

Комбинаторная теория переобучения

Воронцов Константин Вячеславович

ВЦ РАН • МФТИ

Семинар «Стохастический анализ в задачах» НМУ–МФТИ

13 апреля 2013

- 1 Введение**
 - Обучение по прецедентам и проблема переобучения
 - Комбинаторно-вероятностная постановка задачи
 - Теория Вапника–Червоненкиса
- 2 Метод порождающих и запрещающих множеств**
 - Эксперименты с переобучением
 - Граф расслоения–связности
 - Оценки расслоения–связности
- 3 Разработка методов обучения по прецедентам**
 - Линейные алгоритмы классификации
 - Логические алгоритмы классификации
 - Классификаторы ближайшего соседа

Основные определения и обозначения

\mathbb{X} — объекты; \mathbb{Y} — ответы;

$y^* : \mathbb{X} \rightarrow \mathbb{Y}$ — неизвестная зависимость.

Дано: $x_i = (x_i^1, \dots, x_i^n)$ — обучающие объекты с известными ответами $y_i = y^*(x)$, $i = 1, \dots, \ell$:

$$\begin{pmatrix} x_1^1 & \dots & x_1^n \\ \dots & \dots & \dots \\ x_\ell^1 & \dots & x_\ell^n \end{pmatrix} \xrightarrow{y^*} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}$$

Найти: алгоритм $a : \mathbb{X} \rightarrow \mathbb{Y}$, способный давать правильные ответы на новых объектах $\tilde{x}_i = (\tilde{x}_i^1, \dots, \tilde{x}_i^n)$, $i = 1, \dots, k$:

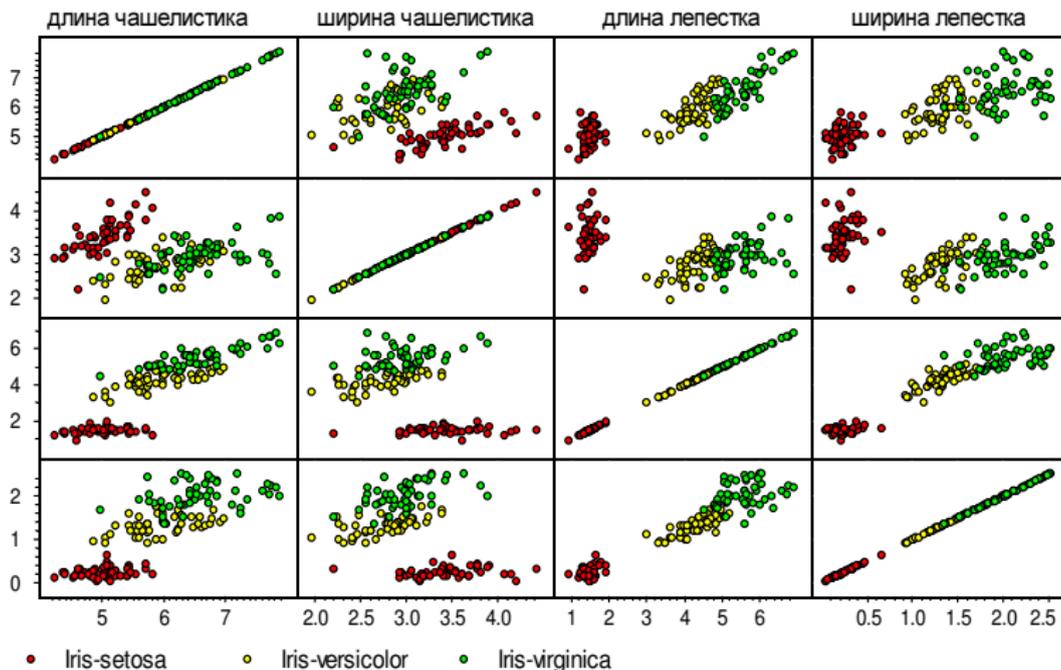
$$\begin{pmatrix} \tilde{x}_1^1 & \dots & \tilde{x}_1^n \\ \dots & \dots & \dots \\ \tilde{x}_k^1 & \dots & \tilde{x}_k^n \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \dots \\ a(\tilde{x}_k) \end{pmatrix}$$

Примеры прикладных задач обучения по прецедентам

- Распознавание, классификация, принятие решений ($|\mathbb{Y}| < \infty$):
 - x — пациент; y — долгосрочный результат лечения;
 - x — заёмщик; y — кредит выдать / не выдать;
 - x — курсы акций; y — купить / продать.
 - x — абонент; y — уйдёт / не уйдёт к другому оператору;
 - x — фотопортрет; y — идентификатор личности;
 - x — фрагмент ДНК; y — функция: промотор / ген;
 - x — фрагмент белка; y — тип вторичной структуры;
 - x — текстовое сообщение; y — спам / не спам;
- Регрессия и прогнозирование ($\mathbb{Y} = \mathbb{R}$ или \mathbb{R}^m):
 - x — структура химического соединения; y — его свойство;
 - x — параметры технолог. процесса; y — свойство продукции;
 - x — история продаж; y — прогноз потребительского спроса;
 - x — данные о недвижимости; y — продажная стоимость;
 - x — пара \langle клиент, товар \rangle ; y — рейтинг товара;
 - x — пара \langle баннер, фраза \rangle ; y — вероятность клика.

Пример: задача классификации цветков ириса [Фишер, 1936]

$n = 4$ признака, $|\mathbb{Y}| = 3$ класса, длина выборки $\ell = 150$.



Модель алгоритмов и метод обучения

Модель алгоритмов — параметрическое семейство отображений

$$A = \{g(x, \theta) \mid \theta \in \Theta\},$$

где $g: \mathbb{X} \times \Theta \rightarrow \mathbb{Y}$ — фиксированная функция,
 Θ — множество допустимых значений параметра θ .

В задачах обучения по прецедентам выделяются два этапа:

- 1 Метод обучения $\mu: (\mathbb{X} \times \mathbb{Y})^\ell \rightarrow A$ по обучающей выборке $X = (x_i, y_i)_{i=1}^\ell$ выбирает из A алгоритм $a = \mu(X)$.
- 2 Найденный алгоритм a применяется для вычисления прогнозов $\tilde{y}_i = a(\tilde{x}_i)$ на новой выборке $\bar{X} = \{\tilde{x}_1, \dots, \tilde{x}_k\}$.

Принцип минимизации эмпирического риска

Эмпирический риск — частота ошибок алгоритма a на X :

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i].$$

Минимизация эмпирического риска — пример метода обучения:

$$\mu(X) = \arg \min_{a \in A} Q(a, X).$$

Проблема обобщающей способности:

- будет ли алгоритм $a = \mu(X)$ приближать y^* на всём \mathbb{X} ?
- найдём ли мы «закон природы» или *переобучимся*, т. е. подгоним функцию $g(x, \theta)$ под заданные точки (x_i, y_i) ?
- будет ли $Q(a, \bar{X})$ мало на новых данных — *контрольной выборке* $\bar{X} = (\tilde{x}_i, \tilde{y}_i)_{i=1}^k$, $\tilde{y}_i = y^*(\tilde{x}_i)$?

Пример переобучения. Модельная задача регрессии

Зависимость $y^*(x) = \frac{1}{1 + 25x^2}$ на отрезке $x \in [-2, 2]$.

Признаковое описание $x \mapsto (1, x^1, x^2, \dots, x^n)$.

Алгоритм полиномиальной регрессии

$$a(x, \theta) = \theta_0 + \theta_1 x + \dots + \theta_n x^n \quad \text{— полином степени } n.$$

Обучение методом наименьших квадратов:

$$Q(\theta, X) = \sum_{i=1}^{\ell} (\theta_0 + \theta_1 x_i + \dots + \theta_n x_i^n - y_i)^2 \rightarrow \min_{\theta_0, \dots, \theta_n}.$$

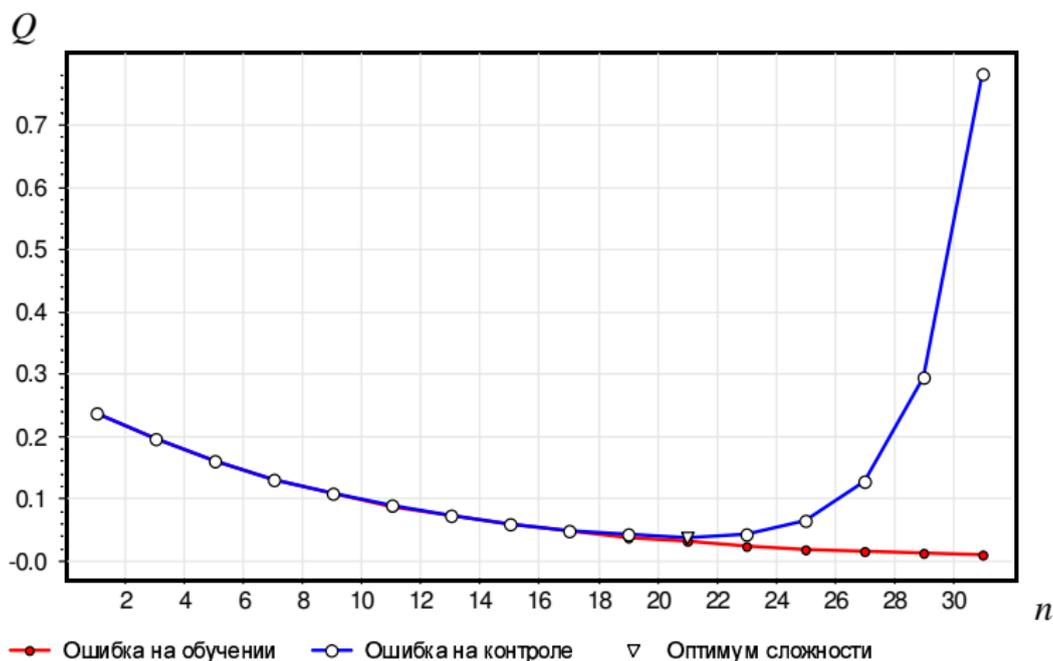
Обучающая выборка: $X = \{x_i = 4 \frac{i-1}{\ell-1} - 2 \mid i = 1, \dots, \ell\}$.

Контрольная выборка: $\bar{X} = \{\tilde{x}_i = 4 \frac{i-0.5}{\ell-1} - 2 \mid i = 1, \dots, \ell-1\}$.

Что происходит с $Q(\mu(X), X)$ и $Q(\mu(X), \bar{X})$ при увеличении n ?

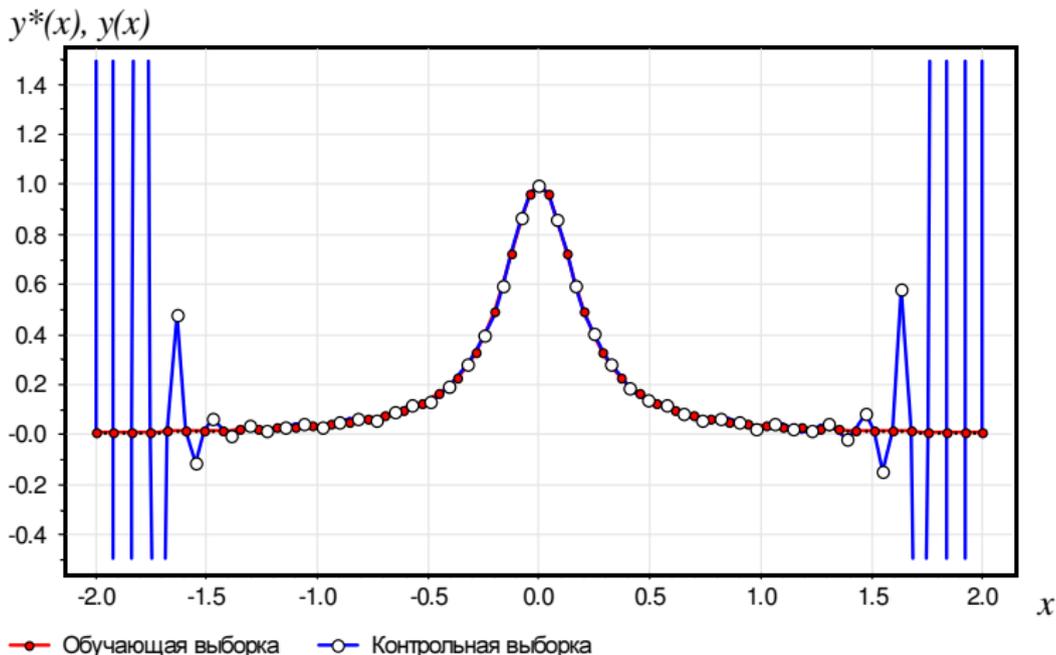
Пример переобучения: эксперимент при $\ell = 50$, $n = 1..31$

Переобучение — это когда $Q(\mu(X), \bar{X}) \gg Q(\mu(X), X)$:



Пример переобучения: эксперимент при $\ell = 50$

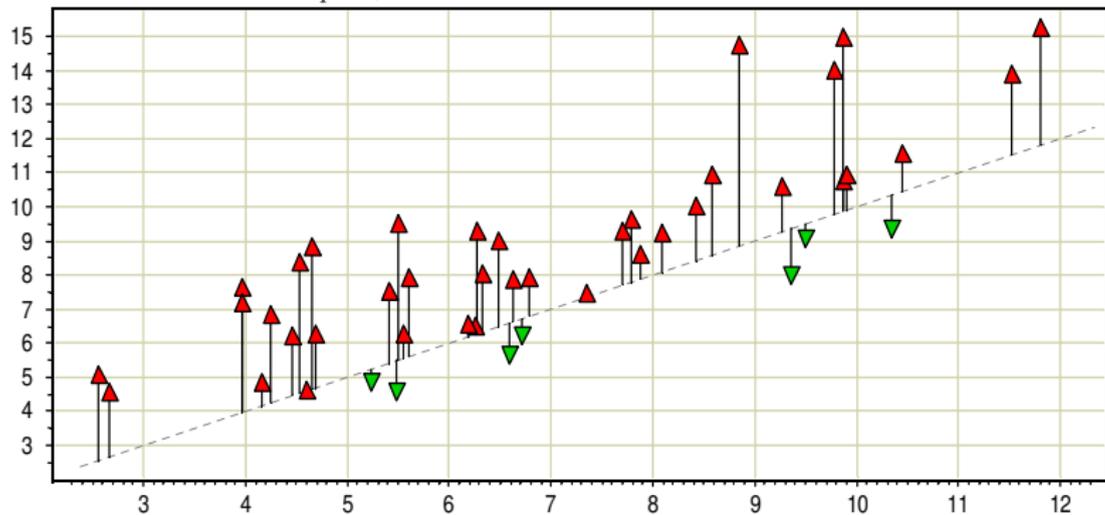
Переобучение, «вид изнутри»: что происходит с полиномами слишком высоких степеней (в данном случае $n = 40$)



Пример переобучения. Реальная задача классификации

Задача предсказания отдалённого результата хирургического лечения атеросклероза, $L = 98$. Точки — различные алгоритмы.

Частота ошибок на контроле, %



Частота ошибок на обучении, %

Задача обучения по прецедентам

$\mathbb{X} = \{x_1, \dots, x_L\}$ — конечное *генеральное множество* объектов;

$A = \{a_1, \dots, a_D\}$ — конечное множество *алгоритмов* (гипотез);

$I(a, x) = [\text{алгоритм } a \text{ ошибается на объекте } x];$

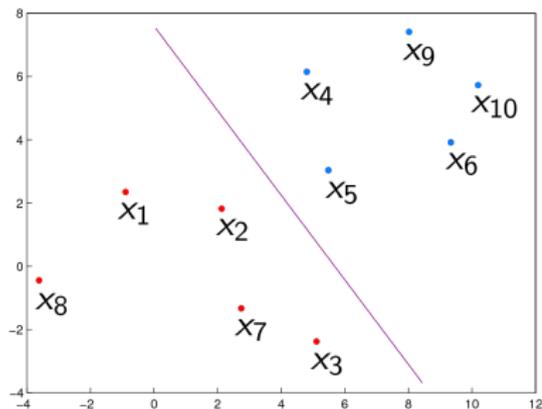
$L \times D$ -матрица ошибок с попарно различными столбцами:

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

$m(a, X) = \sum_{x \in X} I(a, x)$ — число ошибок $a \in A$ на выборке $X \subset \mathbb{X}$;

$\nu(a, X) = \frac{1}{|X|} m(a, X)$ — частота ошибок a на выборке X ;

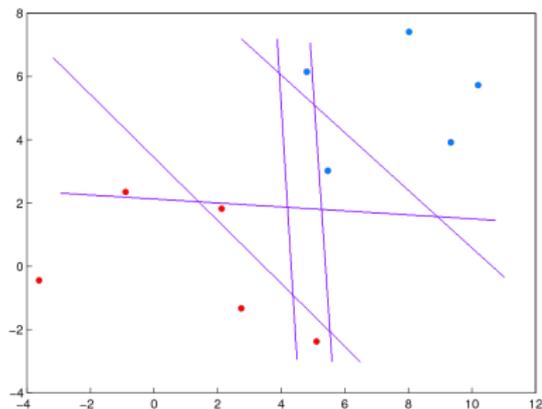
Пример. Матрица ошибок линейных классификаторов



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

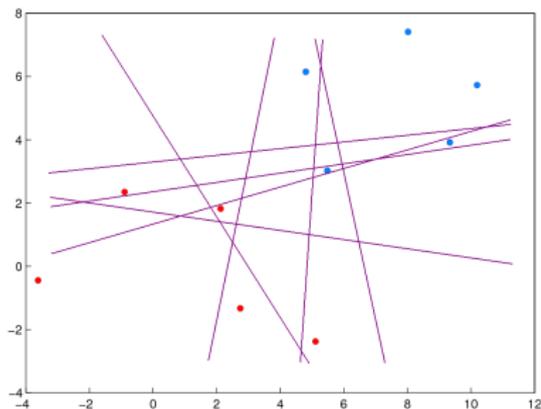
Пример. Матрица ошибок линейных классификаторов



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 обучения по прецедентам

Опр. Метод обучения $\mu: 2^{\mathbb{X}} \rightarrow A$ произвольной выборке $X \subset \mathbb{X}$ ставит в соответствие некоторый алгоритм $a \in A$.

Опр. Метод минимизации эмпирического риска:

$$\mu X = \arg \min_{a \in A} m(a, X).$$

Обучение по прецедентам — это принятие решений на основе неполной информации.

Проблема надёжности обучения по прецедентам:
насколько большим может оказаться $m(\mu X, \bar{X})$?

Единственная вероятностная аксиома:

пусть все разбиения $X \sqcup \bar{X} = \mathbb{X}$ равновероятны,
 X — наблюдаемая обучающая выборка, $|X| = \ell$,
 \bar{X} — скрытая контрольная выборка, $|\bar{X}| = k$.

Обобщающая способность метода обучения

$P \equiv E \equiv \frac{1}{C_L} \sum_{X \subset \mathbb{X}}$ — доля разбиений выборки.

$\delta(\mu, X, \bar{X}) = \nu(\mu X, \bar{X}) - \nu(\mu X, X)$ — переобученность μ на X .

Функционалы обобщающей способности:

- Полный скользящий контроль (Complete Cross-Validation):

$$CCV(\mu, \mathbb{X}) = E \nu(\mu X, \bar{X}).$$

- Ожидаемая переобученность (Expected OverFitting):

$$EOF(\mu, \mathbb{X}) = E \delta(\mu, X, \bar{X}).$$

- Вероятность большой частоты ошибок на контроле:

$$R_\varepsilon(\mu, \mathbb{X}) = P[\nu(\mu X, \bar{X}) \geq \varepsilon].$$

- Вероятность переобучения:

$$Q_\varepsilon(\mu, \mathbb{X}) = P[\delta(\mu, X, \bar{X}) \geq \varepsilon].$$

Отличия от общепринятой постановки задачи

Функционалы обобщающей способности:

- Вероятность равномерного отклонения частоты $\nu(a, X)$ от вероятности ошибки $P(a)$ [Вапник, Червоненкис, 1971]

$$S_\varepsilon(A, \mathbb{X}) = P \left[\sup_{a \in A} (P(a) - \nu(a, X)) \geq \varepsilon \right]$$

- Вероятность переобучения

$$Q_\varepsilon(\mu, \mathbb{X}) = P \left[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geq \varepsilon \right] \ll S_\varepsilon(A, \mathbb{X}).$$

Основные отличия:

- 1 Отказываемся от завышенной оценки \sup по всему A .
- 2 Стараемся учитывать особенности метода обучения μ .
- 3 Отказываемся оценивать ненаблюдаемую величину $P(a)$.
- 4 Отказываемся от лишних вероятностных допущений.

Несколько цитат

А. Н. Колмогоров. Теория информации и теория алгоритмов:

«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.»

В. Д. Голпа. Введение в алгебраическую теорию информации:

«Надобность в вероятностной модели отпадает, поскольку теория информации оказывается достаточно интересной и богатой приложениями в алгебраической постановке. Одним из таких приложений является распознавание образов.»

И ещё пара цитат

Ю. К. Беляев. Вероятностные методы выборочного контроля: «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.»

А. Н. Колмогоров. Теория информации и теория алгоритмов: «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.*»

Две альтернативные интерпретации

Задача комбинаторной теории переобучения:

- 1 Оценивание вероятности переобучения, математического ожидания переобученности или частоты ошибок на контроле в *слабой вероятностной аксиоматике*, предполагающей, что все разбиения $X \sqcup \bar{X} = \mathbb{X}$ равновероятны.
- 2 Вычисление функционалов полного скользящего контроля Q_ε , CCV и др., основанных на усреднении *по всем* разбиениям заданной конечной выборки \mathbb{X} на обучающую X и контрольную \bar{X} подвыборки, *без явного применения метода обучения μ* .

Закон больших чисел в слабой аксиоматике

Пусть $A = \{a\}$ — одноэлементное множество, $m \equiv m(a, \mathbb{X})$.

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

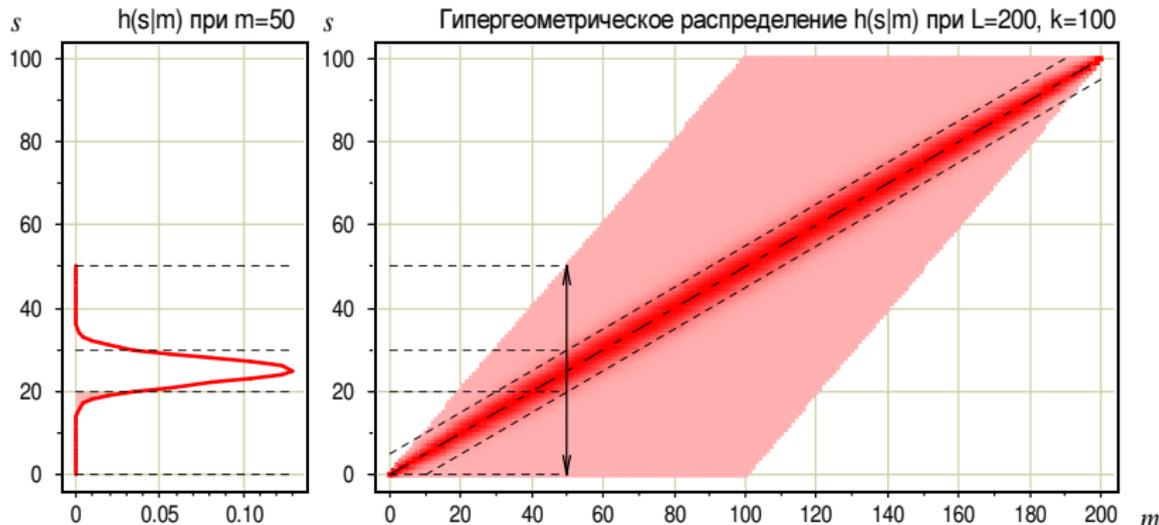
$$Q_\varepsilon(a, \mathbb{X}) = P[\nu(a, \bar{X}) - \nu(a, X) \geq \varepsilon].$$

Теорема

Пусть $A = \{a\}$, $m = m(a, \mathbb{X})$. Для любого \mathbb{X} , любого $\varepsilon \in [0, 1]$

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

где $\mathcal{H}_L^{\ell, m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ — функция гипергеометрического распределения.

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

Предсказание $m = m(a, \mathbb{X})$ по $s = m(a, X)$ возможно благодаря узости гипергеометрического пика (концентрации вероятности).

Закон больших чисел: $\nu(a, X) \rightarrow \nu(a, \bar{X})$ при $\ell, k \rightarrow \infty$.

Теория Вапника–Червоненкиса в слабой аксиоматике

Теорема

Для любых \mathbb{X} , μ , A и $\varepsilon \in (0, 1)$

$$\begin{aligned}
 Q_\varepsilon(\mu, \mathbb{X}) &\stackrel{\text{uniform bound}}{\leq} P \max_{a \in A} [\delta(a, X, \bar{X}) \geq \varepsilon] \stackrel{\text{union bound}}{\leq} \sum_{a \in A} Q_\varepsilon(a, \mathbb{X}) \\
 &\leq |A| \cdot \max_a Q_\varepsilon(a, \mathbb{X}) \leq \\
 &\leq |A| \cdot \max_m \mathcal{H}_L^{\ell, m} \left(\frac{\ell}{L} (m - \varepsilon k) \right) \leq \\
 &\leq |A| \cdot \frac{3}{2} \exp(-\varepsilon^2 \ell), \quad \text{при } \ell = k.
 \end{aligned}$$

Обычно эта оценка завышена в 10^8 – 10^{11} раз. Почему?

- 1) uniform bound завышена, когда A *расслаивается* по $m(a, \mathbb{X})$
- 2) union bound завышена, когда в A много *схожих* алгоритмов

Два мысленных эксперимента

1. Пусть в семействе есть один очень хороший алгоритм, $m(a_0, \mathbb{X}) = 0$, и много плохих алгоритмов a : $m(a, \mathbb{X}) \gg 0$. Тогда a_0 почти всегда будет лучшим и на обучающей выборке.
Результат: можно полагать $|A| \approx 1$.

В общем случае **надо учитывать расслоение семейства A** , $A_m = \{a \in A: m(a, \mathbb{X}) = m\}$, наиболее важны нижние слои.

2. Пусть в семействе есть алгоритм a_0 , и все остальные очень похожи на него. Тогда это «почти один и тот же алгоритм».
Результат: можно полагать $|A| \approx 1$.

В общем случае **надо учитывать связность семейства A** , сколько в A вместе с каждым a содержится b : $\|a - b\| = 1$.

Эволюция подходов в теории статистического обучения

- Uniform convergence bounds [Vapnik, Chervonenkis, 1968]
- Theory of learnable (PAC-learning) [Valiant, 1982]
- Data-dependent bounds [Haussler, 1992]
- Concentration inequalities [Talagrand, 1995]
- Connected function classes [Sill, 1995]
- Similar classifiers VC bounds [Bax, 1997]
- Margin based bounds [Bartlett, 1998]
- Self-bounding learning algorithms [Freund, 1998]
- Rademacher complexity [Koltchinskii, 1998]
- Adaptive microchoice bounds [Langford, Blum, 2001]
- Algorithmic stability [Bousquet, Elisseeff, 2002]
- Algorithmic luckiness [Herbrich, Williamson, 2002]
- Shell bounds [Langford, 2002]
- Localized complexities [Bartlett, Mendelson, Philips, 2004]
- PAC-Bayes bounds [McAllester, 1999; Langford, 2005]
- Oracle inequalities [Koltchinskii, 2011]

Эксперименты с переобучением

Физика — экспериментальная, естественная наука, часть естествознания. Математика — это та часть физики, в которой эксперименты дешёвы. [В.И.Арнольд]

Будем оценивать вероятность методом Монте–Карло — как долю разбиений выборки из случайного подмножества N разбиений, $|N|$ порядка 10^3 – 10^4 :

$$\hat{Q}_\varepsilon(\mu, \mathbb{X}) = \frac{1}{|N|} \sum_{(\bar{X}, X) \in N} [\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geq \varepsilon].$$

Эксперимент №1: измерение факторов завышенности на реальных задачах классификации.

Эксперимент №2: исследование модельных семейств, задаваемых непосредственно матрицами ошибок.

Эксперимент №1. Измерение факторов завышенности VC-оценки

Степень завышенности раскладывается в произведение факторов:

$$\hat{Q}_\varepsilon(\mu, \mathbb{X}) \cdot r_1 \cdot r_2 \cdot r_3 = |A| \cdot \frac{3}{2} \exp(-\varepsilon^2 \ell)$$

Факторы завышенности VC-оценки:

- $r_1 \geq 1$: расслоение (принцип равномерной сходимости)
- $r_2 \geq 1$: сходство (применение неравенства Буля)
- $r_3 \geq 1$: экспоненциальная аппроксимация ГГР

Эффективный локальный коэффициент разнообразия:

такое значение $\Delta = |A|$, при котором оценка была бы точной:

$$\hat{Q}_\varepsilon(\mu, \mathbb{X}) = \Delta \cdot \frac{3}{2} \exp(-\varepsilon^2 \ell)$$

Эксперимент №1. Измерение факторов завышенности VC-оценки

- 7 задач классификации на два класса (из репозитория UCI)
- 20×2 -кратный скользящий контроль, $\ell = k$
- Логический алгоритм классификации Forecsys LogicPro[®]
[Воронцов, Кочедыков, Ивахненко]

Задача	L	n	средняя ошибка на тестовых данных				
			C4.5	C5.0	RIPPER	SLIPPER	LogicPro
crx	690	15	15.5	14.0	15.2	15.7	14.3 ± 0.2
german	1000	20	27.0	28.3	28.7	27.2	28.5 ± 1.0
hepatitis	155	19	18.8	20.1	23.2	17.4	16.7 ± 1.7
horse-colic	300	25	16.0	15.3	16.3	15.0	16.4 ± 0.5
hypothyroid	3163	25	0.4	0.4	0.9	0.7	0.8 ± 0.04
liver	345	6	37.5	31.9	31.3	32.2	29.2 ± 1.6
promoters	106	57	18.1	22.7	19.0	18.9	12.0 ± 2.0

L — объём полной выборки; n — число признаков.

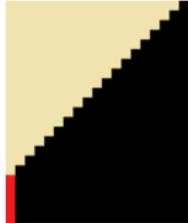
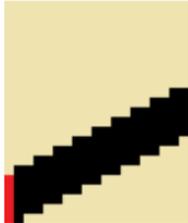
Эксперимент №1. Измерение факторов завышенности VC-оценки

Факторы завышенности: r_1 – расслоение, r_2 – сходство, r_3 – exp.
(пороги $\varepsilon_0, \varepsilon_1, \varepsilon_2$ соответствуют надёжности $\hat{Q}_\varepsilon = 0.05, 0.1, 0.01$).

Задача	класс y	r_1	r_2	r_3	$\Delta[\varepsilon_1, \varepsilon_2]$	$\Delta(\varepsilon_0)$
crx	0	2 759	680	32.6	[10; 41]	24
	1	1 104	1700	11.6	[11; 180]	12
german	1	15 215	1500	10.9	[38; 530]	54
	2	44 400	9000	9.9	[1.0; 2.2]	1.9
hepatitis	0	308	280	9.5	[11; 148]	83
	1	132	680	22.5	[12; 27]	15
horse-colic	1	151	4500	7.2	[2; 9]	7
	2	504	3400	7.3	[3; 6]	6
hypothyroid	0	1 964 200	400	16.5	[3; 220]	21
	1	581 400	460	28.7	[2; 44]	30
promoters	0	555	340	9.8	[36; 230]	72
	1	510	790	6.9	[9; 22]	18

Эксперимент №2. Сравнение четырёх модельных семейств

Матрицы ошибок: строки — объекты, столбцы — алгоритмы;
лучший алгоритм одинаков во всех четырёх семействах.

	есть расслоение по числу ошибок	нет расслоения по числу ошибок
есть связность, соседние алгоритмы отличаются на одном объекте, образуется <i>цепь</i>		
нет связности, соседние алгоритмы существенно различны, <i>цепь</i> не образуется		

Результаты эксперимента (при $\ell = k = 100$, $\varepsilon = 0.05$, $|N| = 10^4$)

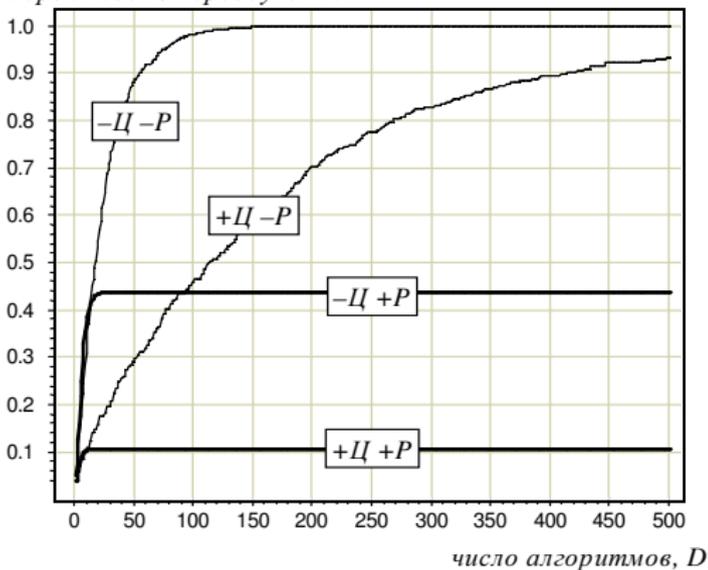
Условные обозначения:

- +Ц — цепь;
- Ц — не цепь;
- +Р — с расслоением;
- Р — без расслоения;

Связность замедляет темп роста $Q_\varepsilon(D)$

Расслоение понижает уровень горизонтальной асимптоты $Q_\varepsilon(D)$

Вероятность переобучения



Вывод: получение точных оценок вероятности переобучения невозможно без учёта эффектов расслоения и связности.

Граф расслоения–связности множества алгоритмов

Определим бинарные отношения на множестве алгоритмов A :
 частичный порядок $a \leq b$: $I(a, x) \leq I(b, x)$ для всех $x \in \mathbb{X}$;
 предшествование $a \prec b$: $a \leq b$ и $\|b - a\| = 1$.

Определение

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

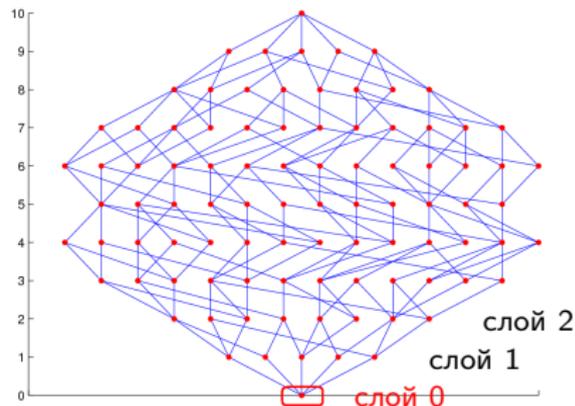
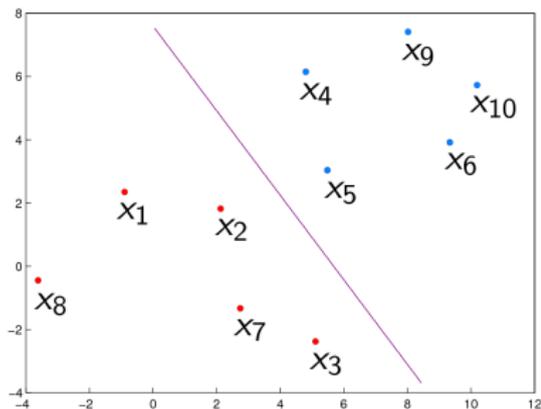
A — множество попарно различных векторов ошибок;

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

Свойства графа расслоения–связности:

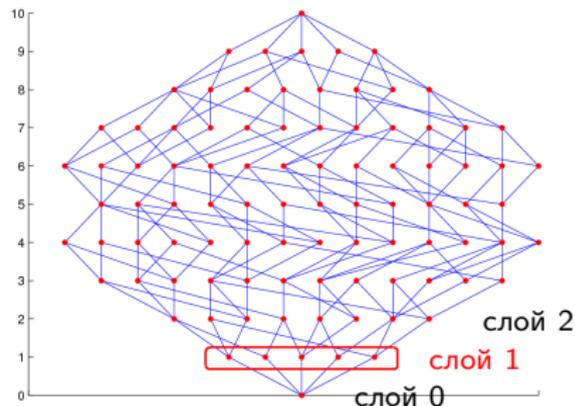
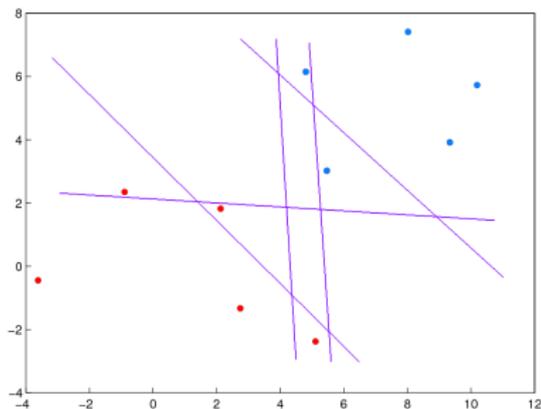
- это подграф диаграммы Хассе отношения порядка \leq на A ;
- каждому ребру (a, b) соответствует объект $x_{ab} \in \mathbb{X}$, такой, что $I(a, x_{ab}) = 0$, $I(b, x_{ab}) = 1$;
- граф является многодольным со слоями
 $A_m = \{a \in A : m(a, \mathbb{X}) = m\}$, $m = 0, \dots, L$;

Пример. Семейство линейных алгоритмов классификации



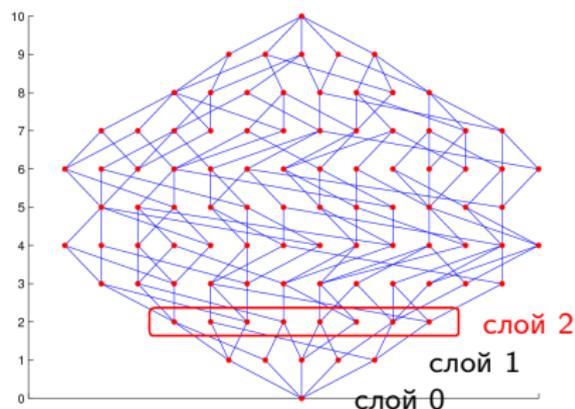
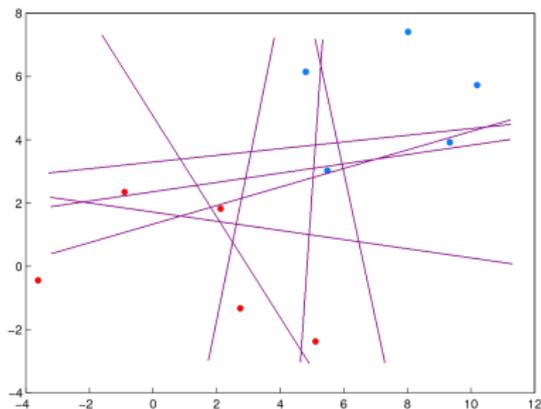
	слой 0
x ₁	0
x ₂	0
x ₃	0
x ₄	0
x ₅	0
x ₆	0
x ₇	0
x ₈	0
x ₉	0
x ₁₀	0

Пример. 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. Семейство линейных алгоритмов классификации



	слой 0	слой 1						слой 2								
X ₁	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	...
X ₂	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	...
X ₃	0	0	0	1	0	0	0	0	1	1	0	0	0	0	1	...
X ₄	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	...
X ₅	0	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	0	1	0	1	0	...
X ₇	0	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	0	...
X ₉	0	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	0	...

Порождающее и запрещающее множества алгоритма

Определение

Верхняя связность $u(a)$ алгоритма a — это число всех рёбер, исходящих из вершины a :

$$u(a) = |X_a|, \quad X_a = \{x_{ab} \in \mathbb{X} \mid a \prec b\};$$

X_a называется *порождающим множеством* алгоритма a .

Определение

Неполноценность $q(a)$ алгоритма a — это число различных объектов, соответствующих всем рёбрам на путях, ведущих в a :

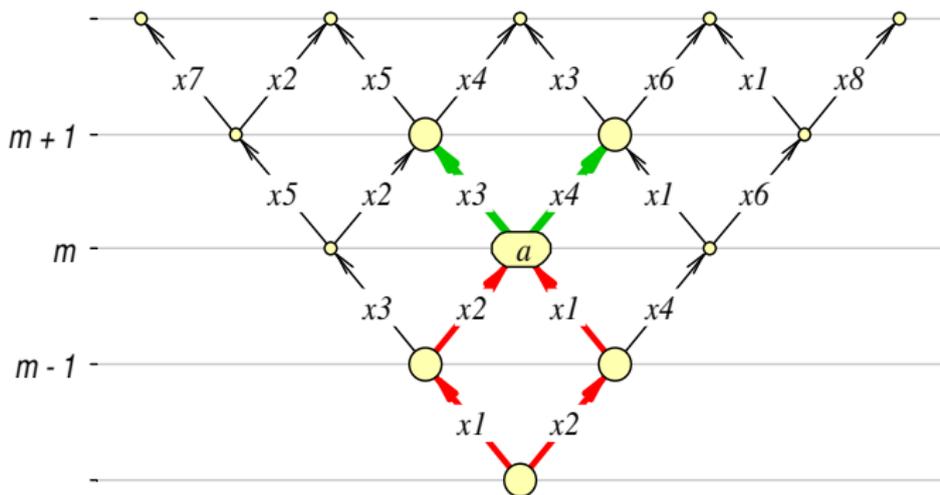
$$q(a) = |X'_a|, \quad X'_a = \{x \in \mathbb{X} \mid \exists b \in A: b \prec a, I(b, x) < I(a, x)\};$$

X'_a называется *запрещающим множеством* алгоритма a .

Пример: двумерная сеть алгоритмов

Верхняя связность алгоритма a : $X_a = \{x3, x4\}$, $u(a) = |X_a| = 2$;

Неполноценность алгоритма a : $X'_a = \{x1, x2\}$, $q(a) = |X'_a| = 2$;



Основная лемма: если $\mu X = a$, то $X_a \subseteq X$ и $X'_a \subseteq \bar{X}$.

Верхняя оценка вероятности переобучения

Теорема (Воронцов, Решетняк, Ивахненко, 2010)

Для любого монотонного метода μ , любых \mathbb{X} , A и $\varepsilon \in (0, 1)$

$$Q_\varepsilon(\mu, \mathbb{X}) \leq \sum_{a \in A} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \mathcal{H}_{L-u-q}^{\ell-u, m-q} \left(\frac{\ell}{L} (m - \varepsilon k) \right),$$

где $u = |X_a|$ — верхняя связность алгоритма a ,

$q = |X'_a|$ — неполноценность алгоритма a ,

$m = m(a, \mathbb{X})$ — число ошибок алгоритма a ,

$$\mathcal{H}_L^{\ell, m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}, \quad z = 0, \dots, \ell$$

— функция гипергеометрического распределения:

Следствие: $P[\mu X = a] \leq C_{L-u-q}^{\ell-u} / C_L^\ell$.

Идея доказательства

Опр. Метод обучения μ называется *монотонным*, если

$$\mu(X) \in A(X) = \underset{a \in A}{\text{Arg min}} K(a, X),$$

где $K(a, X)$ — строго монотонная функция вектора ошибок a :

$$\forall X \subset \mathbb{X}, \forall a, b \in A \text{ если } a < b, \text{ то } K(a, X) < K(b, X).$$

Опр. Метод μ называется *пессимистичным*, если

$$\mu(X) = \underset{a \in A(X)}{\text{arg max}} \delta(a, X).$$

Основная лемма

Если метод обучения μ монотонный и пессимистичный, то

$$[\mu X = a] \leq [X_a \subseteq X] [X'_a \subseteq \bar{X}].$$

Идея доказательства

1. Пусть μ — произвольный монотонный метод обучения, $\bar{\mu}$ — монотонный пессимистичный метод обучения. Тогда

$$Q_\varepsilon(\mu, \mathbb{X}) \leq Q_\varepsilon(\bar{\mu}, \mathbb{X}).$$

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

$$3. P[\bar{\mu}(X) = a] \leq P[\underbrace{X_a \subseteq X \text{ и } X'_a \subseteq \bar{X}}_{S(a, X)}] = \frac{C_{L-|X_a|-|X'_a|}^{\ell-u}}{C_L^\ell} = \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell}.$$

4. По формуле полной вероятности:

$$Q_\varepsilon(\bar{\mu}, \mathbb{X}) \leq \sum_{a \in A} \underbrace{P[S(a, X)]}_{C_{L-u-q}^{\ell-u} / C_L^\ell} \cdot \underbrace{P[\delta(a, X) \geq \varepsilon \mid S(a, X)]}_{\mathcal{H}_{L-u-q}^{\ell-u, m-q} \left(\frac{\ell}{L} (m - \varepsilon k) \right)}. \quad \blacksquare$$

Свойства оценки

$$Q_\varepsilon \leq \sum_{a \in A} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \mathcal{H}_{L-u-q}^{\ell-u, m-q} \left(\frac{\ell}{L} (m - \varepsilon k) \right)$$

- 1 Вклад алгоритма $a \in A$ убывает экспоненциально по $u(a) \Rightarrow$ **связные семейства меньше переобучаются**; по $q(a) \Rightarrow$ **только нижние слои значимы для Q_ε** .
- 2 При $|A| = 1$ это оценка скорости сходимости частот в двух выборках (вариант закона больших чисел):

$$Q_\varepsilon = \mathcal{H}_L^{\ell, m} \left(\frac{\ell}{L} (m - \varepsilon k) \right) \rightarrow 0 \text{ при } \ell, k \rightarrow \infty.$$

- 3 При $q = u = 0$ и $\ell = k$ это оценка Вапника-Червоненкиса:

$$Q_\varepsilon \leq \sum_{a \in A} \mathcal{H}_L^{\ell, m} \left(\frac{\ell}{L} (m - \varepsilon k) \right) \leq |A| \cdot \exp(-\varepsilon \ell^2).$$

- 4 Оценка обращается в равенство в случае многомерных монотонных сетей алгоритмов [**Павел Ботов**].
- 5 Имеется критерий точности оценки [**Никита Животовский**]

Свойства оценки

$$Q_\varepsilon \leq \sum_{a \in A} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \mathcal{H}_{L-u-q}^{\ell-u, m-q} \left(\frac{\ell}{L} (m - \varepsilon k) \right)$$

- 6 При замене неполноценности q на нижнюю связность d это верхняя оценка функционала равномерного отклонения

$$S_\varepsilon(A, \mathbb{X}) = P \left[\sup_{a \in A} (\nu(a, \bar{X}) - \nu(a, X)) \geq \varepsilon \right],$$

которая учитывает связность, но не учитывает расслоение.

- 7 Вероятность получить алгоритм в результате обучения

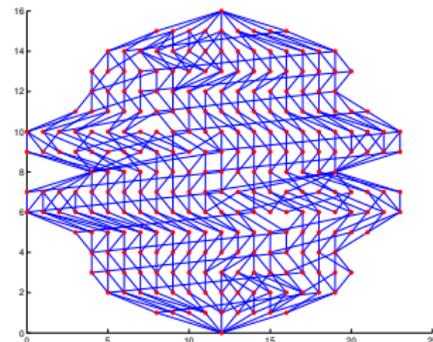
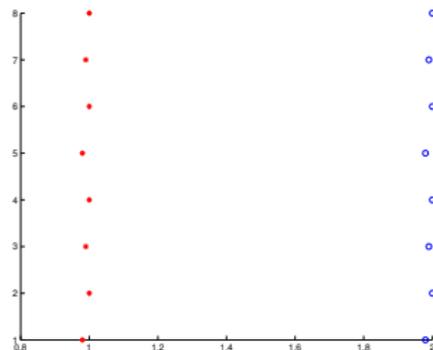
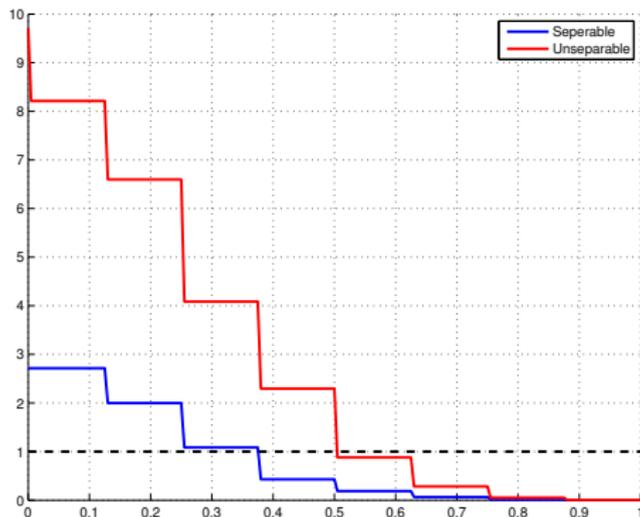
$$P[\mu X = a] \leq P_a = C_{L-u-q}^{\ell-u} / C_L^\ell.$$

- 8 Если $q(a) > k$, то $P_a = 0$ и вклад алгоритма a равен 0
 \Rightarrow при малой длине контроля k оценка вырождается;
 $\Rightarrow k$ надо брать не меньше числа значимых нижних слоёв.

- 9 $\sum_{a \in A} P_a$ — оценка степени завышенности.

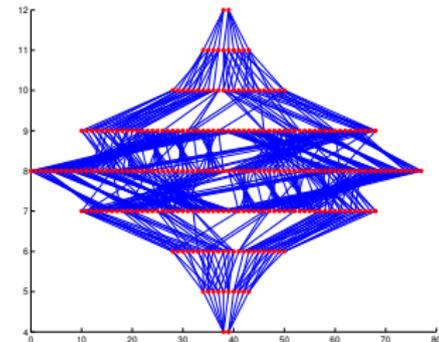
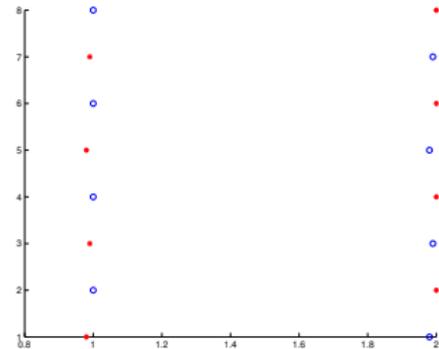
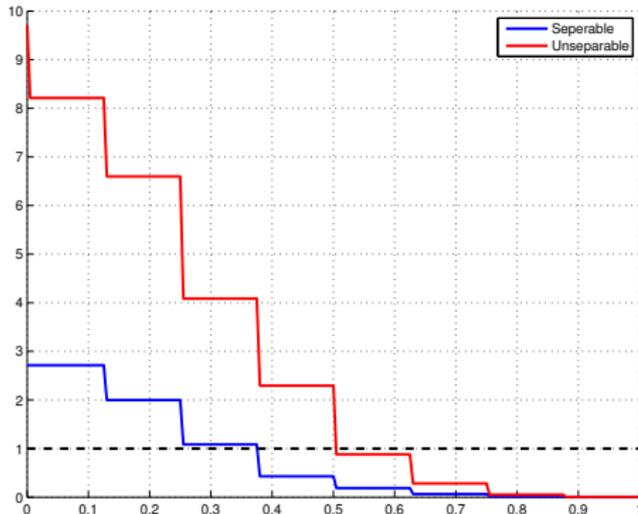
Эксперимент на модельных данных

А — линейные классификаторы,
линейно разделимая выборка



Эксперимент на модельных данных

А — линейные классификаторы,
линейно неразделимая выборка



Верхние оценки средней частоты ошибок на контроле

Теорема

Для любого монотонного метода μ , любых \mathbb{X} и A

$$\text{CCV}(\mu, \mathbb{X}) \leq \sum_{a \in A} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \left(\frac{m}{k} - \frac{(m-q)(\ell-u)}{k(L-u-q)} \right).$$

где $u = |X_a|$ — верхняя связность алгоритма a ,

$q = |X'_a|$ — неполноценность алгоритма a ,

$m = m(a, \mathbb{X})$ — число ошибок алгоритма a .

Преимущество:

оценка CCV вычисляется намного проще, чем оценки Q_ε и R_ε .

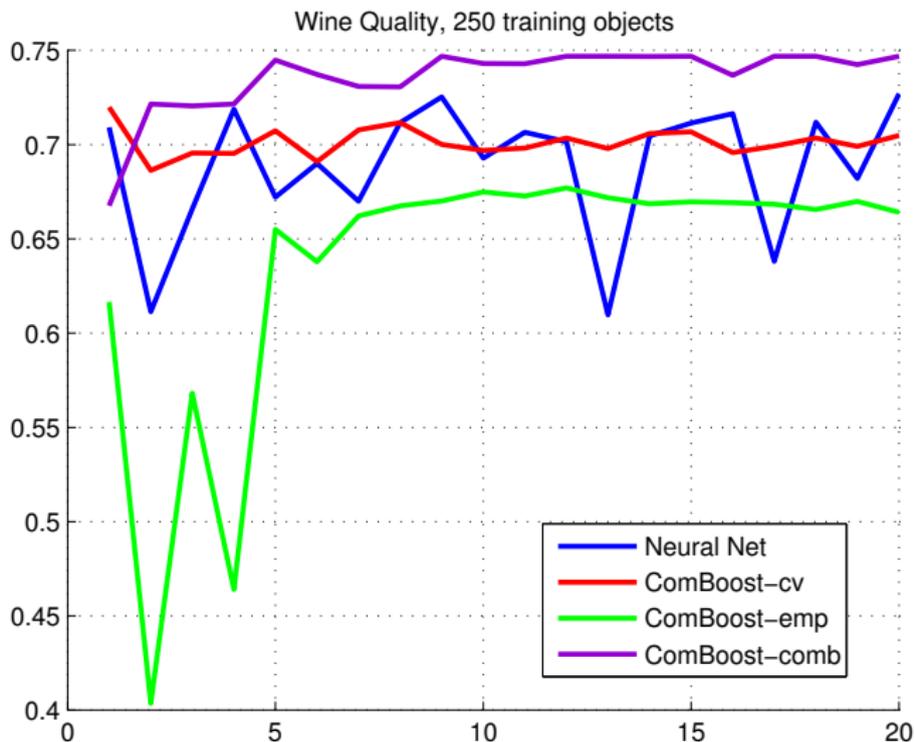
Построение композиций линейных классификаторов

- 1 Получение более точной оценки, учитывающей несколько истоков графа РС
- 2 Приближённое вычисление оценки методом случайных блужданий по графу РС
- 3 Построение линейных классификаторов низкой размерности по подвыборкам
- 4 Их объединение в линейную композицию методом комитетного бустинга
- 5 Эксперименты на реальных задачах из UCI MLR

Результат: сильно разреженная двухслойная нейронная сеть с низкой сложностью и высокой обобщающей способностью.

[Евгений Соколов]

Зависимость ошибки от длины композиции



Понятие логической закономерности

Задача классификации: $y_i = y^*(x_i)$ — класс объекта x_i .

Закономерность класса y — это предикат $r: \mathbb{X} \rightarrow \{0, 1\}$, выделяющий ($r(x) = 1$) как можно больше объектов класса y :

$$p(r, X) = \sum_{x_i \in X} r(x_i) [y_i = y] \rightarrow \max,$$

и как можно меньше объектов всех остальных классов:

$$n(r, X) = \sum_{x_i \in X} r(x_i) [y_i \neq y] \rightarrow \min.$$

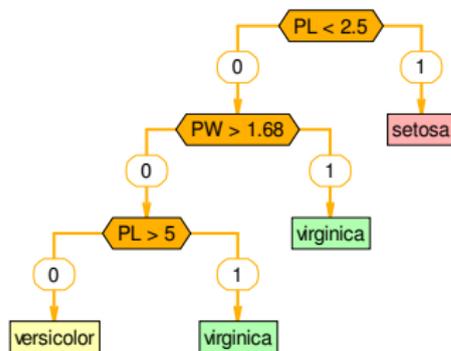
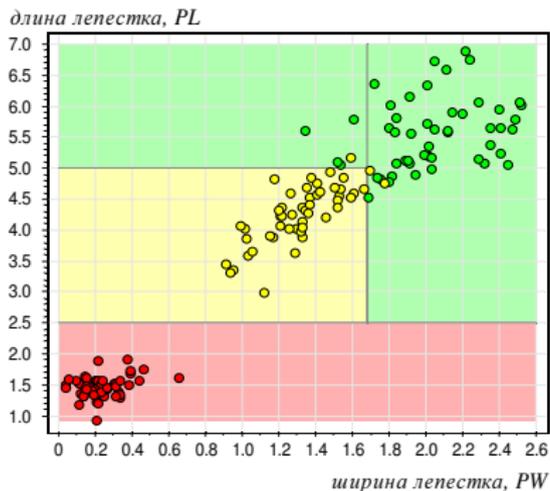
Логическая закономерность — конъюнкция пороговых условий:

$$r(x) = \bigwedge_{j \in J} [f_j(x) \leq \theta_j],$$

где $f_j(x)$ — числовые признаки, θ_j — пороги, $j = 1, \dots, n$;

$J \subseteq \{1, \dots, n\}$ — подмножество признаков, обычно $|J| = 1..5$.

Пример. Закономерности в задаче с ирисами Фишера



setosa	$r_1(x) = [PL \leq 2.5]$
virginica	$r_2(x) = [PL > 2.5] \wedge [PW > 1.68]$
virginica	$r_3(x) = [PL > 5] \wedge [PW \leq 1.68]$
versicolor	$r_4(x) = [PL > 2.5] \wedge [PL \leq 5] \wedge [PW < 1.68]$

Критерии информативности для поиска закономерностей

$$\mathcal{H}(p, n) \rightarrow \max$$

- точность: $\mathcal{H}(p, n) = (p + N - n)/(P + N)$;
- взвешенная точность: $\mathcal{H}(p, n) = p - \lambda n$;
- энтропийный критерий информационного выигрыша:

$$\mathcal{H}(p, n) = h\left(\frac{P}{\ell}\right) - \frac{p+n}{\ell} h\left(\frac{p}{p+n}\right) - \frac{\ell-p-n}{\ell} h\left(\frac{P-p}{\ell-p-n}\right),$$

где $h(q) = -q \log_2 q - (1 - q) \log_2 (1 - q)$;

- индекс Джини (Gini impurity): то же, при $h(q) = 2q(1 - q)$;
- точный тест Фишера: $\mathcal{H}(p, n) = -\log C_P^p C_N^n / C_{P+N}^{p+n}$;
- критерий бустинга: $\mathcal{H}(p, n) = \sqrt{p} - \sqrt{n}$;

$P = |\{x_i: y_i = y\}|$ — число «своих» в обучающей выборке;

$N = |\{x_i: y_i \neq y\}|$ — число «чужих» в обучающей выборке.

Поиск и композиция логических закономерностей

Поиск закономерностей:

- 1: **пока** не построено множество закономерностей R
- 2: из множества правила-кандидатов R породить R' ;
- 3: выкинуть из R' неинформативные правила;
- 4: $R := R'$;

Композиция — взвешенное голосование:

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{r \in R_y} w_r r(x),$$

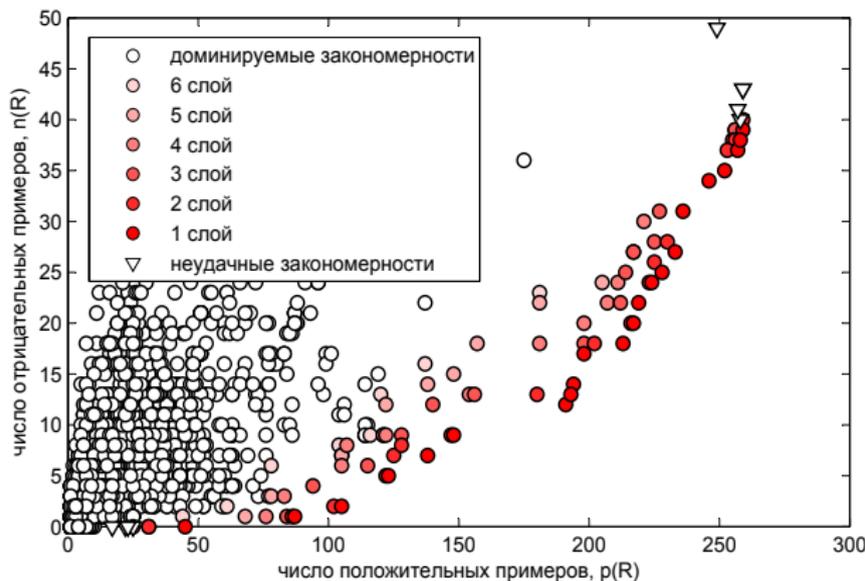
где \mathbb{Y} — множество классов,

R_y — множество правил, голосующих за класс $y \in \mathbb{Y}$,

$r: X \rightarrow \{0, 1\}$ — правило, w_r — вес правила.

Двухкритериальная оптимизация для поиска закономерностей

Парето-фронт — множество недоминируемых закономерностей (точка (p, n) недоминируема, если правее и ниже точек нет)

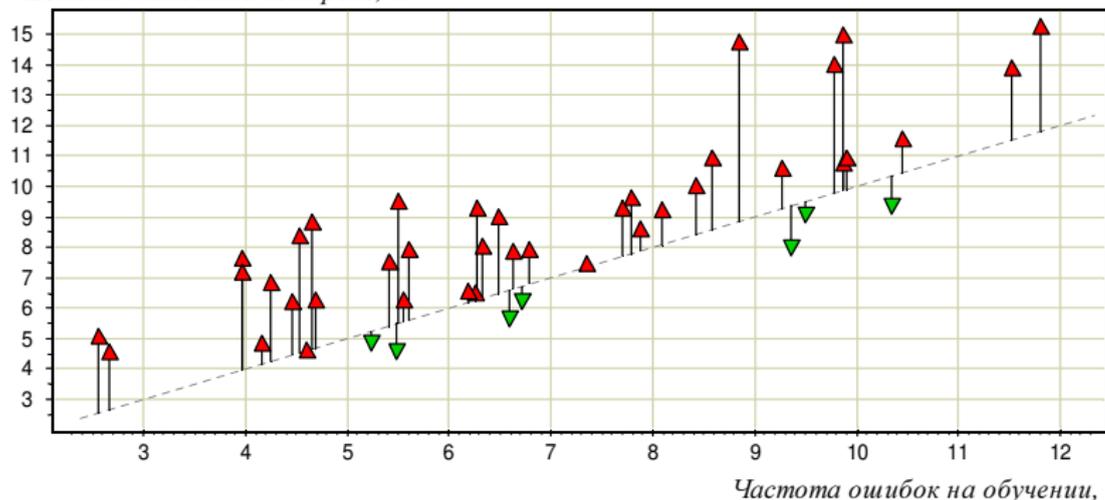


Задача кредитного скоринга german из репозитория UCI.

Пример. Переобучение закономерностей

Как отбросить переобученные закономерности на этапе обучения?

Частота ошибок на контроле, %



Задача предсказания отдалённого результата хирургического лечения атеросклероза. Точки — найденные закономерности.

Модификация критериев (p, n) с поправкой на переобучение

1. Для каждого тестируемого набора признаков J построить семейство конъюнкций с параметрами $(\theta_j)_{j \in J}$ и вычислить оценки расслоения–связности, как функции от ε ,

$$P\left[\frac{1}{k}n(r, \bar{X}) - \frac{1}{\ell}n(r, X) \geq \varepsilon\right] \leq \eta_n(\varepsilon);$$

$$P\left[\frac{1}{\ell}p(r, X) - \frac{1}{k}p(r, \bar{X}) \geq \varepsilon\right] \leq \eta_p(\varepsilon);$$

2. Обращение оценок: с вероятностью $1 - \eta$ (рекомендация: $\eta = \frac{1}{2}$)

$$n(r, \bar{X}) \leq \hat{n}(r, X) \equiv \frac{k}{\ell}n(r, X) + k\varepsilon_n(\eta);$$

$$p(r, \bar{X}) \geq \hat{p}(r, X) \equiv \frac{k}{\ell}p(r, X) - k\varepsilon_p(\eta).$$

3. Для выбора набора признаков J вместо $(p \rightarrow \max, n \rightarrow \min)$ использовать модифицированный критерий $(\hat{p} \rightarrow \max, \hat{n} \rightarrow \min)$.

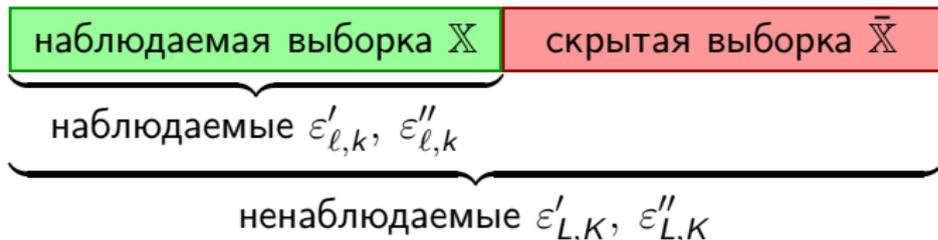
Три проблемы

- 1 Как получить оценки для ошибок I и II рода?

$$\text{I рода: } l'(r, x_i) = [r(x_i) = 0] [y_i = y], \quad m' = P - p, \quad \nu';$$

$$\text{II рода: } l''(r, x_i) = [r(x_i) = 1] [y_i \neq y], \quad m'' = n, \quad \nu''.$$

- 2 Как эффективно перебрать конъюнкции в нижних слоях графа расслоения–связности?
- 3 Имея оценку для разбиений полной выборки $\mathbb{X} = X \sqcup \bar{X}$, как оценить частоту ошибок на **действительно скрытой** контрольной выборке $\bar{\mathbb{X}}$ длины K ?



Проблема 1. Оценка переобучения для ошибок I и II рода

Теорема

Для метода обучения $\mu: \mathcal{H}(p, n) \rightarrow \max$, где \mathcal{H} строго возрастает по p и строго убывает по n ; любых $\mathbb{X}, R, \varepsilon \in (0, 1)$

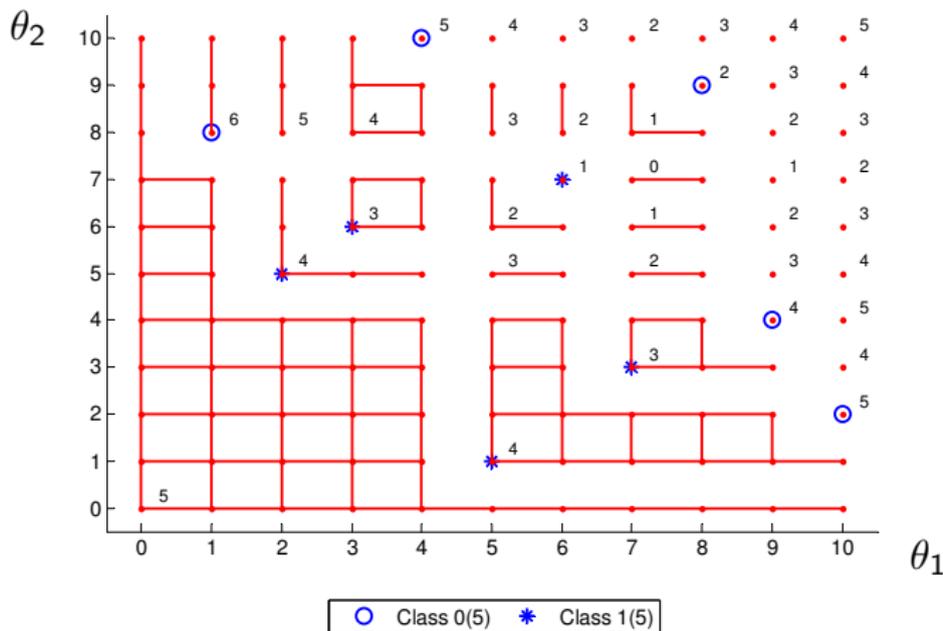
$$Q'_\varepsilon \leq \sum_{r \in R} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \mathcal{H}_{L-u-q}^{\ell-u, m'-q'} \left(\frac{\ell}{L} (m' - \varepsilon k) \right);$$

$$Q''_\varepsilon \leq \sum_{r \in R} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \mathcal{H}_{L-u-q}^{\ell-u, m''-q''} \left(\frac{\ell}{L} (m'' - \varepsilon k) \right);$$

где $u = |X_r|$ — верхняя связность правила r , $q = |X'_r|$,
 $q' = |X'_r \cap \mathbb{X}_y|$, $q'' = |X'_r \cap \mathbb{X}_{\bar{y}}|$ — неполноценность правила r
 относительно индикаторов ошибки $I' + I''$, I' , I'' ,
 $m' = m'(r, \mathbb{X})$, $m'' = m''(r, \mathbb{X})$ — число ошибок правила r
 относительно индикаторов ошибки I', I'' .

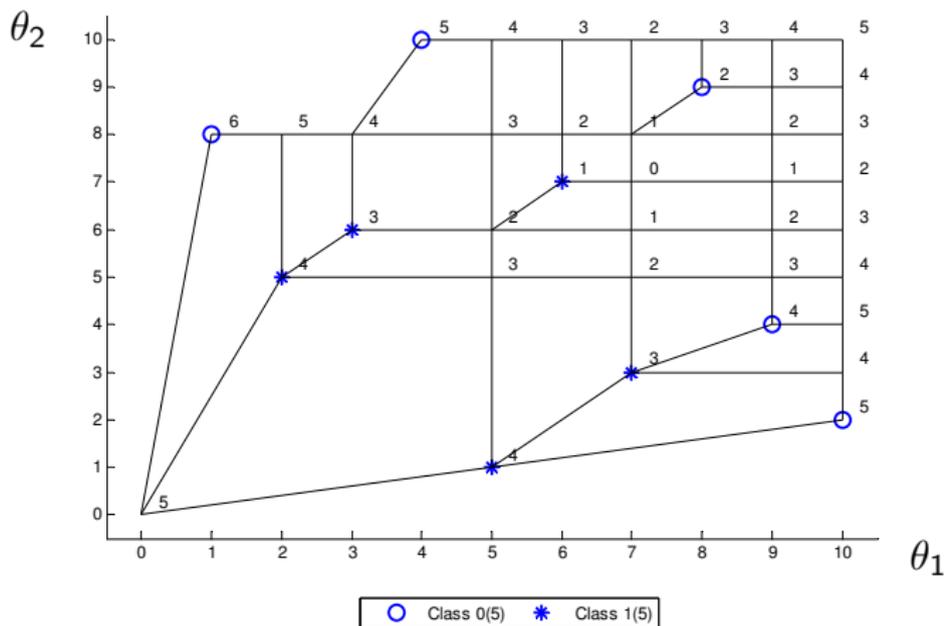
Проблема 2. Перебор конъюнктивных закономерностей

Пример: разделимая 2-мерная выборка, $L = 10$, два класса.
закономерности: $r(x) = [f_1(x) \leq \theta_1] \wedge [f_2(x) \leq \theta_2]$.



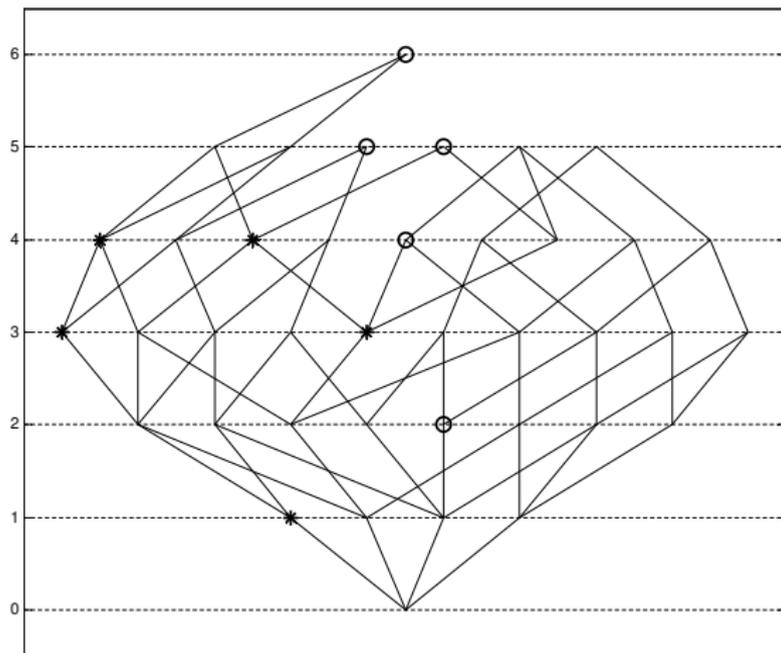
Проблема 2. Классы эквивалентности закономерностей

Пример: разделимая 2-мерная выборка, $L = 10$, два класса.
закономерности: $r(x) = [f_1(x) \leq \theta_1] \wedge [f_2(x) \leq \theta_2]$.



Проблема 2. Классы эквивалентности закономерностей

Пример: граф расслоения–связности, изоморфный графу классов эквивалентности с предыдущего слайда.



Проблема 3. Наблюдаемые и ненаблюдаемые оценки

Мы построили критерий предсказанной информативности:

$$\widehat{\mathcal{H}} = \mathcal{H} \left(\frac{k}{\ell} p(r, X) - k\varepsilon'_{\ell,k} \left(\frac{1}{2} \right), \frac{k}{\ell} n(r, X) + k\varepsilon''_{\ell,k} \left(\frac{1}{2} \right) \right).$$

Но нам нужен критерий, предсказывающий информативность на скрытой выборке \bar{X} длины K

$$\widehat{\mathcal{H}}_{\text{ненабл}} = \mathcal{H} \left(\frac{K}{L} p(r, \bar{X}) - K\varepsilon'_{L,K} \left(\frac{1}{2} \right), \frac{K}{L} n(r, \bar{X}) + K\varepsilon''_{L,K} \left(\frac{1}{2} \right) \right).$$

Однако мы не можем вычислить $\varepsilon'_{L,K}$ и $\varepsilon''_{L,K}$, т.к. \bar{X} скрыта.

Эвристика: заменив $\varepsilon'_{L,K}$ на $\varepsilon'_{\ell,k}$ и $\varepsilon''_{L,K}$ на $\varepsilon''_{\ell,k}$, получим наблюдаемый критерий предсказанной информативности:

$$\widehat{\mathcal{H}}_{\text{набл}} = \mathcal{H} \left(\frac{K}{L} p(r, \bar{X}) - K\varepsilon'_{\ell,k} \left(\frac{1}{2} \right), \frac{K}{L} n(r, \bar{X}) + K\varepsilon''_{\ell,k} \left(\frac{1}{2} \right) \right).$$

Почему мы имеем право на такую замену?

Эксперимент: зависимость $\widehat{\mathcal{H}}_{\text{ненабл}}$ от $\widehat{\mathcal{H}}_{\text{набл}}$

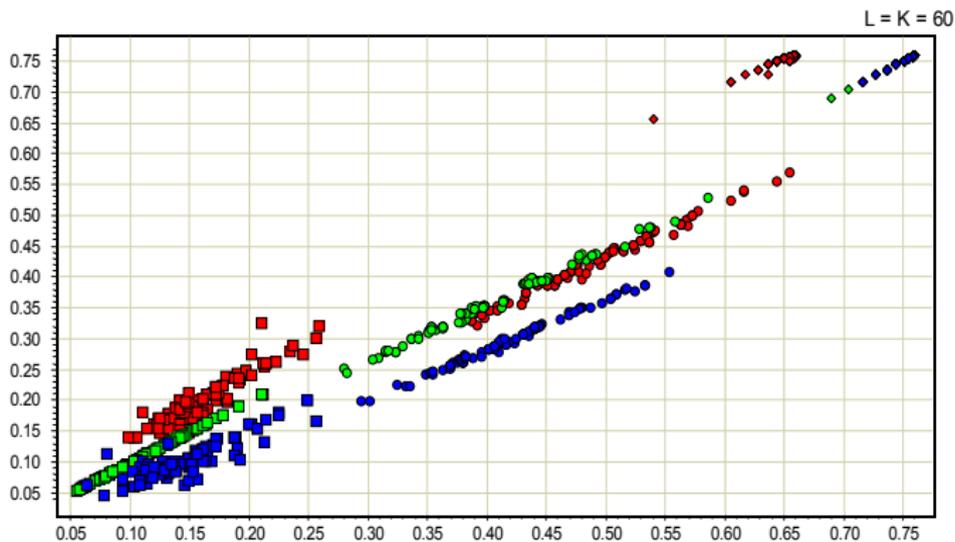
Двумерные модельные данные, $L = K = 60$;

уровень шума: \diamond — 0%, \circ — 10%, \square — 50%;

где шум: на границе классов, равномерно, внутри классов.

Точки соответствуют разбиениям $\mathbb{X} \sqcup \bar{\mathbb{X}}$.

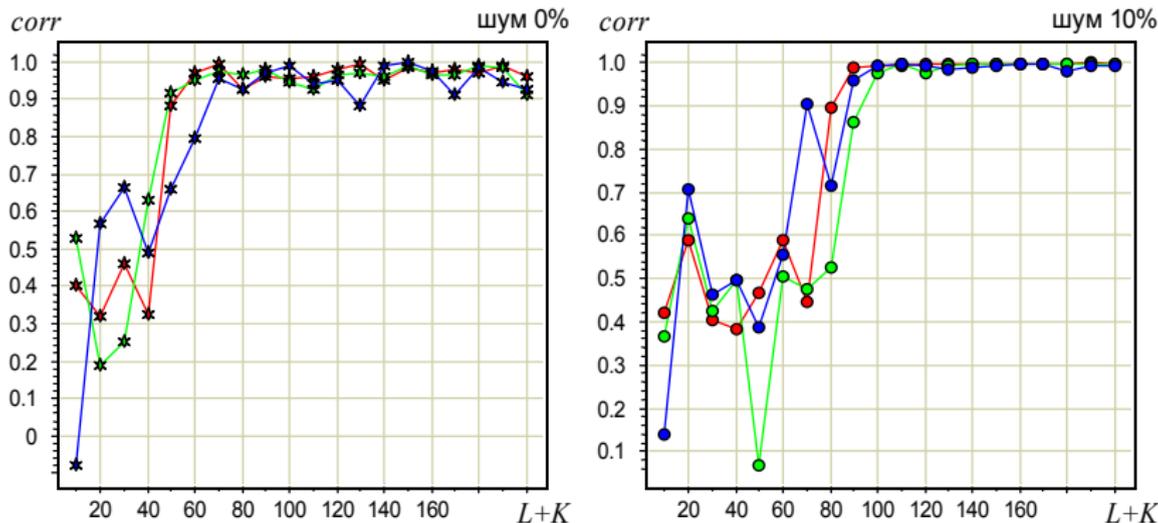
$\widehat{\mathcal{H}}_{\text{ненабл}}$



$\widehat{\mathcal{H}}_{\text{набл}}$

Эксперимент: корреляция $\hat{\mathcal{H}}_{\text{ненабл}}$ и $\hat{\mathcal{H}}_{\text{набл}}$

Зависимость корреляции $\hat{\mathcal{H}}_{\text{ненабл}}$ и $\hat{\mathcal{H}}_{\text{набл}}$ от длины супер-выборки при различном уровне и расположении шума.



Открытая проблема: теоретического обоснования пока нет.

Эксперименты на реальных данных

Реальные задачи классификации из репозитория UCI:

задачи	объектов	признаков
australian	690	14
echo cardiogram	74	10
heart disease	294	13
hepatitis	155	19
labor relations	40	16
liver	345	6

Методы обучения композиций логических закономерностей:

- WV (weighted voting) — взвешенное голосование;
- DL (decision list) — решающий список.

Методика тестирования: 10-кратный скользящий контроль.

Результаты эксперимента на реальных данных

методы	задачи					
	austr	echo	heart	hepa	labor	liver
RIPPER-opt	15.5	2.97	19.7	20.7	18.0	32.7
RIPPER+opt	15.2	5.53	20.1	23.2	18.0	31.3
C4.5 (Tree)	14.2	5.51	20.8	18.8	14.7	37.7
C4.5 (Rules)	15.5	6.87	20.0	18.8	14.7	37.5
C5.0	14.0	4.30	21.8	20.1	18.4	31.9
SLIPPER	15.7	4.34	19.4	17.4	12.3	32.2
LR	14.8	4.30	19.9	18.8	14.2	32.0
WV	14.9	4.37	20.1	19.0	14.0	32.3
DL	15.1	4.51	20.5	19.5	14.7	35.8
WV модиф.	14.1	3.2	19.3	18.1	13.4	30.2
DL модиф.	14.4	3.6	19.5	18.6	13.6	32.3

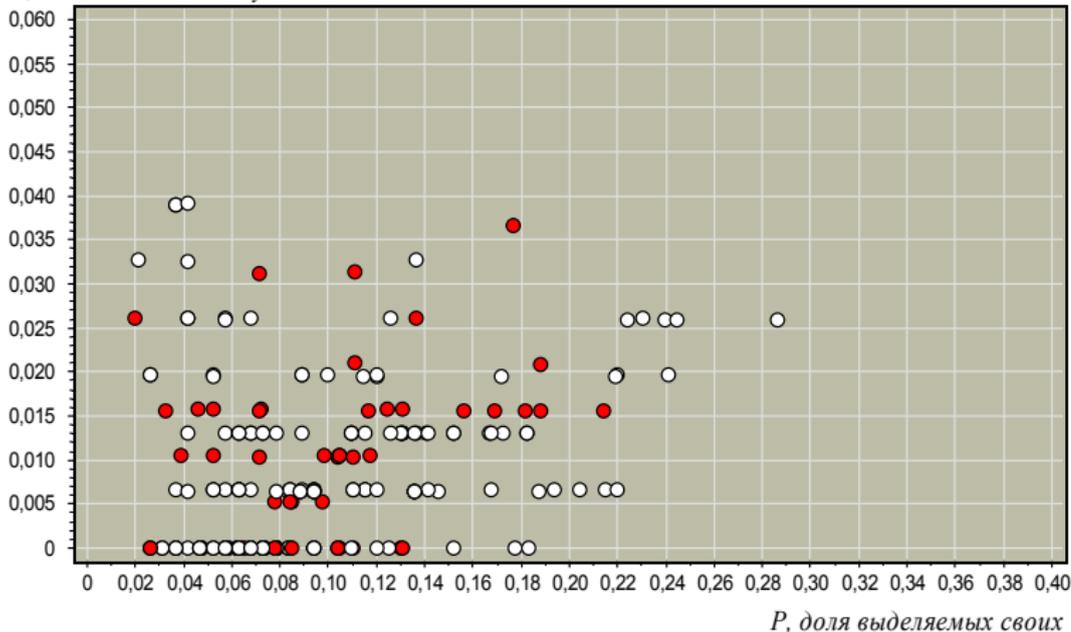
По каждой задаче выделено два лучших результата.

Закономерности в (p, n) -пространстве

Задача UCI:australian

 N , доля выделяемых чужих

Закономерности после модификации, контроль

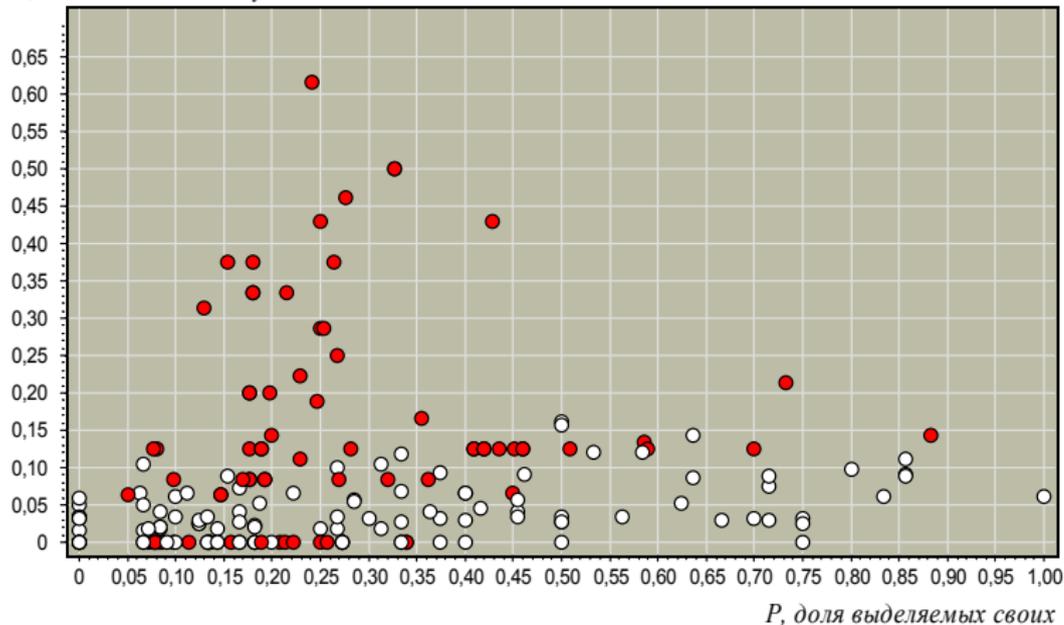


Закономерности в (p, n) -пространстве

Задача UCI:hepatitis

 N , доля выделяемых чужих

Закономерности до модификации, контроль

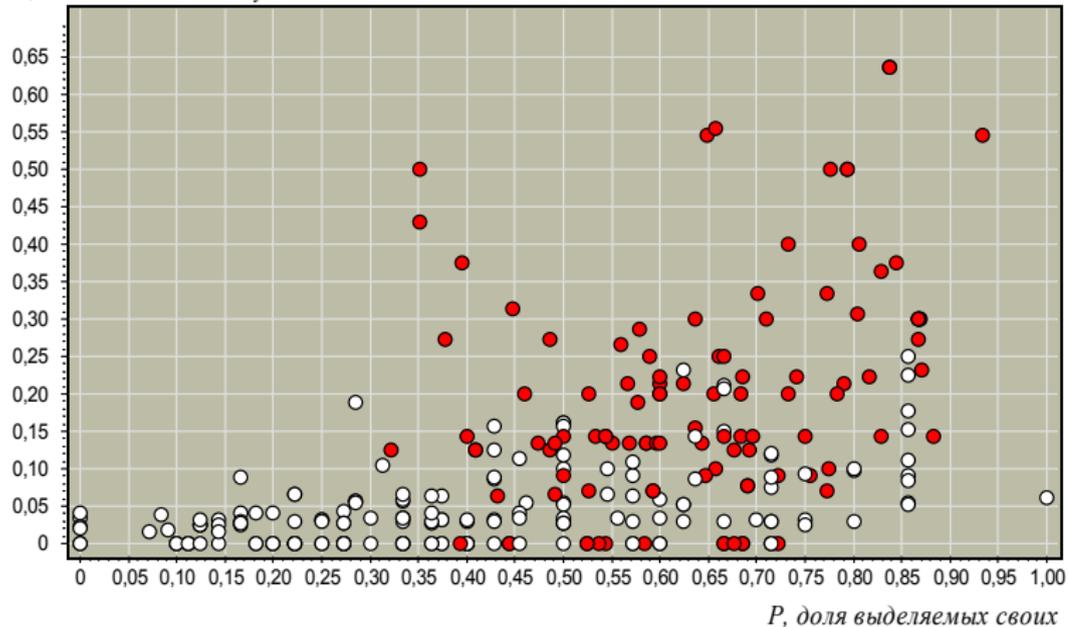


Закономерности в (p, n) -пространстве

Задача UCI:hepatitis

 N , доля выделяемых чужих

Закономерности после модификации, контроль



Обобщения метода порождающих и запрещающих множеств

Гипотеза ПЗМ (базовый вариант): $\forall a \in A \exists (X_a, X'_a)$:

$$[\mu X = a] \leq [X_a \subseteq X][X'_a \subseteq \bar{X}], \quad \forall X.$$

Обобщение 1: $\forall a \in A \exists (X_{av}, X'_{av}, c_{av})_{v \in V_a}$:

$$[\mu X = a] \leq \sum_{v \in V_a} c_{av} [X_{av} \subseteq X][X'_{av} \subseteq \bar{X}], \quad \forall X.$$

Обобщение 2: $\forall x_i \in \mathbb{X} \exists (X_{iv}, X'_{iv}, c_{iv})_{v \in V_i}$:

$$[x_i \in \bar{X}] I(\mu X, x_i) \leq \sum_{v \in V_i} c_{iv} [X_{iv} \subseteq X][X'_{iv} \subseteq \bar{X}], \quad \forall X.$$

Второе обобщение очень удобно для оценивания CCV.

Оценивание CCV методом ПЗМ по объектам

$$\begin{aligned} \text{CCV} &= E\nu(\mu X, \bar{X}) = \\ &= E \frac{1}{k} \sum_{i=1}^L [x_i \in \bar{X}] I(\mu X, x_i) = \\ &= \sum_{i=1}^L \sum_{v \in V_i} \frac{c_{iv}}{k} P[X_{iv} \subseteq X] [X'_{iv} \subseteq \bar{X}] = \\ &= \sum_{i=1}^L \sum_{v \in V_i} \frac{c_{iv}}{k} \frac{C_{L-u-q}^{\ell-u}}{C_L^{\ell}}. \end{aligned}$$

где $u = |X_{iv}|$, $q = |X'_{iv}|$.

- ⊕ это сумма по объектам, а не по алгоритмам;
- ⊖ пока только для kNN и монотонных классификаторов.

Классификатор ближайшего соседа (NN, nearest neighbor)

Задача классификации: $y_i = y(x_i)$ для всех $x_i \in \mathbb{X}$;
 $\rho(x, x')$ — функция расстояния на \mathbb{X} ,

$$\mu X(x) = y(\arg \min_{u \in X} \rho(x, u)).$$

Определение (профиль компактности выборки \mathbb{X})

доля объектов x_i , у которых m -й сосед x_{im} в другом классе:

$$K(m, \mathbb{X}) = \frac{1}{L} \sum_{i=1}^L [y(x_{im}) \neq y_i]; \quad m = 1, \dots, L-1,$$

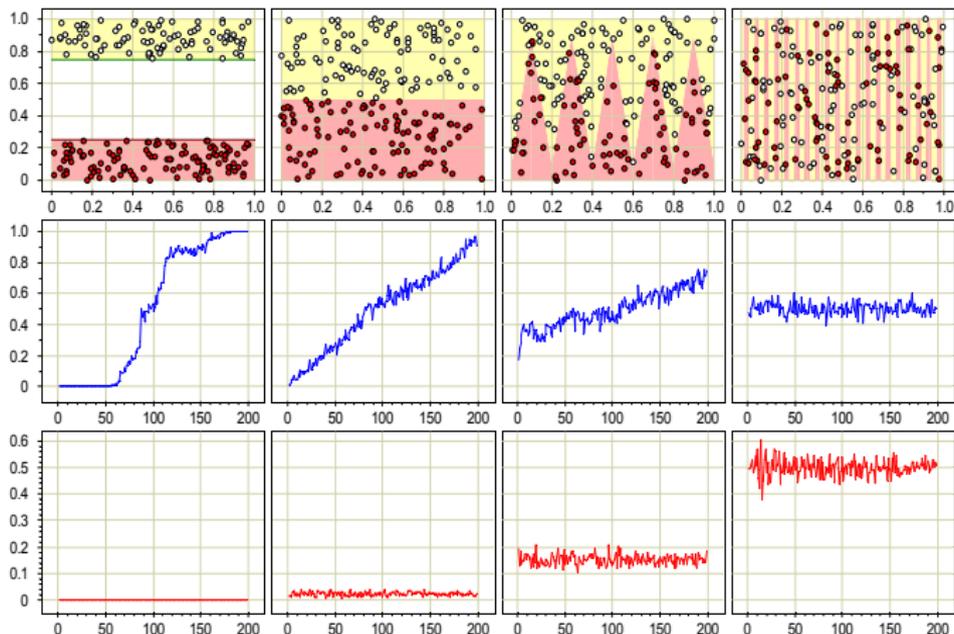
Теорема (точная оценка для метода ближайшего соседа)

$$\text{CCV}(\mu, \mathbb{X}) = \sum_{m=1}^k K(m, \mathbb{X}) \frac{C_{L-1-m}^{\ell-1}}{C_{L-1}^{\ell}}.$$

Профили компактности для серии модельных задач

средний ряд: профили компактности,

нижний ряд: зависимость CCV от длины контроля $k = |\bar{X}|$.



Доказательство через ПЗМ по объектам

x_{im} — m -й сосед объекта x_i в порядке неубывания $\rho(x_i, x_{im})$;
 $y_{im} = y(x_{im})$ — класс m -го соседа объекта x_i .

Система ПЗМ для объекта x_i :

$$[x_i \in \bar{X}] I(\mu X, x_i) = \sum_{m=1}^k [y_{im} \neq y_i] [x_{im} \in X] [x_i, x_{i1}, \dots, x_{i,m-1} \in \bar{X}]$$

Подставляем в формулу CCV:

$$\begin{aligned} \text{CCV} &= \sum_{i=1}^L \sum_{m=1}^k \frac{[y_{im} \neq y_i]}{k} \frac{C_{L-1-m}^{\ell-1}}{C_L^\ell} = \\ &= \sum_{m=1}^k \frac{1}{L} \underbrace{\sum_{i=1}^L [y_{im} \neq y_i]}_{K(m, X)} \frac{C_{L-1-m}^{\ell-1}}{C_{L-1}^\ell}. \end{aligned}$$

Задача отбора эталонов $\Omega \subseteq \mathbb{X}$ (prototype learning)

Модификация NN $\mu_\Omega: X \mapsto a$, $a(x) = y(\arg \min_{x' \in \Omega} \rho(x, x'))$.

Определение (профиль компактности относительно Ω)

$$K(m, \Omega) = \frac{1}{L} \sum_{i=1}^L [y(x_i) \neq y(x_{im}^\Omega)]; \quad m = 1, \dots, |\Omega|.$$

где x_{im}^Ω — m -й сосед объекта x_i из множества Ω ;

Теорема (точное выражение CCV относительно Ω)

$$\text{CCV}(\mu_\Omega, \mathbb{X}) = \sum_{i=1}^L \underbrace{\sum_{m=1}^k [y(x_i) \neq y(x_{im}^\Omega)]}_{T(x_i, \Omega) \text{ — вклад объекта } x_i \text{ в CCV}} \frac{C_{L-1-m}^{\ell-1}}{LC_{L-1}^\ell}.$$

Жадные алгоритмы отбора эталонов

Задача: найти Ω : $CCV(\mu_{\Omega}, \mathbb{X}) \rightarrow \min$.

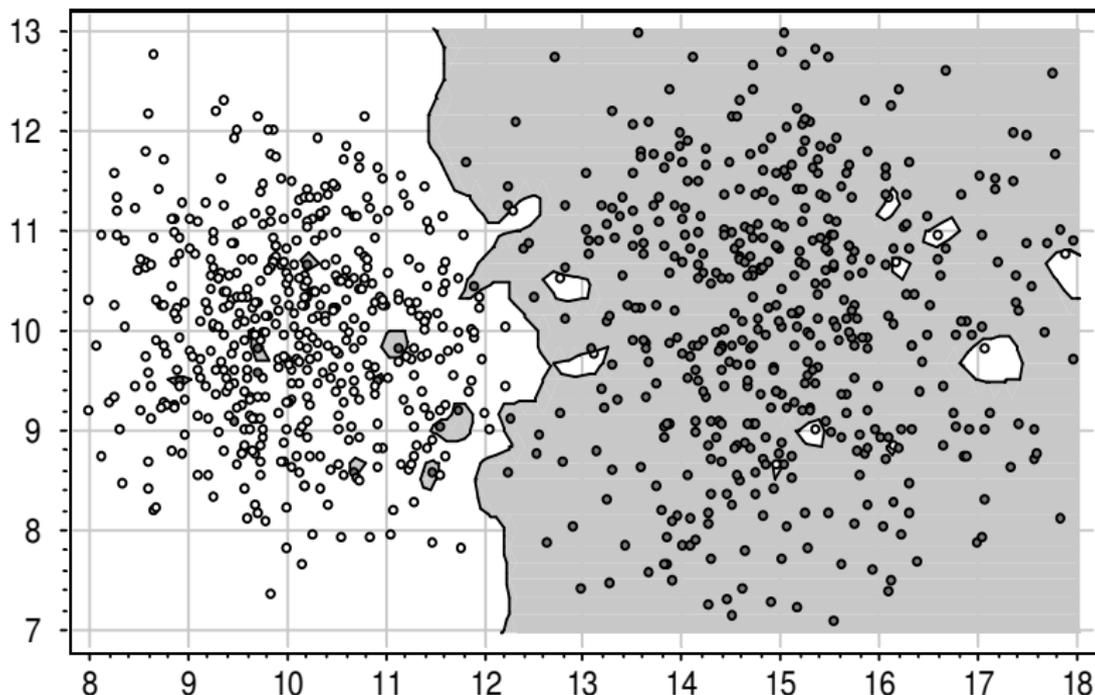
Жадный алгоритм удаления не-эталонов

- 1: $\Omega := \mathbb{X}$;
- 2: **повторять**
- 3: найти $x \in \Omega$: $CCV(\mu_{\Omega \setminus \{x\}}, \mathbb{X}) \rightarrow \min$;
- 4: $\Omega := \Omega \setminus \{x\}$;
- 5: **пока** CCV уменьшается или почти не увеличивается;

Жадный алгоритм добавления эталонов

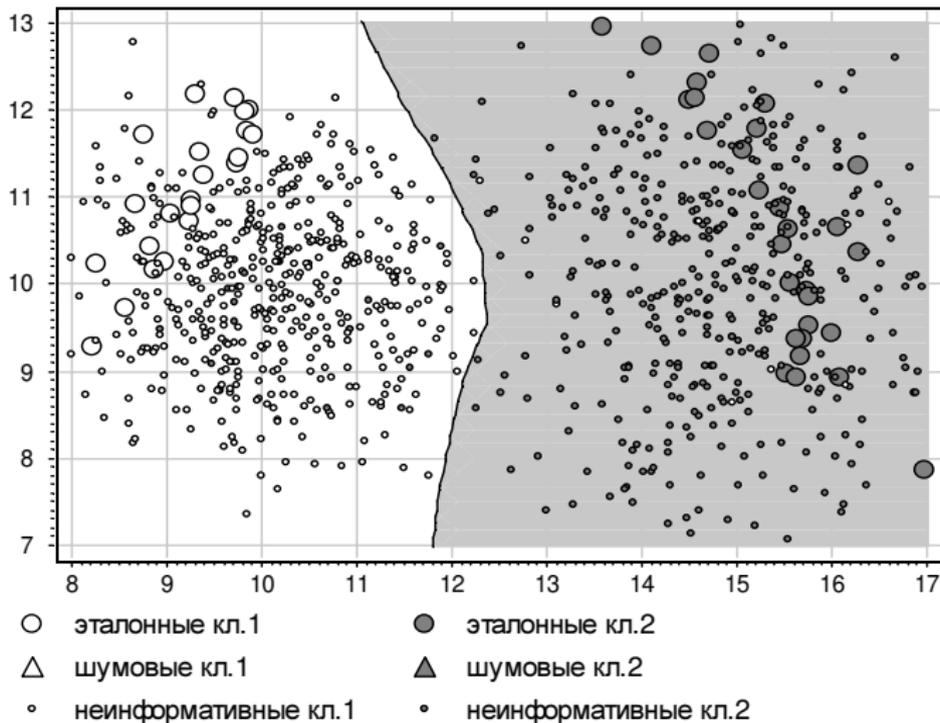
- 1: $\Omega := \{\text{по одному объекту от каждого класса}\}$;
- 2: **повторять**
- 3: найти $x \in \mathbb{X} \setminus \Omega$: $CCV(\mu_{\Omega \cup \{x\}}, \mathbb{X}) \rightarrow \min$;
- 4: $\Omega := \Omega \cup \{x\}$;
- 5: **пока** CCV уменьшается;

Модельные данные

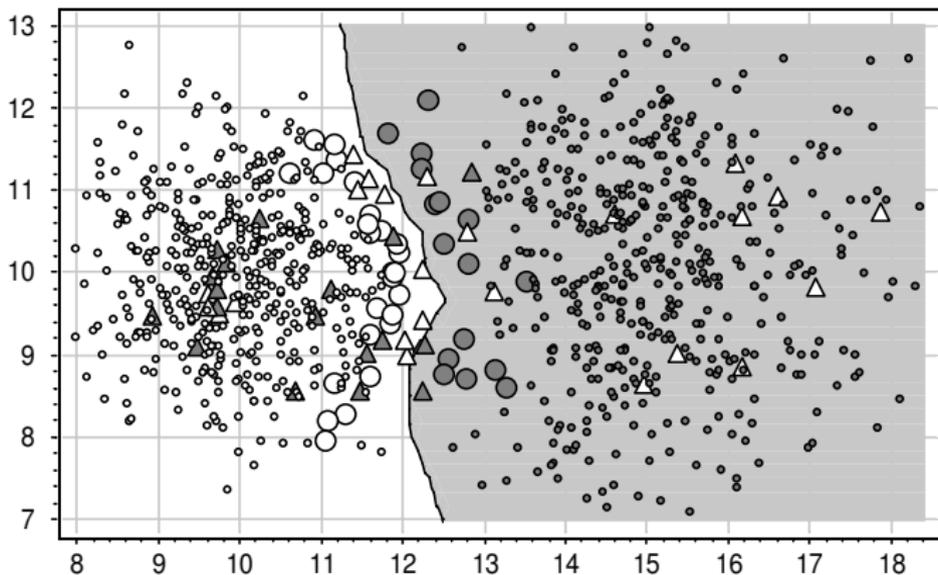


Модельная задача классификации: 1000 объектов, метод NN.

Жадное добавление эталонных объектов



Жадное удаление не-эталонных объектов



○ эталонные кл.1

△ шумовые кл.1

◦ неинформативные кл.1

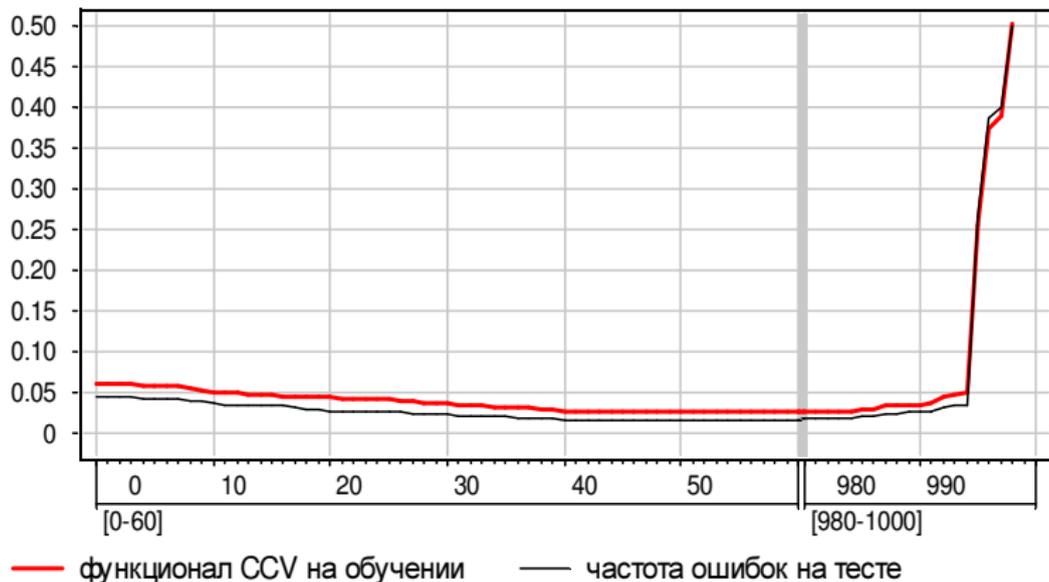
● эталонные кл.2

▲ шумовые кл.2

◦ неинформативные кл.2

Жадное удаление не-эталонных объектов

Зависимость CCV от числа удалённых неэталонных объектов.



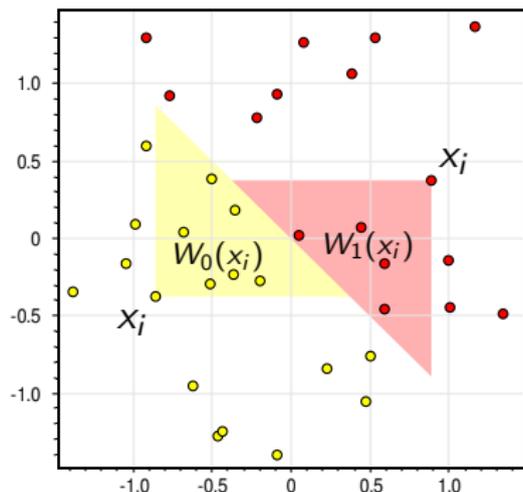
Чудо: при отборе эталонов переобучения вообще нет!

Монотонные алгоритмы классификации: определения

Задача классификации: \mathbb{X} — ч.у. множество, $\mathbb{Y} = \{0, 1\}$,
 A — множество монотонных отображений $a: \mathbb{X} \rightarrow \mathbb{Y}$.

Верхний клин объекта x_j : $W_0(x_j) = \{x \in \mathbb{X}: x_j < x \text{ и } y(x) = 0\}$.

Нижний клин объекта x_j : $W_1(x_j) = \{x \in \mathbb{X}: x < x_j \text{ и } y(x) = 1\}$.



Профиль монотонности выборки

Определение (профиль монотонности выборки \mathbb{X})

доля объектов $x_i \in \mathbb{X}$ с клином мощности m :

$$M(m, \mathbb{X}) = \frac{1}{L} \sum_{i=1}^L [|W_{y(x_i)}(x_i)| = m]; \quad m = 0, \dots, L-1.$$

Теорема

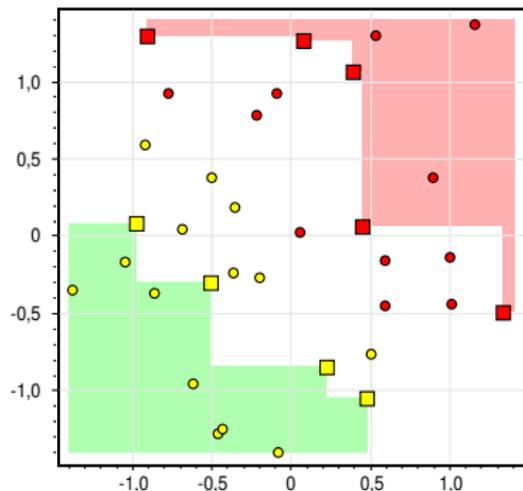
Пусть μ — метод минимизации эмпирического риска в классе всех монотонных функций, θ — степень немонотонности выборки \mathbb{X} . Тогда

$$\text{CCV}(\mu, \mathbb{X}) \leq \sum_{m=0}^{\theta L + k - 1} M(m, \mathbb{X}) \mathcal{H}_{L-1}^{\ell, m}(\theta L).$$

Монотонный классификатор ближайшего соседа

Если учесть, как именно μ_X классифицирует контрольные объекты из зазора, то получается точная оценка.

Минимум оценки достигается, когда разделяющая поверхность проходит по середине, т.е. по принципу ближайшего соседа.



Свойства профиля монотонности и оценки CCV

Выводы

- Невырожденность: $CCV(\mu, \mathbb{X}) \leq 1$.
- Для минимизации CCV важен только начальный участок профиля, т. к. $\mathcal{H}_{L-1}^{\ell, m}(\theta L) \rightarrow 0$ по m при малых θ .
- Для минимизации CCV отношение порядка на множестве объектов \mathbb{X} должно быть близко к линейному вблизи границы классов.
- Минимизация CCV приводит к повышению обобщающей способности алгоритмической композиции с монотонной корректирующей операцией [И. Гуз, 2008, 2011].

Замечание. VC -теория даёт сильно завышенные оценки для монотонных семейств алгоритмов (эффективная ёмкость определяется максимальной длиной антицепи).

Спасибо за внимание!

Воронцов Константин Вячеславович
voron@forecsys.ru

Страницы на www.MachineLearning.ru:

- Теория надёжности обучения по прецедентам (курс лекций, К. В. Воронцов)
- Расслоение и сходство алгоритмов (виртуальный семинар)
- Слабая вероятностная аксиоматика
- Участник:Vokov