Комбинаторный подход к оцениванию качества алгоритмов, обучаемых по прецедентам

Воронцов К. В.

voron@ccas.ru

Москва
Вычислительный Центр РАН

Постановка задачи

- Восстановление зависимости $y^*: X \to Y$
- Выборка $X^l = \{x_1, ..., x_l\}$ с известными ответами $y^*(x_i)$
- Семейство алгоритмов $A = \{a : X \to Y\}$
- Метод обучения — отображение $\mu : X^l \mapsto a$ из A
- Частота ошибок алгоритма a на выборке X^l:

$$\nu(a, X^l) = \frac{1}{l} \sum_{i=1}^{l} I(a, x_i)$$

Задача:

Оценить качество обобщения $\nu\left(\mu(X^l), X^k\right)$,

где X^k — произвольная (неизвестная) выборка.

Отличия статистической и комбинаторной теории

1. Выборка:

   i.i.d. согласно неизвестной мере $P(X)$

   произвольная

2. Модель процесса обучения:

   минимизация эмпирического риска в семействе A

   конкретный метод обучения $\mu : X^l \mapsto a$

3. Функционал качества:

$$P_\varepsilon(A) = \mathrm{P}_{X^k, X^l} \left\{ \sup_{a \in A} \left(\nu(a, X^k) - \nu(a, X^l) \right) > \varepsilon \right\}$$

$$Q_\varepsilon(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} \left[\nu(a_n, X_n^k) - \nu(a_n, X_n^l) > \varepsilon \right],$$

где $a_n = \mu(X_n^l), \quad N = C_L^l, \quad L = l + k.$

Отличия статистической и комбинаторной теории

4. Оценка функционала качества:

$$P_\varepsilon(A) \quad \le \quad \Delta^A(L) \quad \cdot \exp\left(-2\varepsilon^2 \frac{lk}{l+k}\right)$$

$$Q_\varepsilon(\mu, X^L) \quad \le \quad \Delta^l_L(\mu, X^L) \quad \cdot \Gamma^l_L(\varepsilon, \sigma^l_L)$$

$EQ_\varepsilon \le P_\varepsilon$ — выполняется «принцип соответствия»

5. Мера сложности алгоритмов:

$\Delta^A(L) = \#\left\{ I(a, x_i)_{i=1}^l \,\middle|\, a \in A \right\}$ — Глобальная функция роста

$\Delta^l_L(\mu, X^L) = \#\left\{ I(a_n, x_i)_{i=1}^l \,\middle|\, n = 1, \ldots, N \right\}$ — Локальная ф. роста

Эффект локализации — снимается «запрет на сложность»

Проблема:

как строить методы обучения с хорошей локализующей способностью?

Причины завышенности оценок

$$P_\varepsilon(A) \leq \Delta^A(L) \cdot \exp\left(-2\varepsilon^2 \frac{lk}{l+k}\right)$$

$$Q_\varepsilon(\mu, X^L) \leq \Delta_L^l(\mu, X^L) \cdot \Gamma_L^l(\varepsilon, \sigma_L^l)$$

1. Пренебрежение эффектом локализации:

$$\Delta_L^l(\mu, X^L) \leq \Delta^A(L)$$

2. Погрешность экспоненциальной оценки:

$$\Gamma_L^l(\varepsilon, \sigma_L^l) \leq \exp\left(-2\varepsilon^2 \frac{lk}{l+k}\right)$$

3. Погрешность разложения
   (перехода от анализа качества к анализу сложности)

Вывод:

радикальное уточнение оценок возможно только
при учете свойств метода обучения более тонких,
чем сложность порождаемого семейства алгоритмов.

О методе структурной минимизации риска

Выбор семейства оптимальной сложности
в структуре вложенных семейств возрастающей ёмкости:

$$A_1 \subset A_2 \subset \ldots \subset A_h \subset \ldots;$$

$$\Delta_1 < \Delta_2 < \ldots < \Delta_h < \ldots$$

- С помощью верхней оценки по Вапнику-Червоненкису:

$$h^* = \arg\min_h \left\{ \nu(a_h, X^l) + \sqrt{\tfrac{1}{l}(\ln \Delta_h - \ln P)} \right\},$$

где $P \square 0.05$ — надежность.

- С помощью скользящего контроля непосредственно (cross-validated model selection)

Выбор метода по скользящему контролю

Процедура выбора метода μ^* из конечного набора $\mu_1, ..., \mu_T$:

$$\mu^* = \arg\min_{\mu_1,...,\mu_T} \frac{1}{\tilde{N}} \sum_{n \in \tilde{N}} \nu\left(\mu(X_n^l), X_n^k\right),$$

Разновидности скользящего контроля (cross-validation, CV):

$\tilde{N} = 1$ — hold-out;

$\tilde{N} = C_L^l$ — complete CV (при $k = 1$ — leave-one-out);

$1 < \tilde{N} < C_L^l$ — bootstrap.

Проблемы:

как оценить качество μ^* ?

в каких случаях возникает переобучение ?

как зависит качество от l, k, T, $\tilde{N}$, средн. $\nu\left(\mu^*(X_n^l), X_n^k\right)$?

какая разновидность CV лучше ?

Выбор метода по скользящему контролю

Вводится третья выборка — рабочая:

$$X^L = X^l \cup X^k \cup X^q.$$

Функционал качества:

$$Q_\varepsilon^{lkq}(\mu^*, X^L) = \frac{1}{M}\sum_{m=1}^{M}\Big[\nu(a_m, X_m^q) - \nu(a_m, X_m^k) > \varepsilon\Big],$$

где $M = \frac{L!}{l!k!q!}$, $\quad a_m = \mu^*(X_m^l)$

Теорема. Для случая hold-out

$$Q_\varepsilon^{lkq} \le T \cdot \Gamma_{k+q}^k(\varepsilon, 1)$$

1. Оценка не зависит от l и сложности семейств $A_1, ..., A_T$.
2. Существует предел $Q_\varepsilon^{lkq} \to Te^{-2\varepsilon^2 k}$ при $q \to \infty$.
3. Переобучение возможно, когда T велико и k мало.

Методы обучения, выделяющие опорные объекты

Опр. Метод обучения μ выделяет *h* опорных объектов на X^l, если $\exists X^h \ \forall X^d \ X^h \subseteq X^d \subseteq X^l$ выполняется $\mu(X^d) = \mu(X^h)$.

Теорема.

Если метод μ корректный и выделяет *h* опорных объектов из $\forall X^l \subseteq X^L$, то

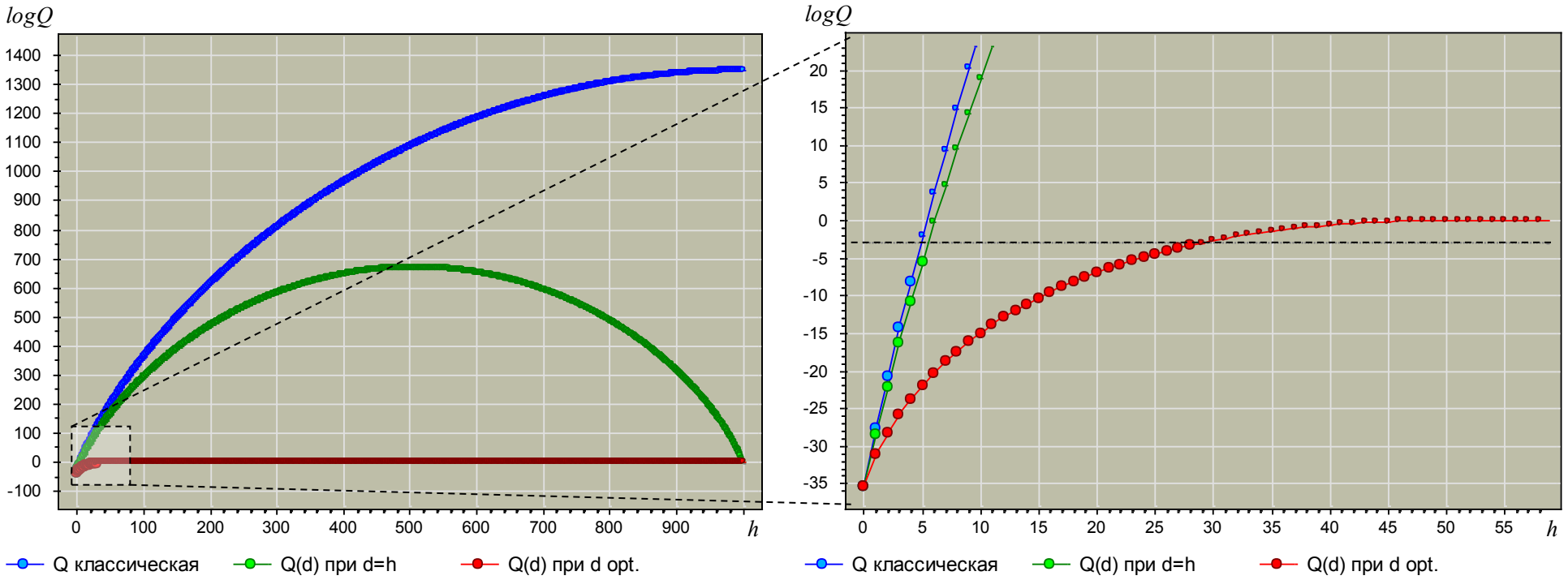
$$Q_\varepsilon(\mu, X^L) \le \frac{C_L^d C_{L-d-m}^{l-d}}{C_L^l C_{l-h}^{d-h}}, \text{ где } d = \left\lfloor \frac{Lh-m}{h+m} \right\rfloor, \ m = \lceil \varepsilon k \rceil.$$

Классическая оценка:

Функция роста не превышает C_L^h, поэтому

$$Q_\varepsilon(\mu, X^L) \le C_L^h \frac{C_{L-m}^l}{C_L^l}.$$

Зависимость оценок $\ln Q_\varepsilon$ от числа опорных объектов *h* (численный расчет при $l = k = 1000, \varepsilon = 0.05$)



Вывод: оценка позволяет обоснованно наращивать число опорных объектов.

Априорное ограничение — компактность

- Простейший метрический алгоритм — метод 1-NN
- Профиль компактности выборки X^L:

$$K(m, X^L) = \frac{1}{L}\sum_{i=1}^{L} I\left(x_i, y^*(x_{im})\right), \qquad m = 1,\ldots, L-1,$$

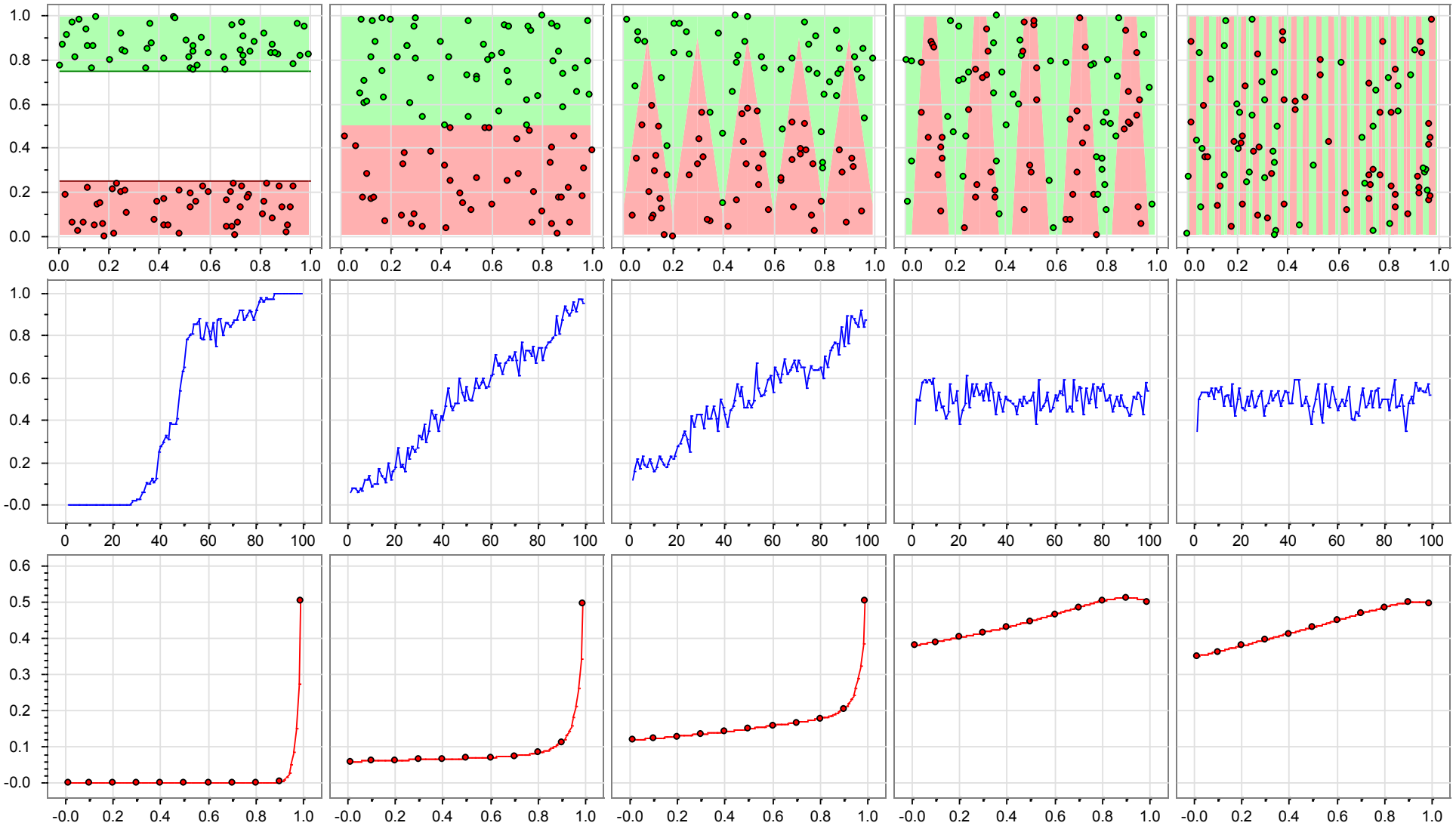
x_{im} — m-й сосед x_i в порядке возрастания $\rho(x_i, x_{im})$.

Теорема. Точное выражение $Q_c(\mu, X^L)$:

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

Достаточно, чтобы $K(m)$ было близко к 0 только при малых m.

Профили компактности выборки

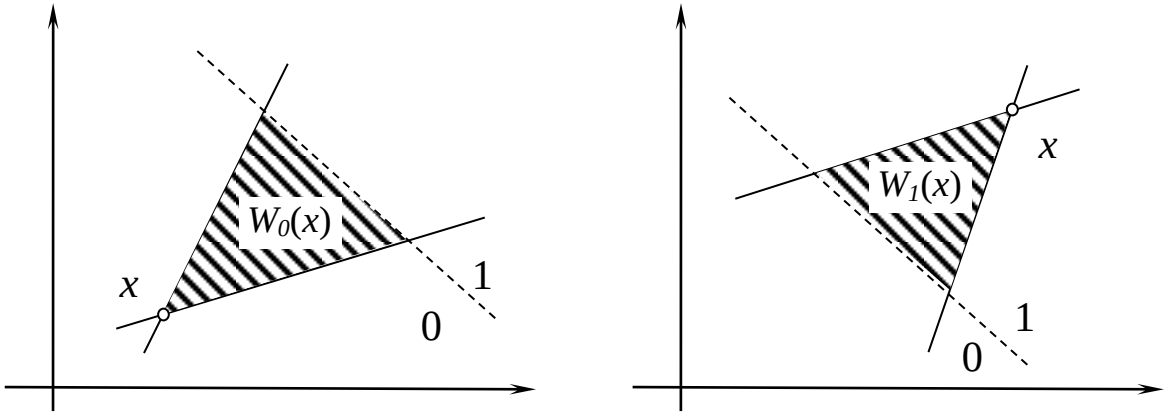


Априорное ограничение — монотонность

- Задача классификации с 2 классами.
  Априорная информация: $y^*: X \to Y$ монотонная
- Клин объекта x_i:

  Верхний клин: $W_0(x_i) = \left\{ x_k \in X^L \,\middle|\, x_i < x_k \text{ и } y_k = 0 \right\}$;

  Нижний клин: $W_1(x_i) = \left\{ x_k \in X^L \,\middle|\, x_i > x_k \text{ и } y_k = 1 \right\}$.



Профиль монотонности

Профиль монотонности выборки X^L:

$$M(m, X^L) = \frac{1}{L} \sum_{i=1}^{L} \left[|W_{y_i}(x_i)| = m \right].$$

Теорема. Если μ — корректный метод обучения монотонного алгоритма классификации, X^L — монотонная выборка, то

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

Свойства этой оценки

- Мощность клина вычисляется за $O(L)$ шагов.
- $Q_c \le 1$ всегда !
- $Q_c = 2/l$ если выборка линейно упорядочена.
- $Q_c = 1$ если точки выборки попарно несравнимы.
- рекомендация: увеличивать мощность клиньев
  (сужать диаметр частичного порядка вблизи границы классов)