Слабая вероятностная аксиоматика и надёжность эмпирических предсказаний

Воронцов Константин Вячеславович
voron@ccas.ru, http://www.ccas.ru/voron

Вычислительный Центр РАН,
Москва, Вавилова 40, 119991

Всероссийская конференция
Математические методы распознавания образов (ММРО-13)
30 сентября – 5 октября, 2007
Санкт-Петербург

Содержание

Слабая вероятностная аксиоматика

1. $X^L = \{x_i\}_{i=1}^L$ — конечная выборка объектов из $\mathbb{X}$.
2. Все разбиения $X^L = X_n^\ell \cup X_n^k$, $n = 1, \ldots, N$, где $N = C_L^k$, $L = \ell + k$, имеют равные шансы реализоваться.

Тогда...

Слабая вероятностная аксиоматика

1. $X^L = \{x_i\}_{i=1}^L$ — конечная выборка объектов из $\mathbb{X}$.
2. Все разбиения $X^L = X_n^\ell \cup X_n^k$, $n = 1, \ldots, N$, где $N = C_L^k$, $L = \ell + k$, имеют равные шансы реализоваться.

Тогда:

- *событие* A есть функция $A\colon \{1, \ldots, N\} \to \{0, 1\}$;
- *вероятность события* есть доля таких разбиений n, для которых $A(n) = 1$:

$$\mathsf{P}_n A(n) = \frac{1}{N} \sum_{i=1}^{N} A(n).$$

Задача эмпирического предсказания

- Реализуется разбиение (X_n^ℓ, X_n^k), $n \in \{1, \ldots, N\}$;
  выборка X_n^ℓ — *наблюдаемая*, выборка X_n^k — *скрытая*.
- Задана функция двух выборок $T\colon \mathbb{X}^k \times \mathbb{X}^\ell \to R$
- Требуется:

  1. Выбрать функцию $\hat{T}\colon \mathbb{X}^\ell \to R$ так, чтобы значение $\hat{T}_n = \hat{T}(X_n^\ell)$ предсказывало бы $T_n = T(X_n^k, X_n^\ell)$.

  2. Оценить точность предсказаний:

$$\mathsf{P}_n\big[d(\hat{T}_n, T_n) > \varepsilon\big] \leqslant \eta(\varepsilon),$$

  где $d(\hat{r}, r)$ — отклонение предсказания $\hat{r}$ от истины r, например $d(\hat{r}, r) = |\hat{r} - r|$.

Слабая вероятностная аксиоматика

- Достаточна для доказательства фундаментальных фактов:
  - закон больших чисел, с точной оценкой скорости сходимости;
  - отклонение эмпирических распределений (критерий Смирнова), с точной оценкой скорости сходимости;
  - непараметрические критерии проверки стат. гипотез;
  - оценки типа Вапника-Червоненкиса.

Слабая вероятностная аксиоматика

- Достаточна для доказательства фундаментальных фактов:
  - закон больших чисел, с точной оценкой скорости сходимости;
  - отклонение эмпирических распределений (критерий Смирнова), с точной оценкой скорости сходимости;
  - непараметрические критерии проверки стат. гипотез;
  - оценки типа Вапника-Червоненкиса.
- Основана на более слабых предположениях:
  - нет определения вероятности как меры на $\mathbb{X}$;
  - нет определения вероятности как частоты при $L \to \infty$;

Слабая вероятностная аксиоматика

- Достаточна для доказательства фундаментальных фактов:
  - закон больших чисел, с точной оценкой скорости сходимости;
  - отклонение эмпирических распределений (критерий Смирнова), с точной оценкой скорости сходимости;
  - непараметрические критерии проверки стат. гипотез;
  - оценки типа Вапника-Червоненкиса.
- Основана на более слабых предположениях:
  - нет определения вероятности как меры на $\mathbb{X}$;
  - нет определения вероятности как частоты при $L \to \infty$;
  - нет необходимости определять понятие «вероятность»!

Пример 1: Закон больших чисел

Опр. *Частота* события $S \subset \mathbb{X}$ на конечной выборке $U \subset \mathbb{X}$:

$$\nu_S(U) = \frac{1}{|U|} \sum_{u \in U} [u \in S].$$

Положим: $R = \mathbb{R}$, $\hat{T}(U) = T(U) = \nu_S(U)$.

Теорема (известный классический факт)

Частота $\nu_S(X_n^\ell)$ *предсказывает частоту* $\nu_S(X_n^k)$, *причём справедлива точная оценка*

$$\mathsf{P}_n\big[\nu_S(X_n^k) - \nu_S(X_n^\ell) \geqslant \varepsilon\big] = H\tbinom{\ell}{L}{s(\varepsilon)}{m},$$

где $H\tbinom{\ell}{L}{s}{m}$ *— левый хвост гипергеометрического распределения,* $s(\varepsilon) = \left\lfloor \frac{\ell}{L}(m - \varepsilon k) \right\rfloor$, $m = L\nu_S(X^L)$.

Пример 1: Закон больших чисел

Левый хвост гипергеометрического распределения:

$$H\binom{\ell\ s(\varepsilon)}{L\ m} = \sum_{t=s_0}^{s(\varepsilon)} \frac{C_m^t C_{L-m}^{\ell-t}}{C_L^\ell}, \quad s_0 = \max\{0, m-k\}$$

Пример 2: Сходимость эмпирических распределений

Опр. *Эмпирическое распределение* функции $\xi\colon \mathbb{X} \to \mathbb{R}$ на конечной выборке $U \subseteq \mathbb{X}$ есть

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

Положим: R — пр-во кус.-пост. невозр. функций $F\colon \mathbb{R} \to [0, 1]$,
$T(U) = \hat{T}(U) = F_\xi(z, U)$, $d(\hat{r}, r) = \max\limits_{z \in \mathbb{R}} |r(z) - \hat{r}(z)|$.

Теорема (малоизвестный факт)

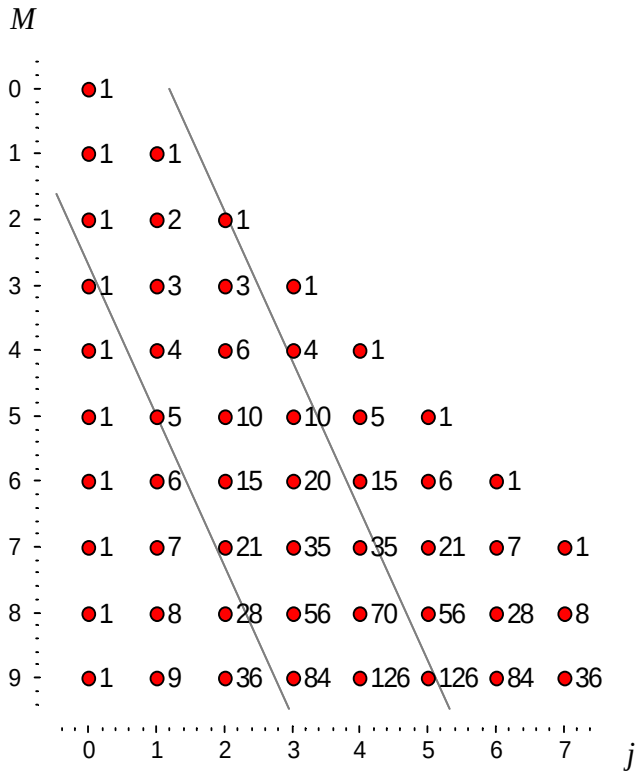
Если значения $\xi(x_i)$ попарно различны на X^L, то

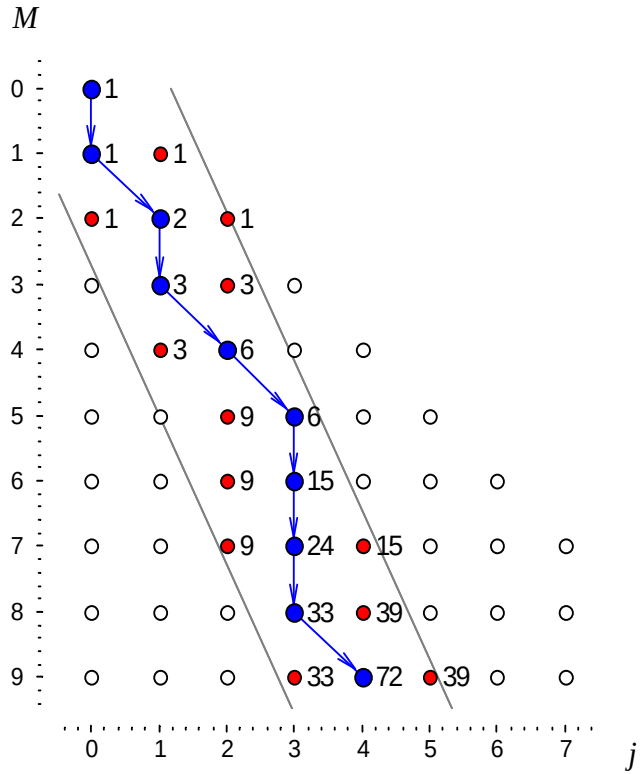
$$\mathsf{P}_n\Big\{\max_{z \in \mathbb{R}} |F_\xi(z, X_n^k) - F_\xi(z, X_n^\ell)| > \varepsilon\Big\} = \frac{G_L^k(\varepsilon)}{C_L^k},$$

где $G_L^k(\varepsilon)$ — значение из усечённого треугольника Паскаля.

Пример 2: Сходимость эмпирических распределений

Усечённый треугольник Паскаля $G_m^j(\varepsilon)$ — между двумя прямыми $j^+(m) = \frac{k}{L}(m + \varepsilon\ell)$, $j^-(m) = \frac{k}{L}(m - \varepsilon\ell)$.





Здесь $L = 16$, $k = 7$, $\varepsilon = 0.3$.

Связь с классической аксиоматикой Колмогорова

Теорема (Принцип соответствия)

Если в слабой аксиоматике для некоторой функции $\phi(X^\ell, X^k)$ получена оценка

$$Q_\varepsilon(X^L) = \mathsf{P}_n\big[\phi(X_n^\ell, X_n^k) > \varepsilon\big] \leqslant \eta(\varepsilon, X^L)$$

то аналогичная оценка верна и в аксиоматике Колмогорова

$$\mathsf{E}_{X^L} Q_\varepsilon(X^L) = \mathsf{P}_{X^L}\big[\phi(X^\ell, X^k) > \varepsilon\big] \leqslant \mathsf{E}_{X^L}\eta(\varepsilon, X^L)$$

Связь с классической аксиоматикой Колмогорова

Теорема (Принцип соответствия)

Если в слабой аксиоматике для некоторой функции $\phi(X^\ell, X^k)$ получена оценка

$$Q_\varepsilon(X^L) = \mathsf{P}_n\big[\phi(X_n^\ell, X_n^k) > \varepsilon\big] \leqslant \eta(\varepsilon, X^L)$$

то аналогичная оценка верна и в аксиоматике Колмогорова

$$\mathsf{E}_{X^L} Q_\varepsilon(X^L) = \mathsf{P}_{X^L}\big[\phi(X^\ell, X^k) > \varepsilon\big] \leqslant \mathsf{E}_{X^L}\eta(\varepsilon, X^L)$$

Трансдукция или индукция?

Если $\eta(\varepsilon, X^L) \equiv \eta(\varepsilon)$,
то оценка справедлива для любой выборки X^L.

Связь с эмпирическим оцениванием

Скользящий контроль (Cross-Validation)

Если оценку не удаётся получить теоретически или если теоретическая оценка слишком завышена:

$$Q_\varepsilon(X^L) = \mathsf{P}_n\big[d(\hat{T}_n, T_n) > \varepsilon\big] \leqslant \boxed{???},$$

то можно измерить Q_ε эмпирически:

$$Q_\varepsilon(X^L) \leqslant \frac{1}{|N'|}\sum_{n\in N'}\big[d(\hat{T}_n, T_n) > \varepsilon\big] + \underbrace{\delta(N, |N'|)}_{\to 0 \text{ при } |N'|\to N},$$

где множество разбиений $N' \subset \{1, \ldots, N\}$ выбирается
— либо случайно (метод Монте-Карло),
— либо по блокам (k-fold Cross-Validation)

Слабая вероятностная аксиоматика: за и против

+ Лучше подходит для задач анализа данных и обучения по прецедентам
− . . . но хуже — для континуальных стохастических явлений

Слабая вероятностная аксиоматика: за и против

+ Лучше подходит для задач анализа данных и обучения по прецедентам
− . . . но хуже — для континуальных стохастических явлений
+ Даёт не асимптотические, не завышенные оценки
− . . . требующие сложных комбинаторных вычислений

Слабая вероятностная аксиоматика: за и против

+ Лучше подходит для задач анализа данных и обучения по прецедентам
− . . . но хуже — для континуальных стохастических явлений
+ Даёт не асимптотические, не завышенные оценки
− . . . требующие сложных комбинаторных вычислений
+ Удовлетворяет «принципу соответствия»
− . . . однако не все классические теоремы имеют аналоги в слабой аксиоматике

Слабая вероятностная аксиоматика: за и против

+ Лучше подходит для задач анализа данных и обучения по прецедентам
− . . . но хуже — для континуальных стохастических явлений
+ Даёт не асимптотические, не завышенные оценки
− . . . требующие сложных комбинаторных вычислений
+ Удовлетворяет «принципу соответствия»
− . . . однако не все классические теоремы имеют аналоги в слабой аксиоматике

Открытая проблема:

Какая часть математической статистики может быть построена в рамках слабой аксиоматики?

Задача обучения по прецедентам

- $\mathbb{X}$ — множество объектов, $\mathbb{Y}$ — множество ответов.
- Бинарная *функция потерь* $\mathscr{L}: \mathbb{X} \times \mathbb{Y} \to \{0, 1\}$.
  - Для задач классификации: $\mathscr{L}(x, y) = \big[y \neq y^*(x)\big]$,
  - Для задач регрессии: $\mathscr{L}(x, y) = \big[|y - y^*(x)| > \delta\big]$,

  где $y^*(x)$ — неизвестная *целевая функция*.
- *Обучающая выборка* $X^\ell = \{x_i\}_{i=1}^\ell$ с известными $\mathscr{L}(x_i, y)$.
- *Метод обучения* $\mu: X^\ell \mapsto f$.
- *Частота ошибок* алгоритма $f: \mathbb{X} \to \mathbb{Y}$ на выборке $U \subset \mathbb{X}$:

$$\nu(f, U) = \frac{1}{|U|} \sum_{u \in U} \mathscr{L}(u, f(u)).$$

- *Обобщающая способность*:
  средняя ошибка $\nu(\mu X^\ell, U)$ должна быть достаточно мала
  для большинства выборок $U \in \mathbb{X}^*$.

Оценки Вапника-Червоненкиса [1971]

- Для семейства алгоритмов F в сильной аксиоматике:

$$P_\varepsilon(F) = \mathsf{P}_{X^L}\Big[\sup_{f\in F}\big(\nu(f,X^k)-\nu(f,X^\ell)\big) > \varepsilon\Big]$$
$$\leqslant \Delta^F(L)\cdot 1.5\, e^{-\varepsilon^2\ell} \quad \text{(при условии } \ell = k\text{)};$$

- $\Delta^F(L)$ — *функция роста* (shatter coefficient) семейства F — макс. число функций $f \in F$, попарно различимых на X^L;
$\Delta^F(L) \leqslant 1.5\,\frac{L^h}{h!}$, $h = \mathrm{VCdim}(F)$ — ёмкость семейства F.

Оценки Вапника-Червоненкиса [1971]

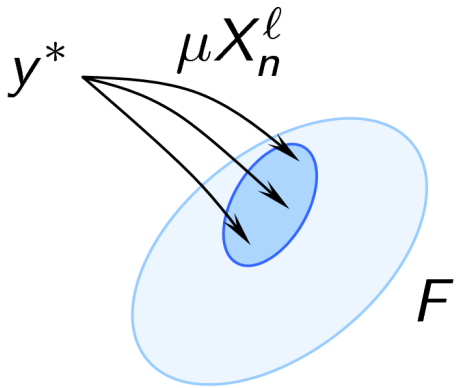
- Для семейства алгоритмов F в сильной аксиоматике:

$$P_\varepsilon(F) = \mathsf{P}_{X^L}\Big[\sup_{f\in F}\big(\nu(f,X^k)-\nu(f,X^\ell)\big) > \varepsilon\Big]$$
$$\leqslant \Delta^F(L)\cdot 1.5\, e^{-\varepsilon^2\ell} \quad \text{(при условии } \ell = k\text{)};$$

- $\Delta^F(L)$ — *функция роста* (shatter coefficient) семейства F — макс. число функций $f \in F$, попарно различимых на X^L; $\Delta^F(L) \leqslant 1.5\,\frac{L^h}{h!}$, $h = \mathrm{VCdim}(F)$ — ёмкость семейства F.
- Оценка крайне завышена, т. к. она не учитывает:
  - особенности выборки X^ℓ;
  - особенности целевой зависимости $y^*(x)$;
  - особенности метода обучения μ;
  - степень различности получаемых алгоритмов;
  - что $1.5\,e^{-\varepsilon^2\ell}$ — лишь асимптотическая аппроксимация.

Оценки, зависящие от данных

- *Эффект локализации:*
  Если зависимость y^*, метод μ и выборка X^L фиксированы, то не все функции из F могут быть результатом обучения.



- Равномерная сходимость [Вапник, Червоненкис, 1969]
- Theory of learnable (PAC-learning) [Valiant, 1982]
- Concentration inequalities [Talagrand, 1995]
- Data-dependent bounds [Haussler, 1992 – Bartlett, 1998 – ...]
- Self-bounding learning algorithms [Freund, 1998]
- PAC-Bayesian model averaging [McAllester, 1999]
- Microchoice bounds [Langford, Blum, 2001]
- Algorithmic luckiness [Herbrich, Williamson, 2002]
- Tight sample complexity bounds [Langford, PhD, 2002]

Оценки, зависящие от данных

- В слабой вероятностной аксиоматике:

$$Q_\varepsilon(\mu, X^L) = \mathsf{P}_n\Big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \varepsilon\Big]$$
$$\leqslant \Delta_L^\ell(\mu, X^L) \cdot \max_m H\binom{\ell\ s(\varepsilon)}{L\ m}$$
$$\Big(\leqslant \Delta^F(L) \cdot 1.5\, e^{-\varepsilon^2 \ell}\Big);$$

  где алгоритм $f_n = \mu X_n^\ell$ — результат обучения на X_n^ℓ;
- $\Delta_L^\ell(\mu, X^L)$ — *локальный коэффициент разнообразия (local shatter coefficient)* множества алгоритмов $\big\{f_n = \mu X_n^\ell \;\big|\; n = 1, \ldots, N\big\}$, которые могут быть результатом обучения для данной задачи $\langle \mu, X^L \rangle$:

Оценки, зависящие от данных: дальнейшее уточнение

- **Идея:** скалярная характеристика разнообразия содержит слишком мало информации о процессе обучения.
- Вводится *профиль разнообразия (shatter profile)* $\{D_m\}_{m=0}^{L}$:

$$\Delta_L^\ell(\mu, X^L) = \sum_{m=1}^{L} D_m(\mu, X^L),$$

где $D_m(\mu, X^L)$ — локальный коэффициент разнообразия множества алгоритмов $\left\{f_n \;\middle|\; \nu(f_n, X^L) = \frac{m}{L},\ n = 1, \ldots, N\right\}$.

Оценки, зависящие от данных: дальнейшее уточнение

- **Идея:** скалярная характеристика разнообразия содержит слишком мало информации о процессе обучения.
- Вводится *профиль разнообразия (shatter profile)* $\{D_m\}_{m=0}^{L}$:

$$\Delta_L^\ell(\mu, X^L) = \sum_{m=1}^{L} D_m(\mu, X^L),$$

где $D_m(\mu, X^L)$ — локальный коэффициент разнообразия множества алгоритмов $\left\{ f_n \;\middle|\; \nu(f_n, X^L) = \frac{m}{L},\ n = 1, \ldots, N \right\}$.

Теорема

В слабой аксиоматике справедлива оценка:

$$Q_\varepsilon(\mu, X^L) \leqslant \sum_{m=1}^{L} D_m(\mu, X^L) \cdot H\binom{\ell\ s(\varepsilon)}{L\ m};$$

Эффективный локальный профиль разнообразия

- **Обратная задача**:
  при каком профиле D_m оценка была бы точной?
- **Теорема**

$$Q_{\varepsilon,m}(\mu, X^L) = \mathsf{P}_n\Big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \varepsilon\Big]\Big[\nu(f_n, X^L) = \tfrac{m}{L}\Big]$$
$$\leqslant D_m(\mu, X^L) \cdot H\tbinom{\ell\ s(\varepsilon)}{L\ m};$$

Заменим здесь «$\leqslant$» на «$=$» и выразим D_m:

Эффективный локальный профиль разнообразия

- **Обратная задача**:
  при каком профиле D_m оценка была бы точной?
- **Теорема**

$$Q_{\varepsilon,m}(\mu, X^L) = \mathsf{P}_n\Big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \varepsilon\Big]\Big[\nu(f_n, X^L) = \frac{m}{L}\Big]$$
$$\leqslant D_m(\mu, X^L) \cdot H\binom{\ell\ s(\varepsilon)}{L\ m};$$

  Заменим здесь «$\leqslant$» на «$=$» и выразим D_m:
- *Эффективный локальный профиль разнообр.* $\{\hat{D}_m\}_{m=0}^L$:

$$\hat{D}_m = \frac{\frac{1}{|N'|}\sum_{n\in N'}\big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \varepsilon\big]\big[\nu(f_n, X^L) = \frac{m}{L}\big]}{H\binom{\ell\ s(\varepsilon)}{L\ m}}.$$

- *Эффективный локальный коэффициент разнообразия*:
  $\hat{\Delta}_L^\ell = \hat{D}_0 + \cdots + \hat{D}_L$.

Эмпирическое исследование

- Сравнить численно причины завышенности:
  - пренебрежение эффектом локализации
  - свёртка профиля в коэффициент разнообразия
  - пренебрежение степенью различности алгоритмов
- На примере логических алгоритмов классификации [см. доклад А. А. Ивахненко]
  - известна функция роста $\Delta^F(L)$
  - легко оценить снизу локальный профиль $\Delta^\ell_L(\mu, X^L)$
  - легко оценить эффективный локальный профиль

Разновидности профилей качества обучения

- Профиль разнообразия (или сложности) [см. выше]
- Профиль разделимости (margin distribution)
  [Mason, Bartlett, Baxter, 1998]
  [Shawe-Taylor, Cristianini, 1999]
  [Garg 2002, 2003]
- Профиль монотонности [Воронцов, 2002]
- Профиль компактности [Воронцов, 2004]
- Профиль устойчивости [???]
- Профиль чего ещё ???

Определение профиля компактности

- *Профиль компактности* выборки X^L — доля объектов, у которых m-й сосед лежит в другом классе:

$$R(m) = \frac{1}{L}\sum_{i=1}^{L}\left[y_i \neq y_i^{(m)}\right], \quad m = 1, \ldots, L-1.$$

- Функционал *полного скользящего контроля*:

$$Q_c(\mu, X^L) = \frac{1}{N}\sum_{n=1}^{N} \nu\left(\mu X_n^\ell, X_n^k\right).$$

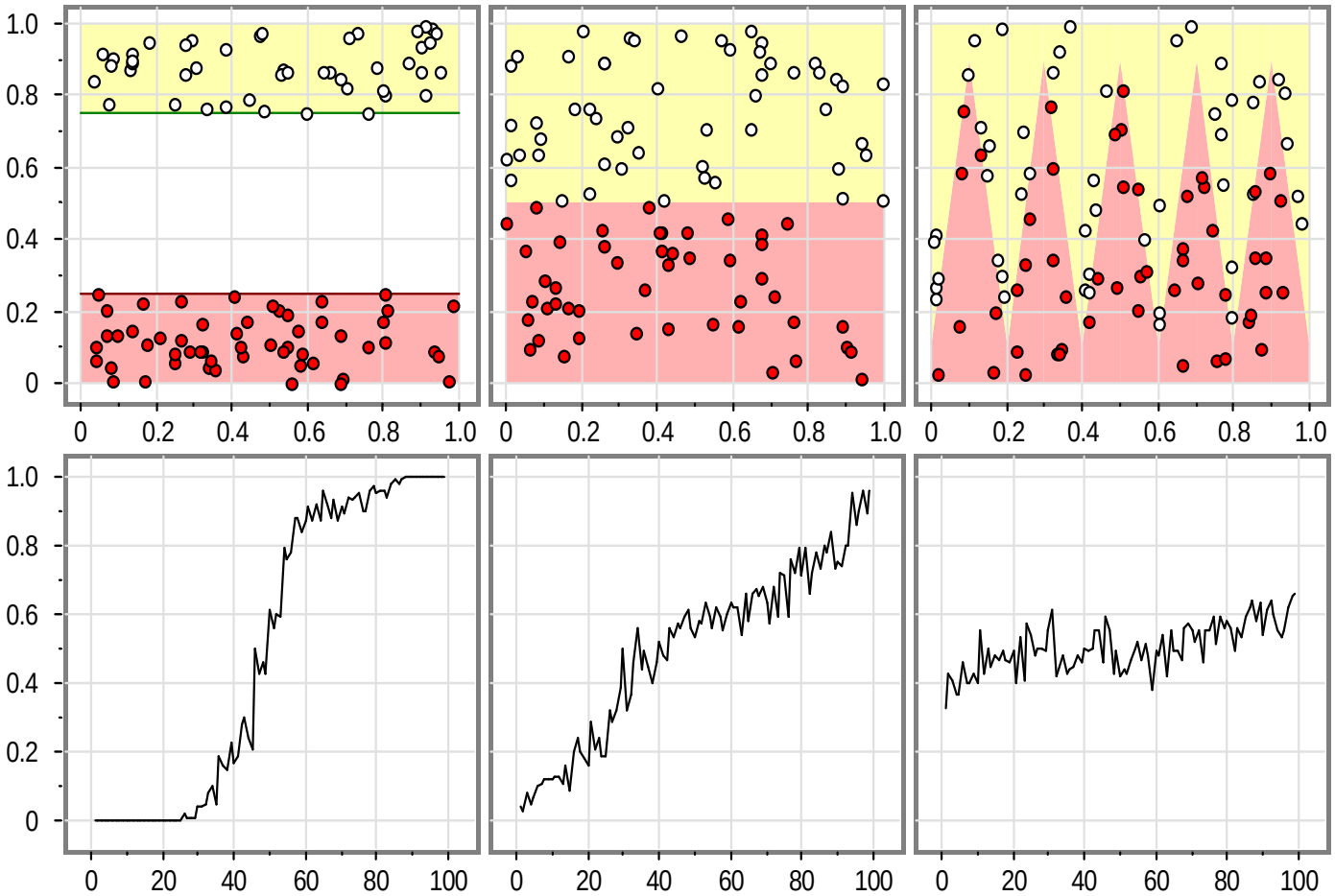
Теорема

Справедлива точная оценка:

$$Q_c(\mu, X^L) = \sum_{m=1}^{k} R(m)\frac{C_{L-1-m}^{\ell-1}}{C_{L-1}^{\ell}}.$$

Пример: профиль компактности

Серия модельных выборок, $L = 100$:



Методика эмпирического измерения качества обучения

- Оценки скользящего контроля
- Профиль разнообразия
- Распределения вариации и смещения по объектам
- Профиль представительности объектов
- Профиль устойчивости
- Профиль разделимости
- Распределение признаков по информативности
- Временны́е показатели эффективности обучения

[см. доклад А. В. Лисицы]