Комбинаторные оценки качества обучения по прецедентам

Воронцов К. В.

voron@ccas.ru

Москва
Вычислительный Центр РАН

О статистической теории Вапника-Червоненкиса

- Восстановление зависимости $y^*: X \to Y$
- Распределение $P(X)$ неизвестно
- Семейство алгоритмов A
- Минимизация эмпирического риска на выборке X^l

$$\nu(a, X^l) \to \min_{a \in A}$$

- Функционал равномерного отклонения (оценка при $l = k$)

$$P_\varepsilon(A) = P\left\{\sup_{a \in A}\left(\nu(a, X^k) - \nu(a, X^l)\right) > \varepsilon\right\} \le \Delta^A(L) \cdot 1.5 e^{-\varepsilon^2 l}$$

где $L = l + k$,
X^k — независимая контрольная выборка,
$\Delta^A(L)$ — функция роста

- Метод СМР — структурной минимизации риска

Завышенность статистических оценок

Достаточная длина обучающей выборки:

	Значение функционала = 0.01				Значение функционала = 1.0 (граница применимости теории)			
h \ ε	0.01	**0.05**	0.1	0.2	0.01	**0.05**	0.1	0.2
1	178787	**5778**	1295	286	129512	**3765**	784	155
2	310361	**9636**	2104	448	260832	**7605**	1588	316
5	710759	**21427**	4585	949	661085	**19384**	4065	815
10	1382392	**41247**	8762	1794	1332678	**39202**	8240	1659
100	13525991	**400226**	84504	17137	13476256	**398178**	83981	17002

Следствие завышенности — переупрощение алгоритмов в СМР

Комбинаторные функционалы качества

- Метод обучения $\mu : \{X^l\} \to A$
- $$Q(\mu, X^L) = \nu\left(\mu(X^l), X^k\right)$$

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

$$Q_\varepsilon(\mu, X^L) = \frac{1}{N}\sum_{n=1}^{N} \left[\nu\left(\mu(X_n^l), X_n^k\right) > \varepsilon\right], \qquad \varepsilon \in [0,1]$$

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

- Взаимозаменяемость функционалов

$$\varepsilon Q_\varepsilon \le Q_c \le \varepsilon + Q_\varepsilon$$

$$\varepsilon Q_{\nu,\varepsilon} \le Q_c \le \varepsilon + Q_{\nu,\varepsilon} + \bar{\nu}_L^l$$

Комбинаторная оценка качества обучения

Теорема. $\forall \mu,\ \forall X^L$ $\boxed{Q_{\nu,\varepsilon}(\mu, X^L) \le \Delta_L^l(\mu, X^L) \cdot \Gamma_L^l\left(\varepsilon, \sigma_L^l(\mu, X^L)\right)}$

- Локальная функция роста $\Delta_L^l(\mu, X^L)$

  локальное семейство алгоритмов

$$A_L^l(\mu, X^L) = \left\{ \mu(X_n^l) \middle| n = 1, \ldots, N \right\}$$

- Степень некорректности метода

$$\sigma_L^l(\mu, X^L) = \max_{n=1,\ldots,N} \nu\left(\mu(X_n^l), X_n^l\right)$$

- Комбинаторный множитель

$$\Gamma_L^l\left(\varepsilon, \sigma\right) = \max_{m \in M(\varepsilon,\sigma)} \sum_{s \in S(\varepsilon,\sigma)} \frac{C_m^s C_{L-m}^{l-s}}{C_L^l}$$

$$M(\varepsilon,\sigma) = \{\, m:\ \varepsilon k < m \le k + \sigma l \},$$

$$S(\varepsilon,\sigma) = \{\, s:\ \max(0, m-k) \le s \le \sigma l,\ \ s < (m - \varepsilon k) l / L \}$$

Связь комбинаторного подхода с вероятностным

- Принцип соответствия

$$\mathbf{E}Q_c(\mu, X^L) = \mathbf{P}\left\{\mu(X^l) \text{ ошибается на } x\right\}$$

$$\mathbf{E}Q_\varepsilon(\mu, X^L) = \mathbf{P}\left\{\nu\left(\mu(X^l), X^k\right) > \varepsilon\right\}$$

$$\mathbf{E}Q_{\nu,\varepsilon}(\mu, X^L) = \mathbf{P}\left\{\nu\left(\mu(X^l), X^k\right) - \nu\left(\mu(X^l), X^l\right) > \varepsilon\right\}$$

Из комбинаторных оценок сразу следуют вероятностные:

$$\mathbf{E}Q_{\nu,\varepsilon}(\mu, X^L) \le \mathbf{E}\Delta_L^l(\mu, X^L) \cdot \Gamma_L^l\left(\varepsilon, \sigma_L^l(\mu, X^L)\right) \le$$

$$\le \Delta^A(L) \cdot 1.5e^{-\varepsilon^2 l} \quad (\text{при } l = k)$$

- Нет требования i.i.d.
- Произвольное $P(X) \rightarrow$ произвольная X^L
- Независимость выборки $\rightarrow$ симметричность функционала

Отличия от оценки Вапника-Червоненкиса

- Комбинаторный функционал *точнее* вероятностного

$$\mathbf{E}Q_{\nu,\varepsilon} = \mathbf{P}\left\{\nu(\mu(X_n^l), X^k) - \nu(\mu(X_n^l), X^l) > \varepsilon\right\} \leq$$

$$\leq \mathbf{P}\left\{\sup_{a\in A}\left(\nu(a, X^k) - \nu(a, X^l)\right) > \varepsilon\right\}$$

- Локальная функция роста *точнее* глобальной

(эффект локализации — снимается «запрет на сложность»)

$$\underbrace{\Delta_L^l(\mu, X^L)}_{\leq C_L^l} \leq \underbrace{\Delta^L(A)}_{\leq 2^L}$$

- Комбинаторный множитель *точнее* экспоненциального

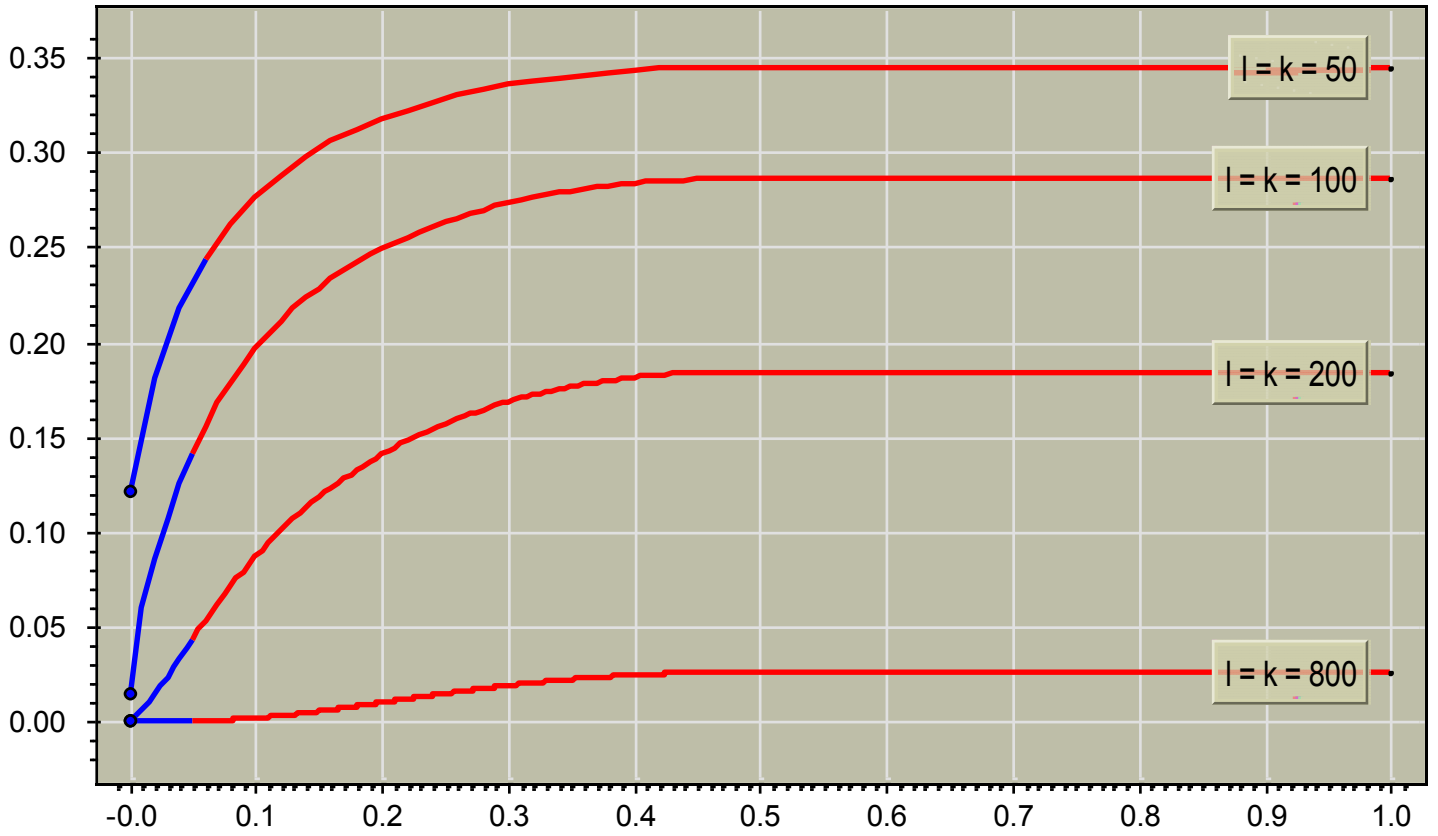
$$\Gamma_L^l(\varepsilon,\sigma) \leq \Gamma_L^l(\varepsilon,1) \leq 1.5\cdot e^{-\varepsilon^2 l} \qquad (\text{при } l = k)$$

Учитывается степень некорректности σ

О важности требования корректности

Зависимость $\Gamma_L^l(\varepsilon,\sigma)$ от σ

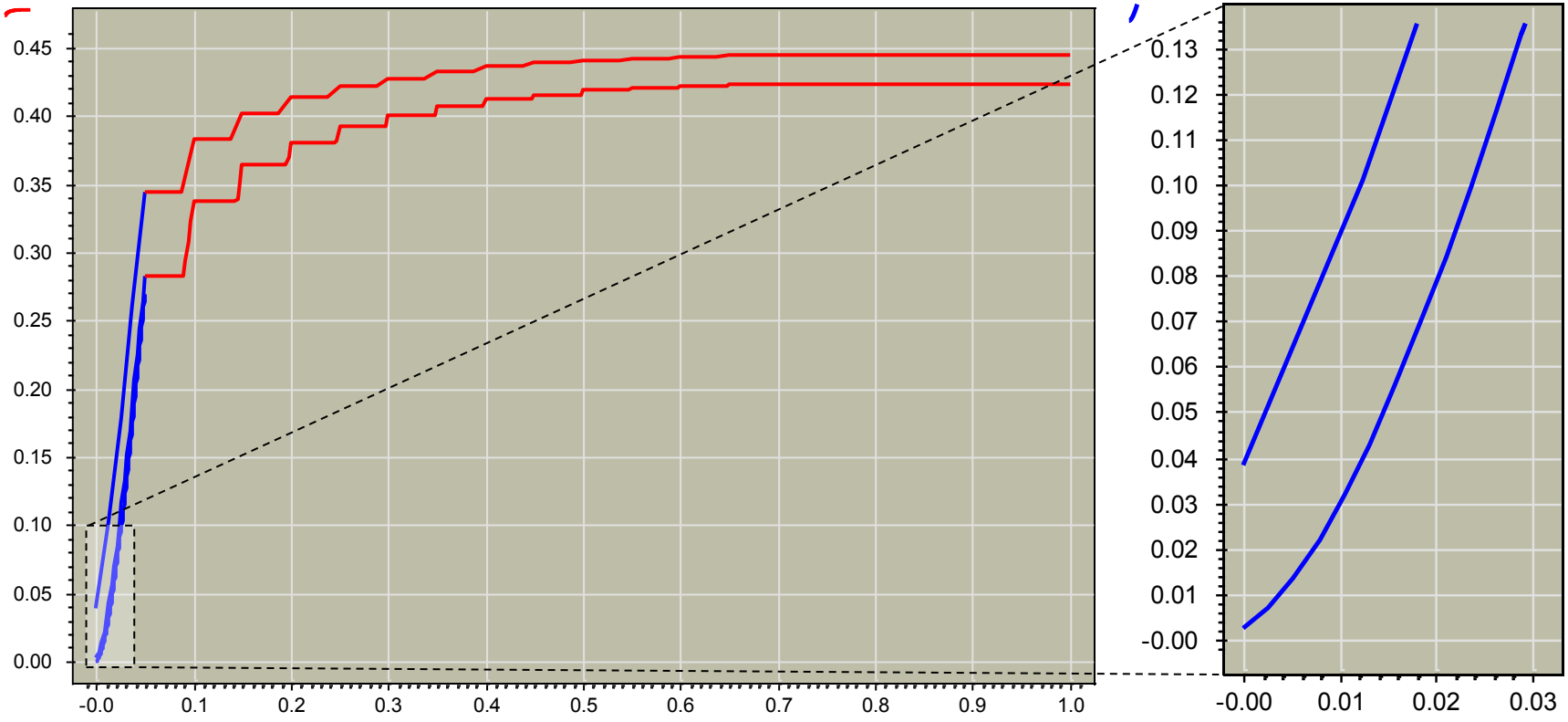
при $\varepsilon = 0.05,\ \ l = k = 50,100,200,800$



О важности требования корректности

Зависимость $\Gamma_L^l(\varepsilon,\sigma)$ от σ

при постоянном $k=20,\ \ \varepsilon=0.05,\ \ L=100,400,1600$



Оценка качества в случае корректности (при $\sigma = 0$)

$$\Gamma_L^l(\varepsilon, 0) = \frac{C_{L-\lceil \varepsilon k \rceil}^l}{C_L^l} \le \left(\frac{k}{L} \right)^{\varepsilon k}$$

Достаточная длина обучающей выборки:

	Значение функционала = 0.01				Значение функционала = 1.0 (граница применимости теории)			
h \ ε	0.01	**0.05**	0.1	0.2	0.01	**0.05**	0.1	0.2
1	1901	**321**	141	66	1101	**181**	71	31
2	3101	**501**	231	96	2401	**361**	151	61
5	6801	**1081**	471	201	6101	**941**	401	161
10	13101	**2061**	901	371	12401	**1921**	821	336
100	107001	**19821**	8561	3521	107001	**19661**	8481	3481

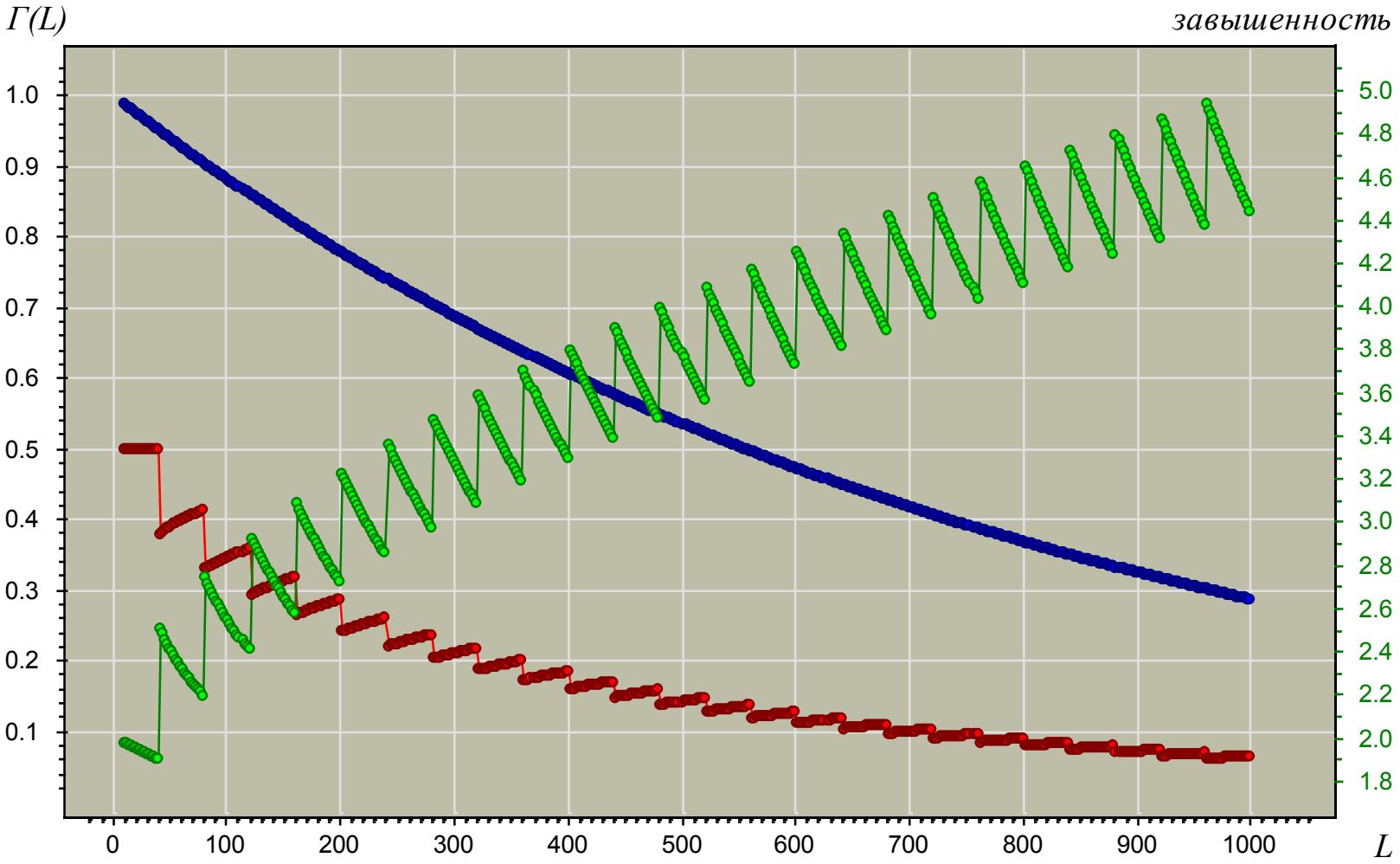
Три причины завышенности сложностных оценок

$$\frac{\Delta^A(L)\cdot 1.5\,e^{-\varepsilon^2 l}}{Q_{\nu,\varepsilon}} = \left(\frac{\Delta^A(L)}{\Delta_L^l}\right)\cdot\left(\frac{1.5\,e^{-\varepsilon^2 l}}{\Gamma_L^l}\right)\cdot\left(\frac{\Delta_L^l\cdot\Gamma_L^l}{Q_{\nu,\varepsilon}}\right)$$

1. Пренебрежение эффектом локализации
2. Экспоненциальная аппроксимация Γ_L^l
3. Погрешность разложения
   (перехода от анализа качества к анализу сложности)

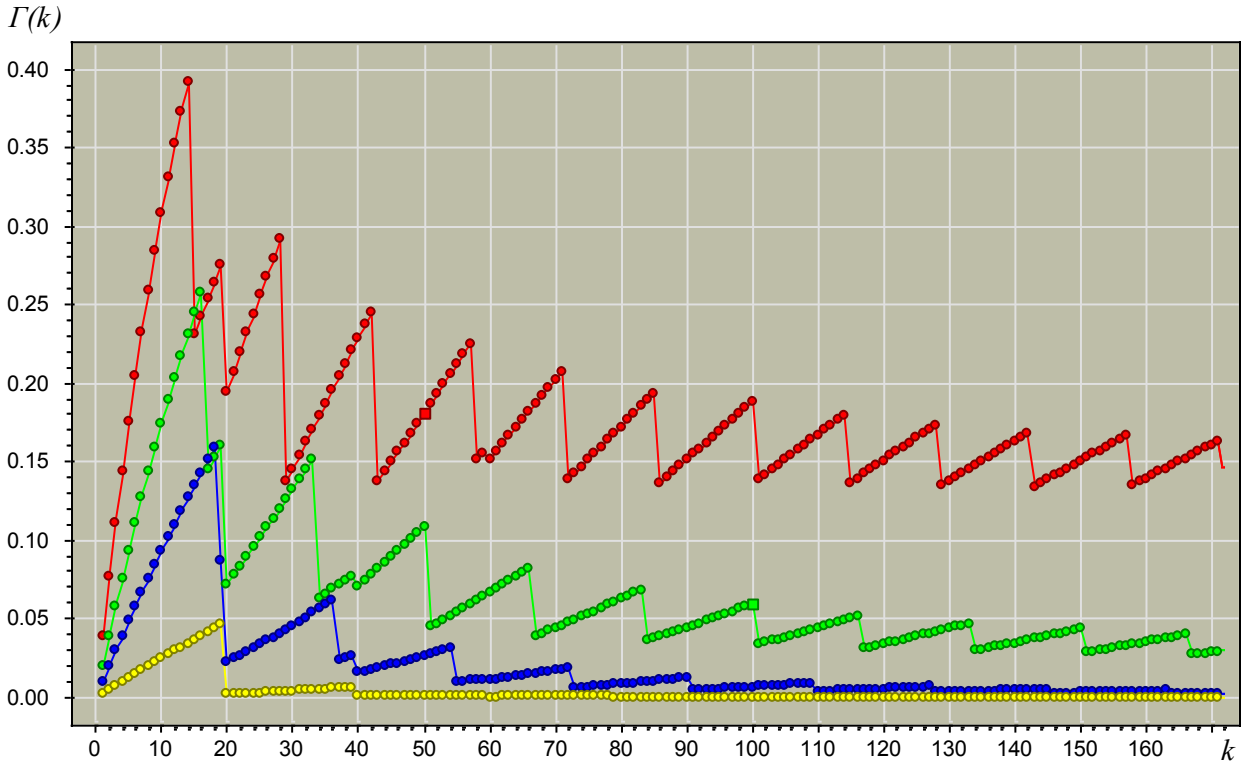
Завышенность экспоненциальной аппроксимации

Зависимость Γ_L^l от L при $\varepsilon = 0.05,\ l = k$



О длине контрольной выборки

- Зависимость Γ_L^l от k при $\varepsilon = 0.05,\ l = k = 50, 100, 200, 800$



- Существует предел Γ_L^l при $k \to \infty$
- Дискретные эффекты при малых k

О методе структурной минимизации риска

Выбор семейства оптимальной сложности
в структуре вложенных семейств возрастающей ёмкости:

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

- С помощью верхней оценки по Вапнику-Червоненкису:

$$\mathbf{P}\left\{\sup_{a \in A}\left(\nu(a, X^k) - \nu(a, X^l)\right) > \varepsilon\right\} < \eta(h, l, \varepsilon)$$

$$a_h \in A_h$$

$$\nu(a_h, X^k) < \nu(a_h, X^l) + \varepsilon(h, l, \eta) \rightarrow \min_h$$

- С помощью скользящего контроля непосредственно:

$$Q(\mu_h, X^L) \rightarrow \min_h$$

(cross-validated model selection)

Об относительном уклонении частот

- Оценка Вапника-Червоненкиса:

$$\mathbf{P}\left\{\sup_{a\in A}\frac{\nu(a,X^k)-\nu(a,X^l)}{\sqrt{\nu(a,X^L)}}>\varepsilon\right\}\ \le\ \Delta^A(L)\cdot\tilde{\Gamma}_L^l(\varepsilon)$$

Является результатом верхней оценки комбинаторного множителя при замене переменной

$$\varepsilon\sqrt{\frac{m}{L}}\to\varepsilon$$

- **Вывод:**
Данная оценка не описывает эффект сужения семейства

Об эффективной ёмкости

- Метод измерения [Вапник, Ботту, Кортес, 1994]:

$$Q_{\sup} = \frac{1}{N}\sum_{n=1}^{N}\left[\sup_{a\in A}\left(\nu(a, X^k) - \nu(a, X^l)\right) > \varepsilon\right] \le C \cdot \frac{L^h}{h!} \cdot e^{-\varepsilon^2 l}$$

sup достигается при $a_n = \mu(\tilde{X}_n^k \cup X_n^l)$.

Учитываются особенности:	ёмкость	эффективная ёмкость	локальная эффективная ёмкость
распределения объектов	–	+	+
самой зависимости	–	–	+
метода обучения	–	–	+

Не-сложностные оценки качества — компактность

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

$x_i : x_{i1}\ x_{i2}\ \ldots\ x_{i,L-1}$ по убыванию $\rho(x_i, x_{im}), \quad i = 1,\ldots,L,$

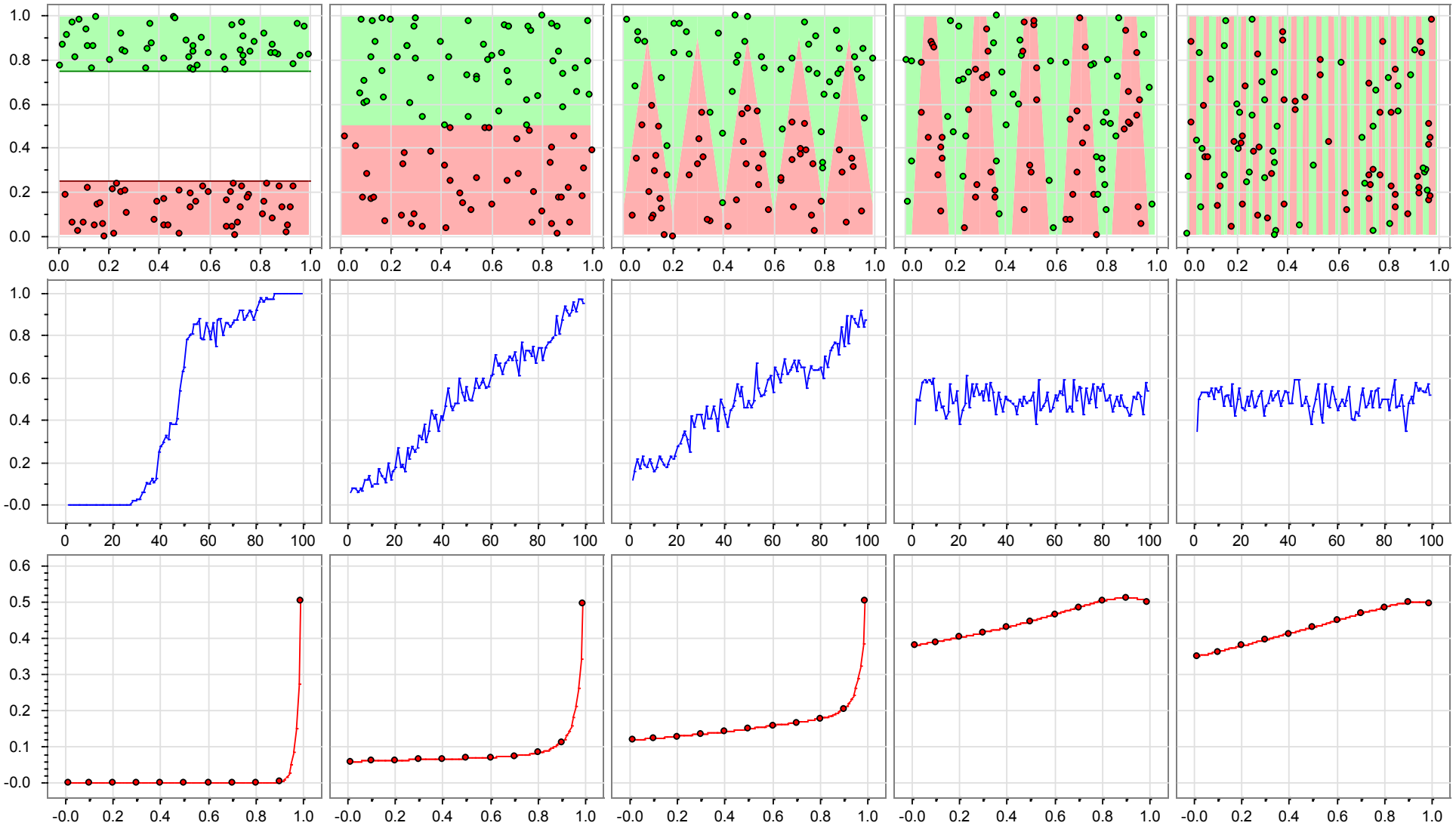
$$r_m(x_i) = I\left(x_i, y^*(x_{im})\right),$$

$$K(m, X^L) = \frac{1}{L}\sum_{i=1}^{L} r_m(x_i), \qquad m = 1,\ldots,L-1.$$

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

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

Профили компактности выборки

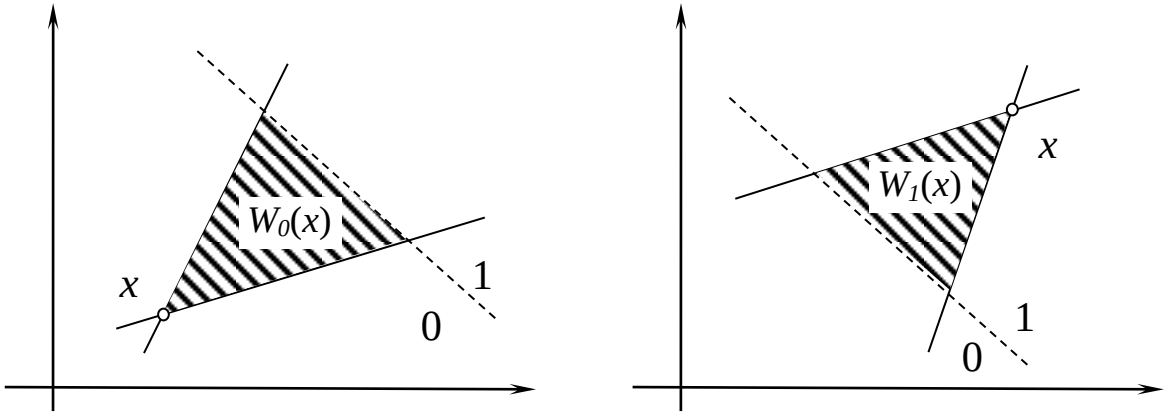


Не-сложностные оценки качества — монотонность

- 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:

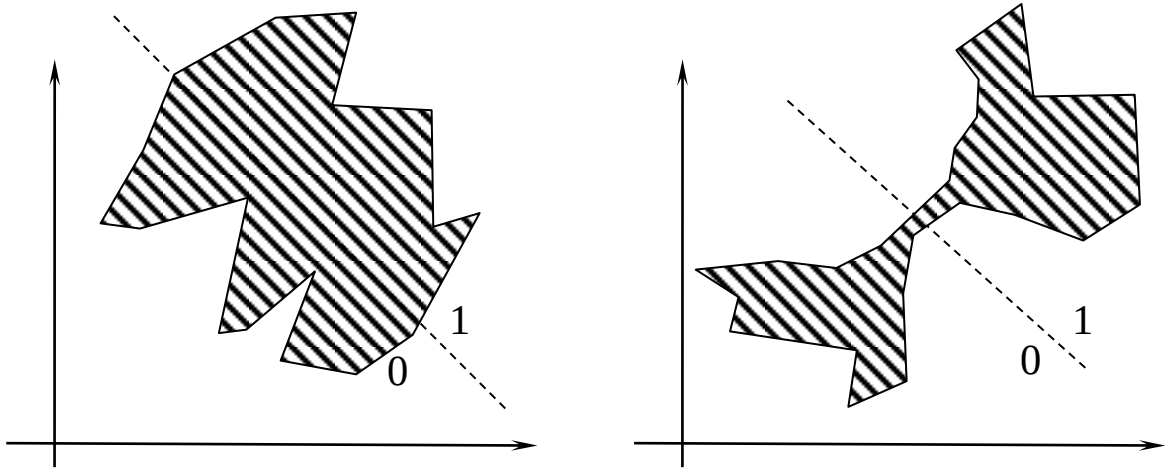
$$M(m, X^L) = \frac{1}{L} \sum_{i=1}^{L} \left[|W_{y_i}(x_i)| = m \right].$$

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)



Некоторые открытые проблемы

- Оценка локальной функции роста для конкретных методов обучения
- Оценка локализующей способности различных методов
- Получение не-сложностных оценок качества