X^L, μ

$$R_\varepsilon(\mu, X^L) = \mathsf{P}\big[\delta(\mu, X^\ell, X^k) \geqslant \varepsilon\big] \leqslant$$
$$\leqslant \mathsf{P}\Big[\max_{a\in A} \delta(a, X^\ell, X^k) \geqslant \varepsilon\Big] = \widetilde{R}_\varepsilon(A, X^L).$$

2. Оценим вероятность объединения событий суммой их вероятностей (неравенство Буля, union bound):

$$\widetilde{R}_\varepsilon(A, X^L) = \mathsf{P} \max_{a\in A} \big[\delta(a, X^\ell, X^k) \geqslant \varepsilon\big] \leqslant$$
$$\leqslant \mathsf{P} \sum_{a\in A} \big[\delta(a, X^\ell, X^k) \geqslant \varepsilon\big] = \sum_{a\in A} \underbrace{\mathsf{P}\big[\delta(a, X^\ell, X^k) \geqslant \varepsilon\big]}_{R_\varepsilon(a, X^L)}.$$

Оценка Вапника–Червоненкиса (VC bound)

Таким образом, доказана

Лемма. Для любых X^L, μ, конечного A и $\varepsilon \in [0,1]$

$$\widetilde{R}_\varepsilon(A, X^L) \leqslant \sum_{a \in A} \mathcal{H}_L^{\ell,\, m}\left(\tfrac{\ell}{L}\big(m - \varepsilon k\big)\right), \quad m = n(a, X^L).$$

Теорема (Вапник и Червоненкис, 1968)

Для любых X^L, μ, конечного A и $\varepsilon \in [0,1]$

$$\widetilde{R}_\varepsilon(A, X^L) \leqslant |A| \cdot \max_m \mathcal{H}_L^{\ell,\, m}\left(\tfrac{\ell}{L}\big(m - \varepsilon k\big)\right) \leqslant$$
$$\leqslant |A| \cdot \tfrac{3}{2} \exp\big(-\varepsilon^2 \ell\big), \quad \text{при } \ell = k.$$

В.Н.Вапник, А.Я.Червоненкис. О равномерной сходимости частот появления событий к их вероятностям. 1968.

Обобщение на случай бесконечных семейств A

Функция роста $\Delta^A(L)$ семейства A — это максимальное по X^L число различных векторов ошибок $\vec{a} = \big(I(a, x_1), \ldots, I(a, x_L)\big)$.
В оценке можно заменить $|A|$ на функцию роста $\Delta^A(L)$.

Ёмкость (размерность Вапника-Червоненкиса) семейства A — это максимальная длина выборки h, для которой $\Delta^A(h) = 2^h$.

Теорема

Если такое h существует, то $\Delta^A(L) \leqslant C_L^0 + \cdots + C_L^h \leqslant \frac{3}{2}\frac{L^h}{h!}$.

Теорема

Ёмкость семейства линейных классификаторов на два класса

$$a(x) = \operatorname{sign}(w_1 x^1 + \cdots + w_n x^n), \quad x = (x^1, \ldots, x^n) \in X.$$

равна размерности пространства параметров, $\operatorname{VCdim}(A) = n$.

Обращение оценки Вапника-Червоненкиса (при $\ell = k$)

1. Оценка: $\mathsf{P}\Big[\max\limits_{a\in A}\big(\nu(a,X^k)-\nu(a,X^\ell)\big) \geqslant \varepsilon\Big] \leqslant \Delta\, \frac{3}{2} \exp\big(-\ell\varepsilon^2\big) = \eta$

Тогда для любого $a \in A$ с вероятностью не менее $(1-\eta)$

$$\nu(a,X^k) \leqslant \underbrace{\nu(a,X^\ell)}_{\text{эмпирический риск}} + \underbrace{\sqrt{\frac{1}{\ell}\ln\Delta + \frac{1}{\ell}\ln\frac{3}{2\eta}}}_{\text{штраф за сложность}}.$$

2. Оценка: $\mathsf{P}\Big[\max\limits_{a\in A}\big(\nu(a,X^k)-\nu(a,X^\ell)\big) \geqslant \varepsilon\Big] \leqslant \frac{3}{2}\frac{L^h}{h!}\frac{3}{2} \exp\big(-\ell\varepsilon^2\big) = \eta$

Тогда для любого $a \in A$ с вероятностью не менее $(1-\eta)$

$$\nu(a,X^k) \leqslant \underbrace{\nu(a,X^\ell)}_{\text{эмпирический риск}} + \underbrace{\sqrt{\frac{h}{\ell}\ln\frac{2e\ell}{h} + \frac{1}{\ell}\ln\frac{9}{4\eta}}}_{\text{штраф за сложность}}.$$

Метод структурной минимизации риска (СМР)

Дано: система вложенных подсемейств возрастающей ёмкости

$$A_0 \subset A_1 \subset \cdots \subset A_h \subset \cdots$$

Найти: оптимальную ёмкость h^*, такую, что

$$\nu(a, X^k) \leqslant \underbrace{\min_{a \in A_h} \nu(a, X^\ell)}_{\text{минимизация эмпирического риска}} + \underbrace{\sqrt{\frac{h}{\ell} \ln \frac{2e\ell}{h} + \frac{1}{\ell} \ln \frac{9}{4\eta}}}_{\text{штраф за сложность}} \to \min_h$$

Недостатки СМР:

— верхняя оценка R_ε очень сильно завышена

— следовательно, h^* может оказаться заниженной

— на практике предпочитают эмпирические оценки CV

В.Н.Вапник, А.Я.Червоненкис. Теория распознавания образов. М.: Наука, 1974.
В.Н.Вапник, А.Я.Червоненкис. Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979.

Причины завышенности оценок Вапника-Червоненкиса

- **Оценка равномерного отклонения** (uniform bound) сильно завышена, когда бóльшая часть алгоритмов имеет исчезающе малую вероятность быть результатом обучения

  На практике распределение

$$q(a) = \mathsf{P}\big[\mu(X^\ell) = a\big], \quad a \in A$$

  как правило, существенно неравномерно!

  Будем называть это эффектом расслоения семейства A.
- **Неравенство Буля** (union bound) сильно завышено, когда среди бинарных векторов ошибок есть много похожих

  Будем называть это эффектом сходства алгоритмов

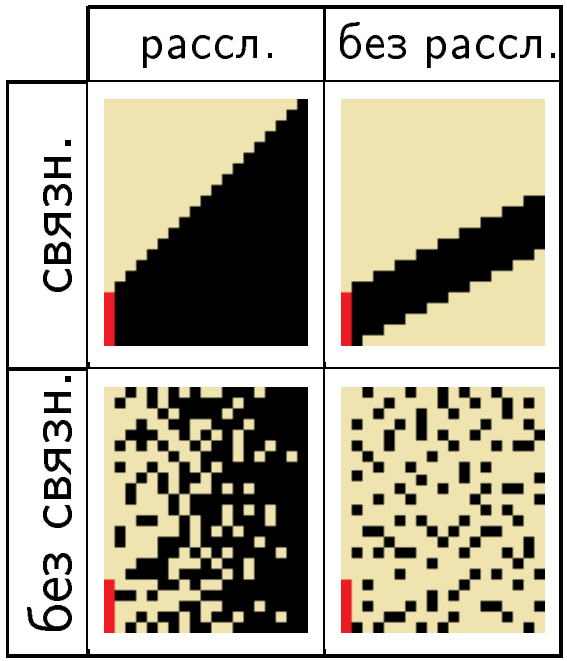
K.V.Vorontsov. Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting. 2008.

Как зависит переобучение от содержимого матрицы ошибок?

Верхние оценки VC-теории зависят только от размера матрицы ошибок $L \times |A|$, и потому сильно завышены.

Эксперимент: сравним R_ε и CCV для четырёх матриц ошибок:

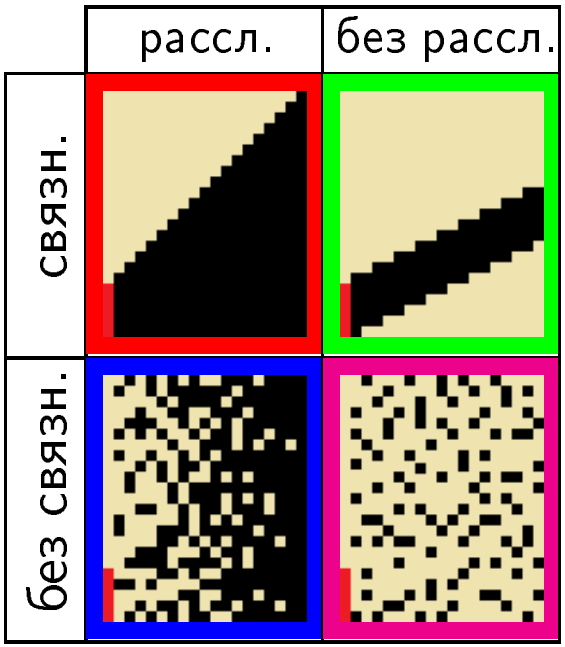
- лучший алгоритм одинаковый
- есть/нет *расслоение* — когда каждый следующий алгоритм допускает на одну ошибку больше, чем предыдущий
- есть/нет *связность* — когда каждый следующий алгоритм лишь на одном объекте отличается от предыдущего

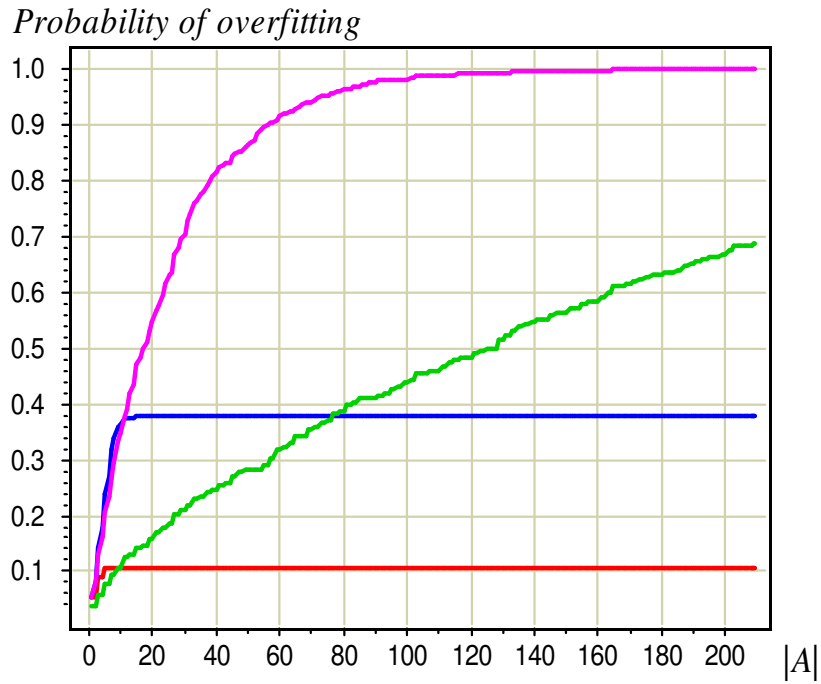


Vorontsov K. V. Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting. PRIA, 2009.

Зависимость вероятности переобучения от числа алгоритмов

$\ell = k = 100,\ m^* = 10,\ \varepsilon = 0.05,\ |N| = 10^3$ разбиений Монте-Карло





- *связность* замедляет темп роста кривой $R_\varepsilon(|A|)$
- *расслоение* понижает уровень горизонтальной асимптоты
- огромные семейства с P&C могут почти не переобучаться
- VC-оценка линейно мажорирует худшую из этих кривых

Семейство из двух алгоритмов $A = \{a_1, a_2\}$

Пусть для алгоритмов a_1, a_2 известны m_0, m_1, m_2, m_3:

$$a_1 = (1, \ldots, 1,\ 1, \ldots, 1,\ 0, \ldots, 0,\ 0, \ldots, 0);$$
$$a_2 = (\underbrace{1, \ldots, 1}_{m_0},\ \underbrace{0, \ldots, 0}_{m_1},\ \underbrace{1, \ldots, 1}_{m_2},\ \underbrace{0, \ldots, 0}_{m_3}).$$

Сходство векторов ошибок измеряется расстоянием Хэмминга:

$$r(a_1, a_2) = \sum_{i=1}^{L} \left| I(a_1, x_i) - I(a_2, x_i) \right| = m_1 + m_2$$

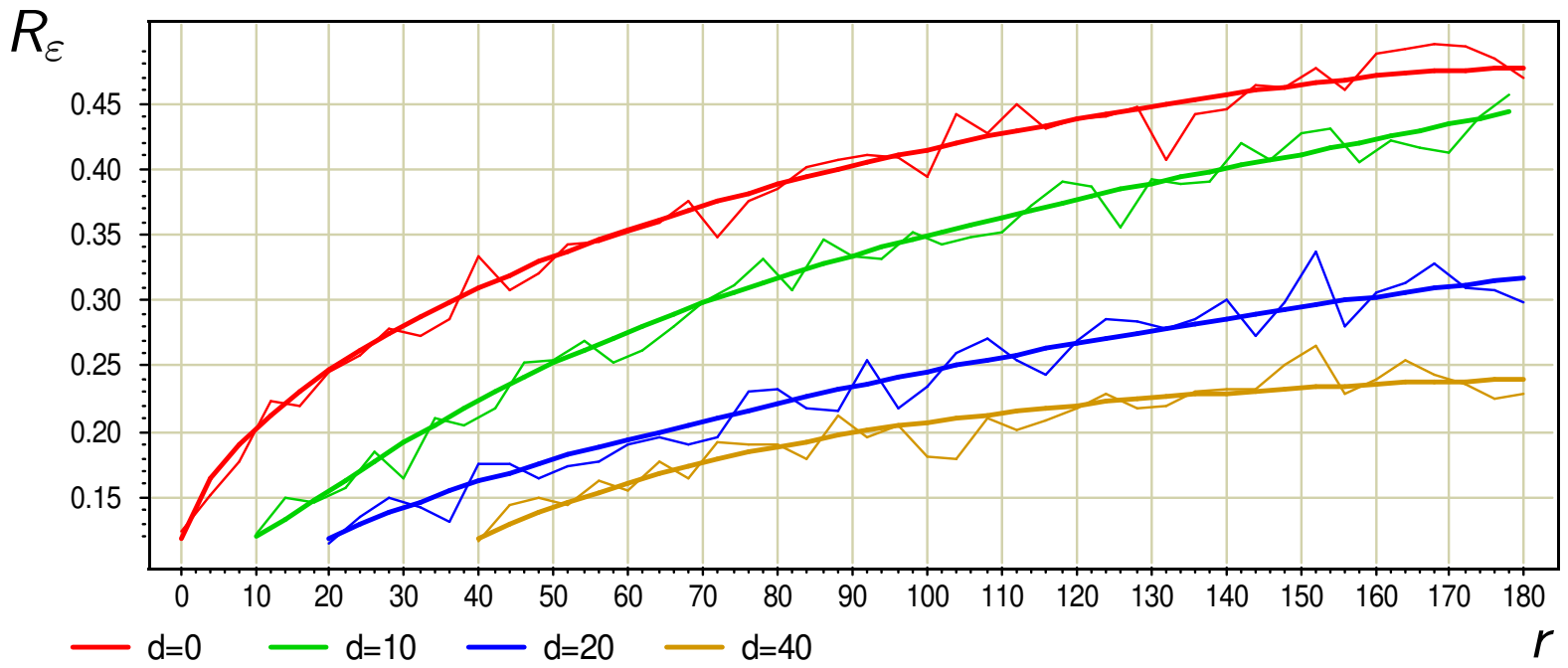
Расслоение измеряется разностью числа ошибок:

$$d(a_1, a_2) = \left| n(a_1, X^L) - n(a_2, X^L) \right| = |m_1 - m_2|$$

Условия эксперимента: $\ell = k = 100$, $m_0 = 20$, $\varepsilon = 0.05$, метод Монте-Карло по $|N| = 10^4$ случайных разбиений.

Эффекты сходства и расслоения для пары алгоритмов

Зависимость вероятности переобучения R_ε
от расстояния Хэмминга r и расслоения d:



- переобучение возникает даже при выборе из двух алгоритмов
- чем более они схожи, тем меньше переобучение
- чем больше расслоение, тем меньше переобучение

Вероятность переобучения семейства из двух алгоритмов

Пусть для алгоритмов a_1, a_2 известны m_0, m_1, m_2, m_3:

$$a_1 = (1,\ldots,1,\ 1,\ldots,1,\ 0,\ldots,0,\ 0,\ldots,0);$$
$$a_2 = (\underbrace{1,\ldots,1}_{m_0},\ \underbrace{0,\ldots,0}_{m_1},\ \underbrace{1,\ldots,1}_{m_2},\ \underbrace{0,\ldots,0}_{m_3}).$$

Теорема (о вероятности переобучения)

Если $A = \{a_1, a_2\}$ и метод μ минимизирует эмпирический риск (число ошибок на обучении), то для любого $\varepsilon \in [0,1]$

$$R_\varepsilon(\mu, X^L) = \sum_{s_0=0}^{m_0} \sum_{s_1=0}^{m_1} \sum_{s_2=0}^{m_2} \frac{C_{m_0}^{s_0} C_{m_1}^{s_1} C_{m_2}^{s_2} C_{L-m_0-m_1-m_2}^{\ell-s_0-s_1-s_2}}{C_L^\ell} \times$$
$$\times \Big(\big[s_1 < s_2\big] \big[s_0 + s_1 \leqslant \tfrac{\ell}{L}(m_0 + m_1 - \varepsilon k)\big] +$$
$$+ \big[s_1 \geqslant s_2\big] \big[s_0 + s_2 \leqslant \tfrac{\ell}{L}(m_0 + m_2 - \varepsilon k)\big] \Big)$$

Граф расслоения–связности множества алгоритмов

Определим бинарные отношения на множестве алгоритмов A:

частичный порядок $a \leqslant b$: $I(a,x) \leqslant I(b,x)$ для всех $x \in X^L$;

предшествование $a \prec b$: $a \leqslant b$ и $\|b-a\| = 1$.

Опр. *Граф расслоения–связности* $\langle A, E \rangle$:

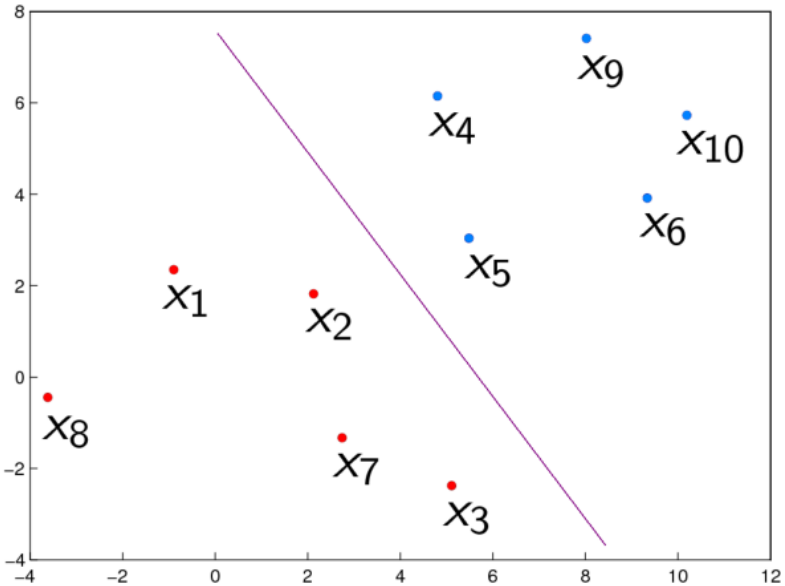
A — множество попарно различных векторов ошибок;

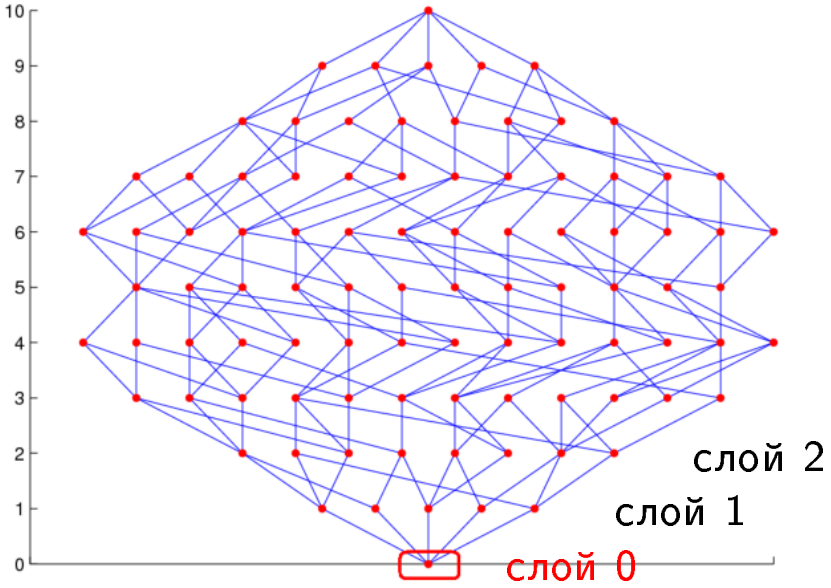
$E = \{(a,b)\colon a \prec b\}$.

Свойства графа расслоения–связности:

- это подграф графа Хассе отношения порядка $\leqslant$ на A;
- каждому ребру (a,b) соответствует *рёберный объект* $x_{ab} \in X^L$, такой, что $I(a,x_{ab}) = 0, \quad I(b,x_{ab}) = 1$;
- граф является многодольным со слоями $A_m = \{a \in A\colon n(a,X^L) = m\}, \quad m = 0,\ldots,L$;

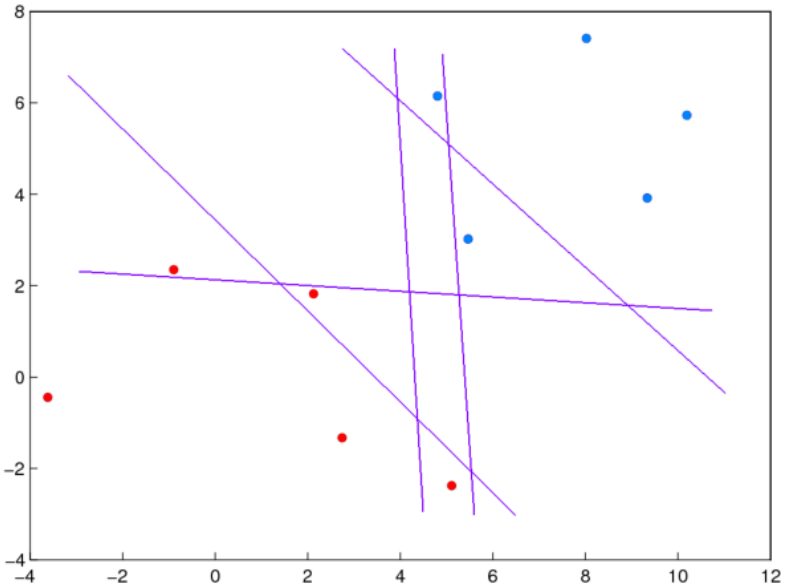
Пример 1. Семейство линейных алгоритмов классификации

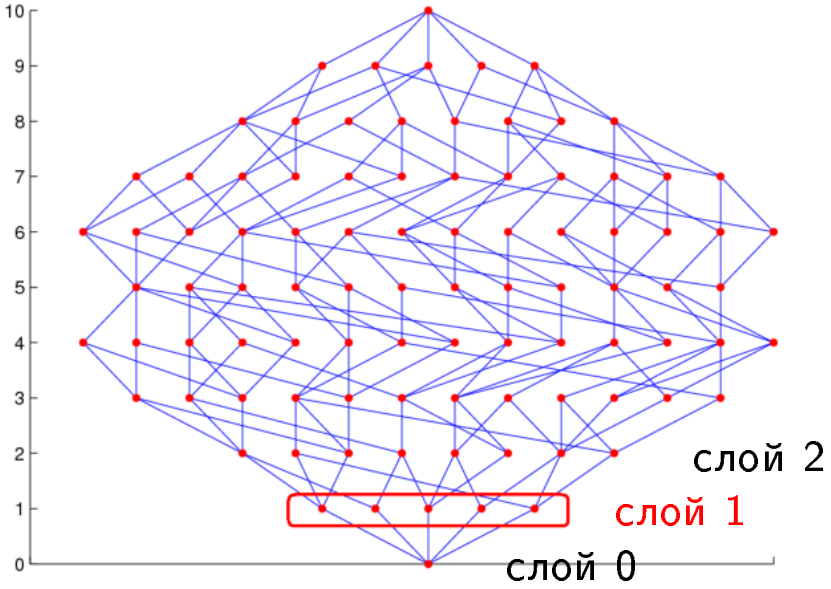




	слой 0
x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0

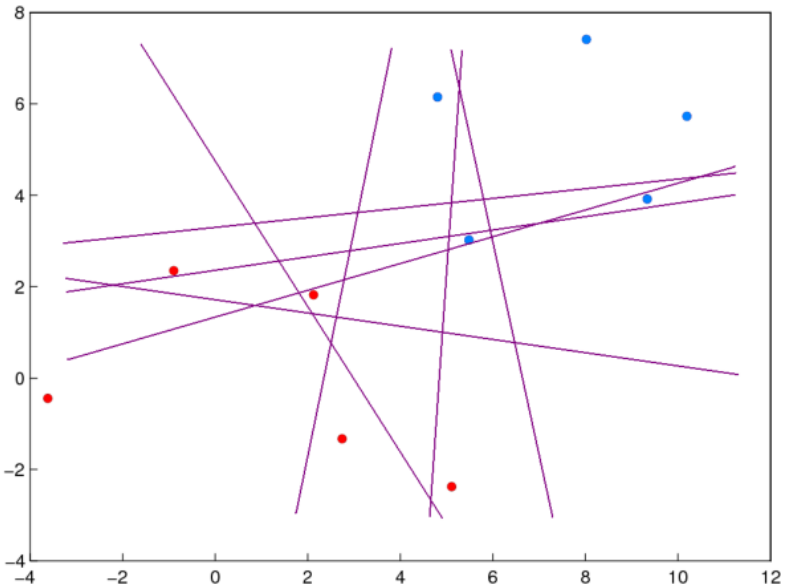
Пример 1. Семейство линейных алгоритмов классификации

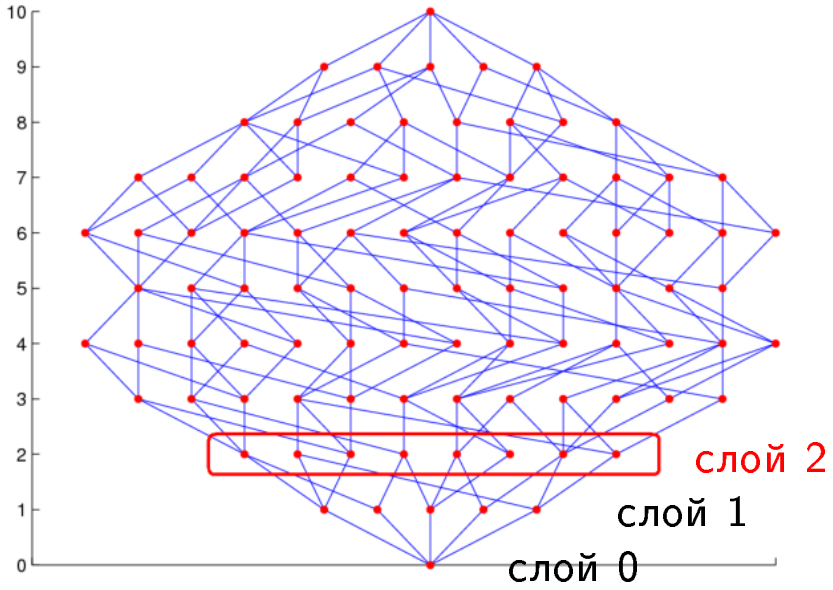




	слой 0	слой 1				
x_1	0	1	0	0	0	0
x_2	0	0	1	0	0	0
x_3	0	0	0	1	0	0
x_4	0	0	0	0	1	0
x_5	0	0	0	0	0	1
x_6	0	0	0	0	0	0
x_7	0	0	0	0	0	0
x_8	0	0	0	0	0	0
x_9	0	0	0	0	0	0
x_{10}	0	0	0	0	0	0

Пример 1. Семейство линейных алгоритмов классификации





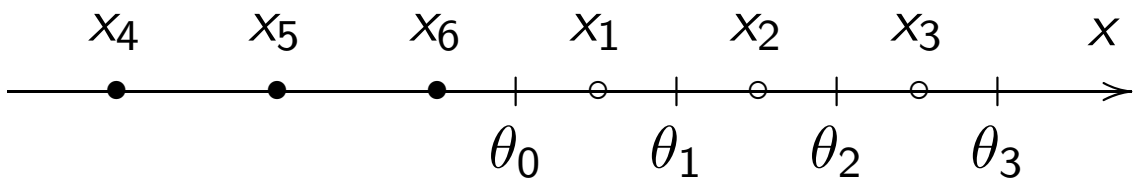
	слой 0	слой 1					слой 2								
x_1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	...
x_2	0	0	1	0	0	0	1	1	0	0	0	0	0	0	...
x_3	0	0	0	1	0	0	0	1	1	0	0	0	0	1	...
x_4	0	0	0	0	1	0	0	0	1	1	0	0	0	0	...
x_5	0	0	0	0	0	1	0	0	0	1	1	1	0	0	...
x_6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	...
x_7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	...
x_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_{10}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...

Пример 2. Монотонная цепь

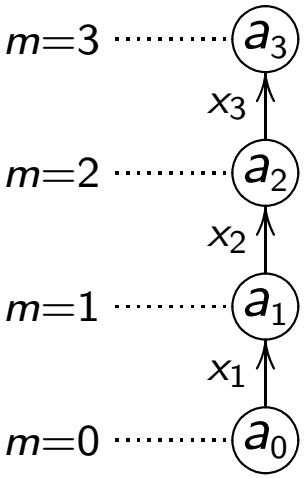
Опр. *Монотонная цепь* алгоритмов: $a_0 \prec a_1 \prec \cdots \prec a_D$.

Пример: 1D пороговый классификатор $a_d(x) = [x - \theta_d]$;

2 класса $\{\bullet, \circ\}$
6 объектов



Граф семейства:



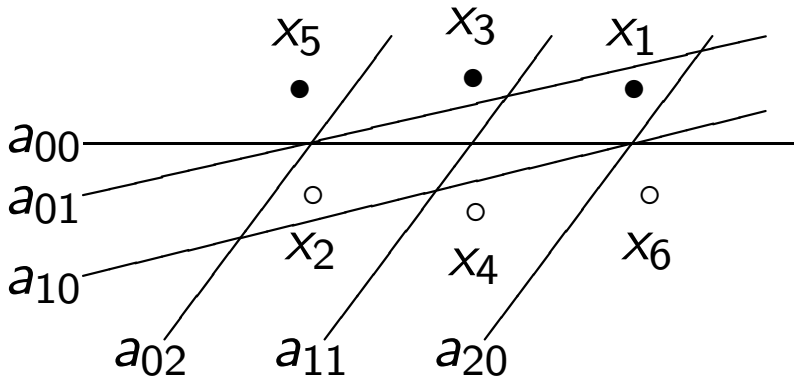
Матрица ошибок:

	a_0	a_1	a_2	a_3
x_1	0	1	1	1
x_2	0	0	1	1
x_3	0	0	0	1
x_4	0	0	0	0
x_5	0	0	0	0
x_6	0	0	0	0

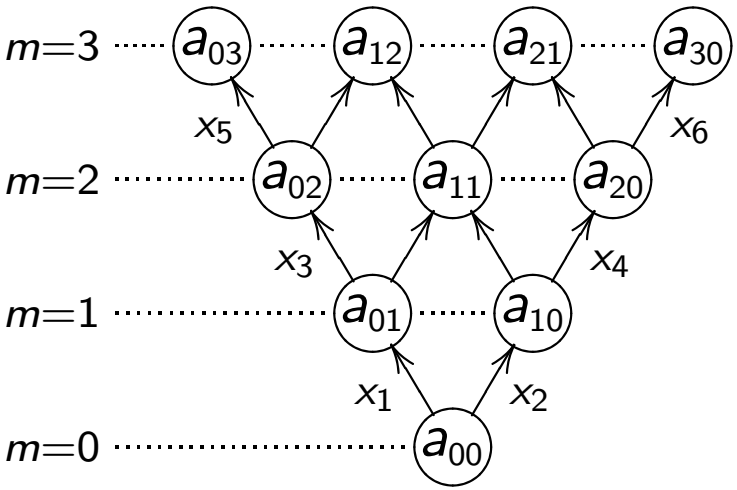
Пример 3. Двумерная сеть классификаторов

Пример:

2D линейный классификатор,
2 класса $\{\bullet, \circ\}$,
6 объектов



Граф семейства:



Матрица ошибок:

	a_{00}	a_{01}	a_{10}	a_{02}	a_{11}	a_{20}	a_{03}	a_{12}	a_{21}	a_{30}
x_1	0	1	0	1	1	0	1	1	1	0
x_2	0	0	1	0	1	1	0	1	1	1
x_3	0	0	0	1	0	0	1	1	0	0
x_4	0	0	0	0	0	1	0	0	1	1
x_5	0	0	0	0	0	0	1	0	0	0
x_6	0	0	0	0	0	0	0	0	0	1

Порождающее и запрещающее множества алгоритмов

Определение

Верхняя связность $u(a)$ алгоритма a — это число всех рёбер, исходящих из вершины a в графе расслоения–связности:

$$u(a) = |X_a|, \quad X_a = \left\{ x_{ab} \in X^L \;\middle|\; a \prec b \right\};$$

X_a называется *порождающим множеством* алгоритма a.

Определение

Ошибочность $q(a)$ алгоритма a — это число различных рёберных объектов на всех путях, ведущих в a:

$$q(a) = |X'_a|, \quad X'_a = \left\{ x \in X^L \;\middle|\; \exists b \in A\colon\, b < a,\; I(b,x) < I(a,x) \right\};$$

X'_a называется *запрещающим множеством* алгоритма a.

Характеристики расслоения и связности алгоритма

Верхняя связность $u(a) = \#\{x_{ab} \in X^L \mid a \prec b\}$

Нижняя связность $d(a) = \#\{x_{ba} \in X^L \mid b \prec a\}$

Ошибочность $q(a) = \#\{x \in X^L \mid \exists b \in A\colon\ b < a,\ I(b,x) < I(a,x)\}$

Число ошибок $m(a) = n(a, X^L)$.

Утв.

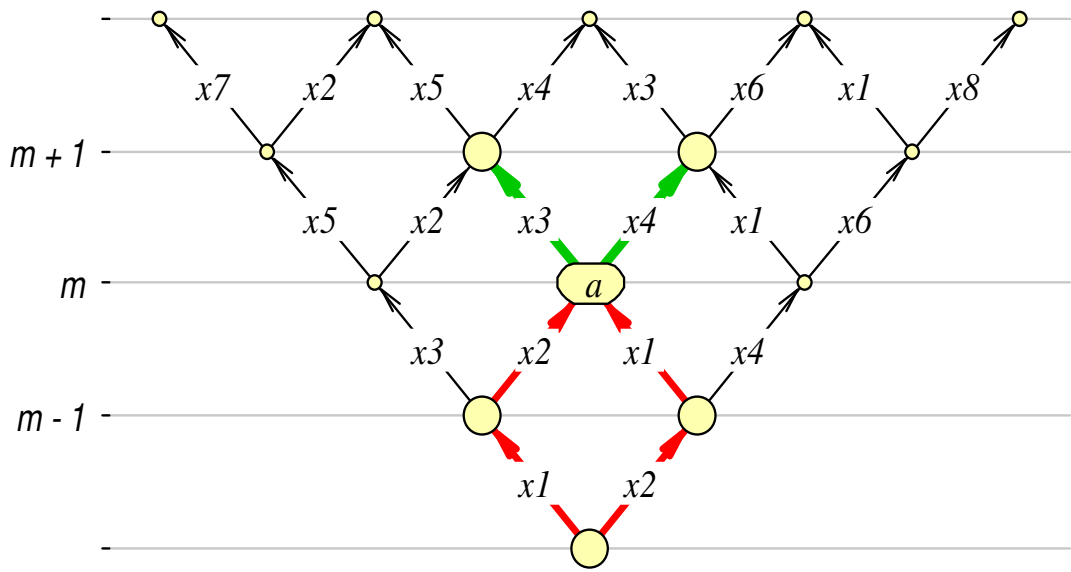
$d(a) \leqslant q(a) \leqslant m(a)$

Пример: двумерная сеть алгоритмов

$u(a) = \#\{x3, x4\} = 2$

$d(a) = \#\{x1, x2\} = 2$

$q(a) = \#\{x1, x2\} = 2$



Верхняя оценка расслоения–связности

Метод минимизации эмпирического риска μ *монотонный*, если

$$\mu(X^\ell) \in A_K(X^\ell) = \operatorname{Arg}\min_{a\in A} K(a, X^\ell),$$

где $K(a, U)$ — строго монотонная функция вектора ошибок a:
для любых $U \subset X^L$, $a, b \in A$ если $a < b$, то $K(a, X) < K(b, X)$.

Пример. Функция $K(a, U) = \nu(a, U)$ — строго монотонная.

Теорема

Для любого монотонного метода μ, любых X^L, A и $\varepsilon \in (0, 1)$

$$R_\varepsilon(\mu, X^L) \leqslant \sum_{a\in A} \frac{C^{\ell-u}_{L-u-q}}{C^\ell_L}\, \mathcal{H}^{\ell-u,\, m-q}_{L-u-q}\left(\tfrac{\ell}{L}\left(m - \varepsilon k\right)\right)$$

K.V.Vorontsov, A.A.Ivahnenko, I.M.Reshetnyak. Generalization bound based on the splitting and connectivity graph of the set of classifiers. 2010.

Идея доказательства

1. Построим по μ *монотонный пессимистичный* метод $\bar{\mu}$ максимизации переобученности: $\bar{\mu}(X^\ell) = \arg\max_{a \in A_K(X^\ell)} \delta(a, X^\ell, X^k)$.

Тогда $R_\varepsilon(\mu, X^L) \leqslant R_\varepsilon(\bar{\mu}, X^L)$ — верхняя оценка.

2. Если $\bar{\mu}(X^\ell) = a$, то $\begin{cases} X_a \subseteq X^\ell \text{ в силу пессимистичности } \bar{\mu}, \\ X'_a \subseteq X^k \text{ в силу монотонности } \bar{\mu}. \end{cases}$

3. $\mathsf{P}\Big[\bar{\mu}(X^\ell) = a\Big] \leqslant \mathsf{P}\big[\underbrace{X_a \subseteq X^\ell \text{ и } X'_a \subseteq X^k}_{S(a, X^\ell)}\big] = \dfrac{C^{\ell-|X_a|}_{L-|X_a|-|X'_a|}}{C^\ell_L} = \dfrac{C^{\ell-u}_{L-u-q}}{C^\ell_L}.$

4. По формуле полной вероятности:

$$R_\varepsilon(\bar{\mu}, X^L) \leqslant \sum_{a \in A} \underbrace{\mathsf{P}\Big[S(a, X^\ell)\Big]}_{C^{\ell-u}_{L-u-q} / C^\ell_L} \cdot \underbrace{\mathsf{P}\Big[\delta(a, X^\ell) \geqslant \varepsilon \;\Big|\; S(a, X^\ell)\Big]}_{\mathcal{H}^{\ell-u,\, m-q}_{L-u-q}\left(\frac{\ell}{L}(m - \varepsilon k)\right)}. \quad \blacksquare$$

Свойства верхней оценки расслоения–связности

1. При $|A| = 1$ функция гипергеометрического распределения:
$$R_\varepsilon = \mathcal{H}_L^{\ell,m}\left(\tfrac{\ell}{L}\big(m - \varepsilon k\big)\right) \to 0 \quad \text{при} \quad \ell, k \to \infty.$$
2. При $q = u = 0$ и $\ell = k$ это оценка Вапника-Червоненкиса:
$$R_\varepsilon \leqslant \sum_{a \in A} \mathcal{H}_L^{\ell,m}\left(\tfrac{\ell}{L}\big(m - \varepsilon k\big)\right) \leqslant |A| \cdot \tfrac{3}{2}\exp(-\varepsilon\ell^2).$$
3. Вклад алгоритма $a \in A$ убывает экспоненциально
по $u(a) \Rightarrow$ **связные семейства меньше переобучаются**;
по $q(a) \Rightarrow$ **только нижние слои вносят вклад в R_ε.**
4. Вероятность получить алгоритм в результате обучения
$$\mathsf{P}\big[\mu(X^\ell) = a\big] \leqslant P_a = C_{L-u-q}^{\ell-u} / C_L^\ell$$
5. Оценка является точной (обращается в равенство)
в случае многомерных монотонных сетей алгоритмов.

М.М.Бонгард ввёл основные принципы машинного обучения

- комбинаторный перебор формул (моделей зависимости)
- критерии информативности и принцип «развала на кучи»
- принцип синтеза информативных признаков
- принцип голосования
- принцип обучения с учётом неразмеченных данных
- принцип экзамена и скользящего контроля
- понятие предрассудка

Переобучение — вывод предрассудков из верных данных

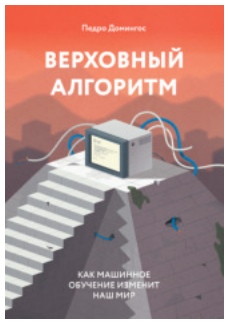
- Происходит даже при выборе лучшего из двух решений
- Без расслоения и связности переобучение наступает уже при нескольких десятках алгоритмов
- Расслоение и связность сильно уменьшают переобучение и способны компенсировать влияние размерности
- На практике семейства, как правило, ими обладают

Бонгард М. М. Проблема узнавания, 1967.

Основные школы машинного обучения

1. *символизм* – поиск логических закономерностей
   - Decision Tree, Rule Induction
2. *коннекционизм* – обучаемые нейронные сети
   - BackPropagation, Deep Belief Nets, Deep Learning
3. *эволюционизм* – саморазвитие сложных моделей
   - Genetic Algorithms, Genetic Programming
4. *байесионизм* – оценивание распределений параметров
   - Naive Bayes, Bayesian Networks, Graphical Models
5. *аналогизм* – «близким объектам близкие ответы»
   - kNN, RBF, SVM, Kernel Smoothing

⊕ *композиционизм* – кооперация моделей
   - Weighted Voting, Boosting, Bagging, Stacking, Random Forest, Яндекс.CatBoost



Домингос П. Верховный алгоритм. 2016 (2015).