Комбинаторная теория переобучения и её применения

Воронцов Константин Вячеславович

ВЦ РАН • МФТИ • ВМК МГУ • Форексис • Яндекс

Семинар лаборатории PreMoLab
http://premolab.ru/laboratory-seminar

24 мая 2012

План семинара

1. **Введение**
   - Проблема надёжности обучения по прецедентам
   - Слабая вероятностная аксиоматика
   - Теория Вапника–Червоненкиса
2. **Метод порождающих и запрещающих множеств**
   - Эксперименты
   - Оценки расслоения–связности
   - Логические алгоритмы классификации
3. **Оценки полного скользящего контроля**
   - Метод ближайшего соседа и отбор эталонов
   - Монотонный классификатор
   - Последние результаты. Открытые проблемы. Планы

Основные понятия и обозначения

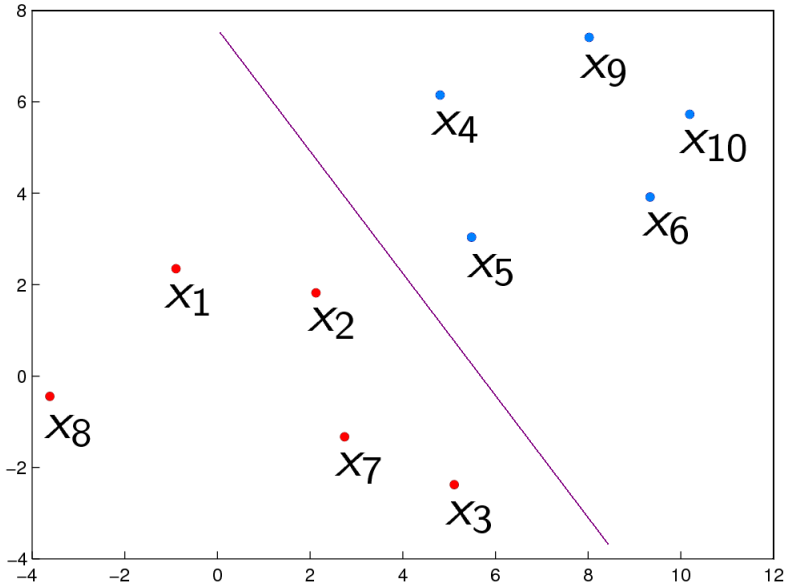
$\mathbb{X} = \{x_1, \ldots, x_L\}$ — конечное *генеральное множество* объектов;
$A = \{a_1, \ldots, a_D\}$ — конечное множество *алгоритмов* (гипотез);
$I(a, x) = [\text{алгоритм } a \text{ ошибается на объекте } x]$;

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

	a_1	a_2	a_3	a_4	a_5	a_6	$\cdots$	a_D	
x_1	1	1	0	**0**	0	1	$\cdots$	1	X — 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	$\bar{X}$ — скрытая
$\ldots$	0	0	0	**1**	0	0	$\cdots$	1	(контрольная) выборка
x_L	0	1	1	**1**	1	1	$\cdots$	0	длины $k = L - \ell$

$m(a, X) = \sum_{x \in X} I(a, x)$ — *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;

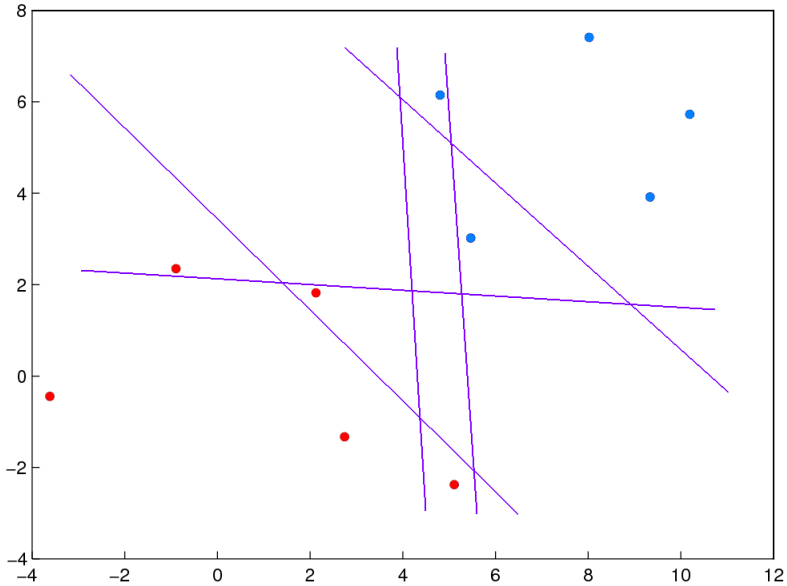
Пример. Матрица ошибок линейных классификаторов



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

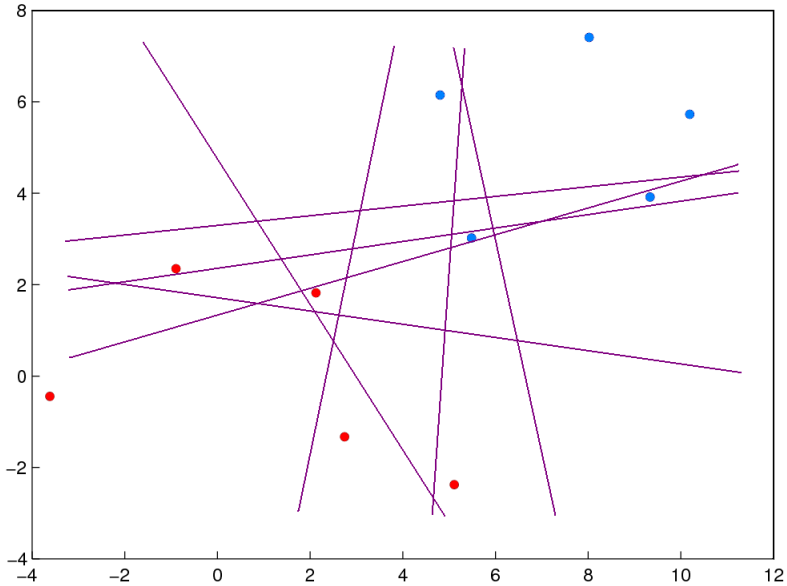


1 вектор с 0 ошибками

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* $\mu\colon 2^{\mathbb{X}} \to A$ произвольной выборке $X \subset \mathbb{X}$ ставит в соответствие некоторый алгоритм $a \in A$.

Пример. Метод *минимизации эмпирического риска*:

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

Задача обучения по прецедентам — это
задача принятия решений на основе неполной информации.

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

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

$\mathsf{P} \equiv \mathsf{E} \equiv \frac{1}{C_L^\ell} \sum_{X \subset \mathbb{X}}$ — 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):
$$\mathrm{CCV}(\mu, \mathbb{X}) = \mathsf{E}\, \nu\big(\mu X, \bar{X}\big).$$
- Ожидаемая переобученность (Expected OverFitting):
$$\mathrm{EOF}(\mu, \mathbb{X}) = \mathsf{E}\, \delta(\mu, X, \bar{X}).$$
- Вероятность большой частоты ошибок на контроле:
$$R_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\nu(\mu X, \bar{X}) \geqslant \varepsilon\big].$$
- Вероятность переобучения:
$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\delta(\mu, X, \bar{X}) \geqslant \varepsilon\big].$$

Отличия от общепринятой постановки задачи

Функционалы обобщающей способности:

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

$$S_\varepsilon(A, \mathbb{X}) = \mathsf{P}\Big[\sup_{a \in A}\big(P(a) - \nu(a, X)\big) \geqslant \varepsilon\Big]$$

- Вероятность переобучения

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon\big] \ll S_\varepsilon(A, \mathbb{X}).$$

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.

Как измерить скорость сходимости в законе больших чисел?

Вероятность большого отклонения частоты ν от вероятности P:

$$P_\varepsilon = \mathsf{P}_X\big[P - \nu(X) > \varepsilon\big].$$

Проблемы

- Не удаётся измерить вероятность P_ε как частоту, если вероятность P неизвестна.
- Не удаётся получить точное выражение для P_ε; оценки P_ε либо завышенные, либо асимптотические.

Причина этих проблем — вероятность P *инфинитарна*.

Заменим P частотой $\nu(\bar{X})$ на неизвестной (скрытой) выборке $\bar{X}$:

$$Q_\varepsilon = \mathsf{P}_{X,\bar{X}}\big[\nu(\bar{X}) - \nu(X) > \varepsilon\big].$$

Тогда $\mathsf{P}_{X,\bar{X}}$ можно понимать как *долю разбиений* $X \sqcup \bar{X} = \mathbb{X}$.

Философия слабой вероятностной аксиоматики

- В анализе данных выборки могут быть только конечными, в том числе скрытые будущие данные, так как конечно время, отпущенное исследователям на решение задачи.
- Собираемся оценивать частоты, которые станут известны в будущем, по частотам, которые наблюдались в прошлом.
- Чтобы что-то предсказывать, достаточно предположить, что все разбиения неслучайной конечной выборки на наблюдаемую и скрытую подвыборки равновероятны. Это аналог гипотезы о независимости наблюдений. Но возможны и более сложные условия перестановочности.
- Остаёмся в рамках комбинаторной вероятности XVII века. Не пользуемся теорией меры на бесконечных множествах.
- Некоторые частоты будем называть вероятностями ради сохранения привычного языка и обозначений.

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. Гоппа. 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*.»

В слабой аксиоматике доказываются:

- закон больших чисел;
- сходимость эмпирических распределений (критерий Смирнова для непрерывных и дискретных случайных величин);
- доверительные интервалы для квантилей;
- критерий знаков, Уилкоксона–Манна–Уитни, и другие непараметрические критерии;
- оценки Вапника–Червоненкиса;

Открытый вопрос:

Насколько значительная часть теории вероятностей, математической статистики, теории информации, теории статистического обучения может быть переформулирована в рамках слабой аксиоматики?

Альтернативная интерпретация

Можно и не говорить о «слабой вероятностной аксиоматике»...

В *комбинаторной теории переобучения* изучаются способы вычисления функционалов Q_ε, CCV и других, 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:

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

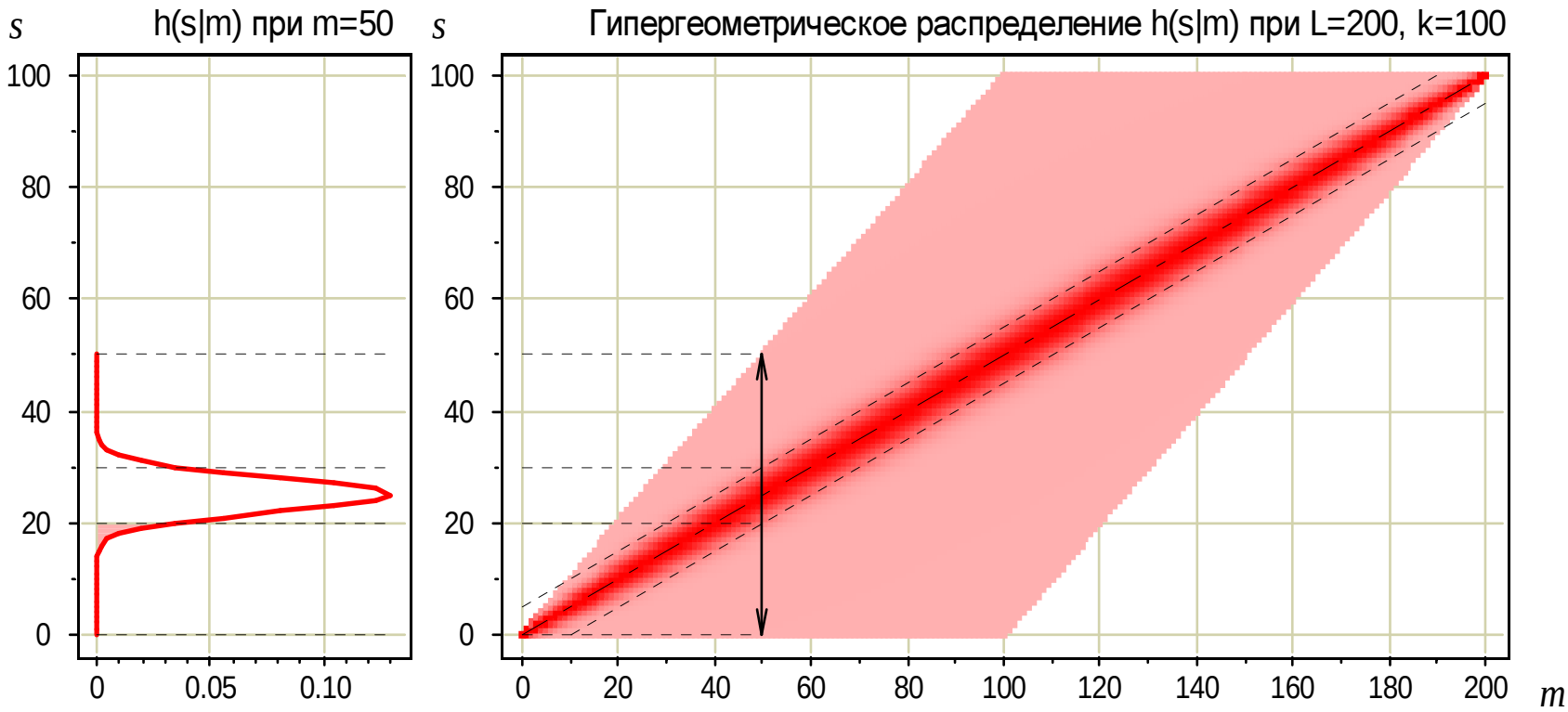
Теорема

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

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

где $\mathcal{H}_L^{\ell,\, m}(z) = \sum\limits_{s=0}^{\lfloor z \rfloor} \dfrac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ — 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: $\nu(a, X) \to \nu(a, \bar{X})$ при $\ell, k \to \infty$.

Теория Вапника–Червоненкиса в слабой аксиоматике

Теорема

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

$$Q_\varepsilon(\mu,\mathbb{X}) \overset{\text{uniform bound}}{\leqslant} \mathsf{P} \max_{a\in A} \big[\delta(a,X,\bar{X}) \geqslant \varepsilon\big] \overset{\text{union bound}}{\leqslant} \sum_{a\in A} Q_\varepsilon(a,\mathbb{X})$$

$$\leqslant |A| \cdot \max_a Q_\varepsilon(a,\mathbb{X}) \leqslant$$

$$\leqslant |A| \cdot \max_m \mathcal{H}_L^{\ell,m}\left(\tfrac{\ell}{L}\big(m-\varepsilon k\big)\right) \leqslant$$

$$\leqslant |A| \cdot \tfrac{3}{2}\exp\big(-\varepsilon^2\ell\big), \quad \text{при } \ell = k.$$

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

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]

Эксперименты с модельными семействами алгоритмов

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

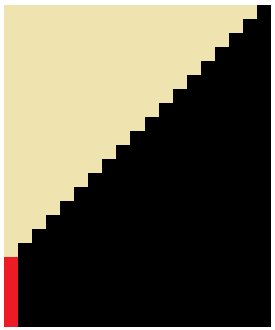
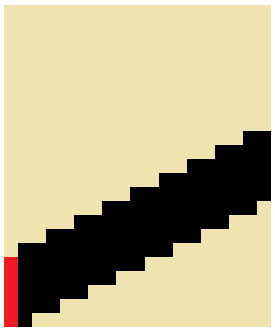
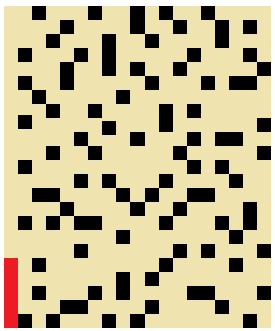
Цель эксперимента — понять, как структура матрицы ошибок влияет на вероятность переобучения.

1. Будем строить *модельные семейства алгоритмов*, задавая их непосредственно своими матрицами ошибок.
2. Будем оценивать вероятность методом Монте–Карло — как долю разбиений выборки из случайного подмножества N разбиений, $|N|$ порядка 10^3–10^4:

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

Эксперимент с четырьмя модельными семействами

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

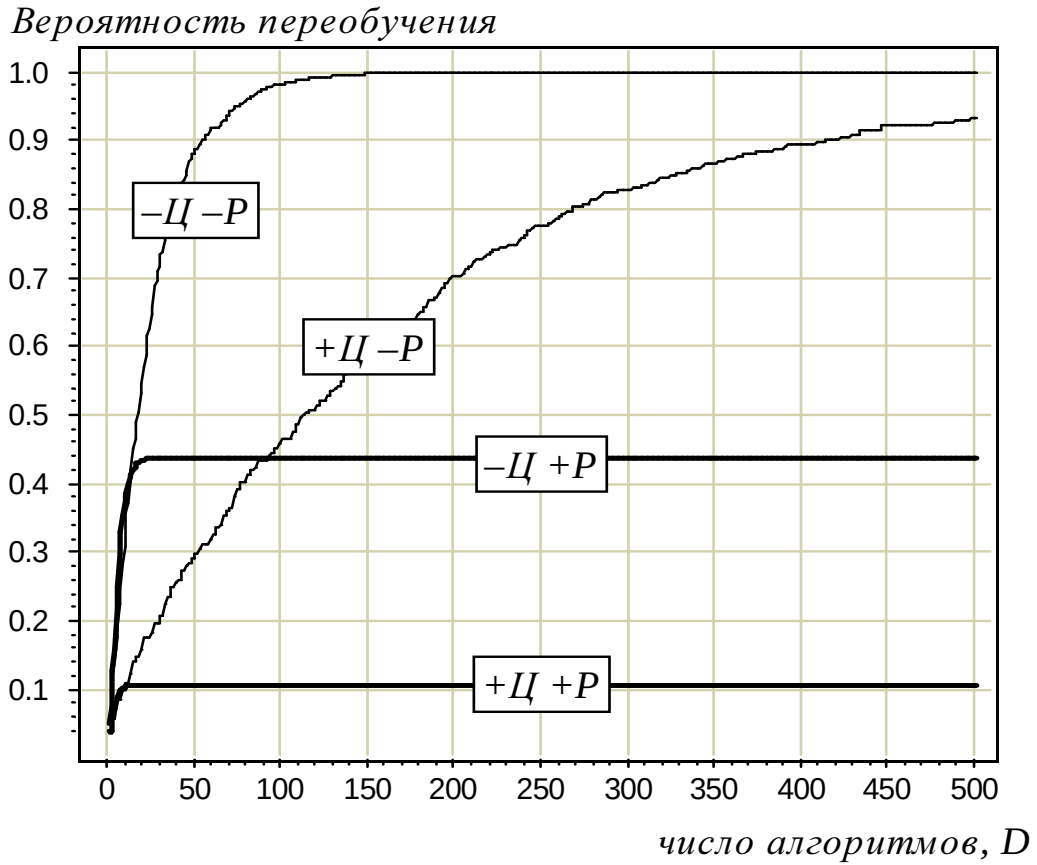
	есть расслоение по числу ошибок	**нет расслоения** по числу ошибок
есть связность, соседние алгоритмы отличаются на одном объекте, образуется *цепь*		
нет связности, соседние алгоритмы существенно различны, *цепь* не образуется		

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 $Q_\varepsilon(D)$



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 $a \leqslant b$: $I(a,x) \leqslant I(b,x)$ для всех $x \in \mathbb{X}$;

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

Определение

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

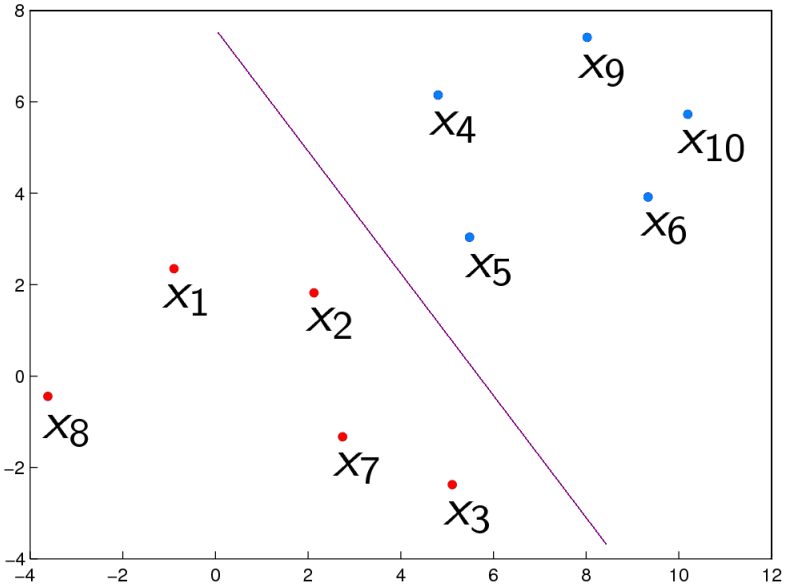
A — множество попарно различных векторов ошибок;

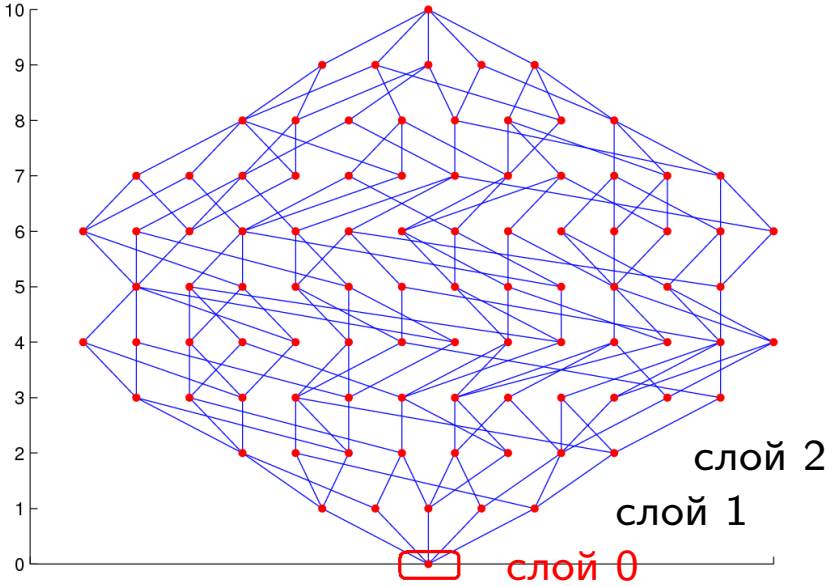
$E = \{(a,b) \colon a \prec b\}$.

Свойства графа расслоения–связности:

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

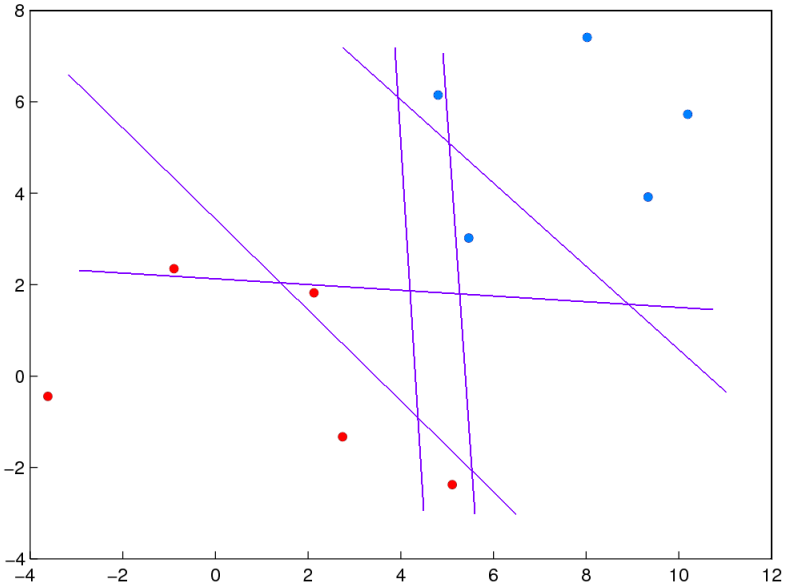
Пример. Семейство линейных алгоритмов классификации

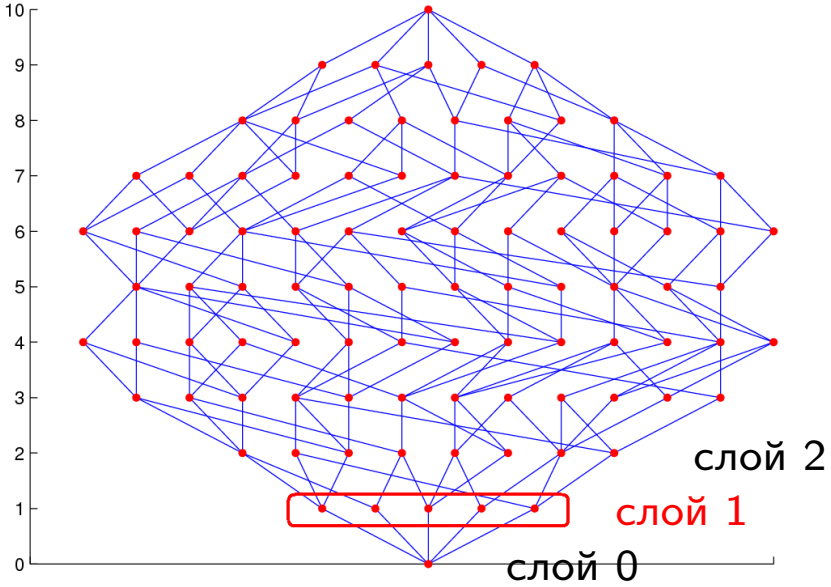




	слой 0
x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0

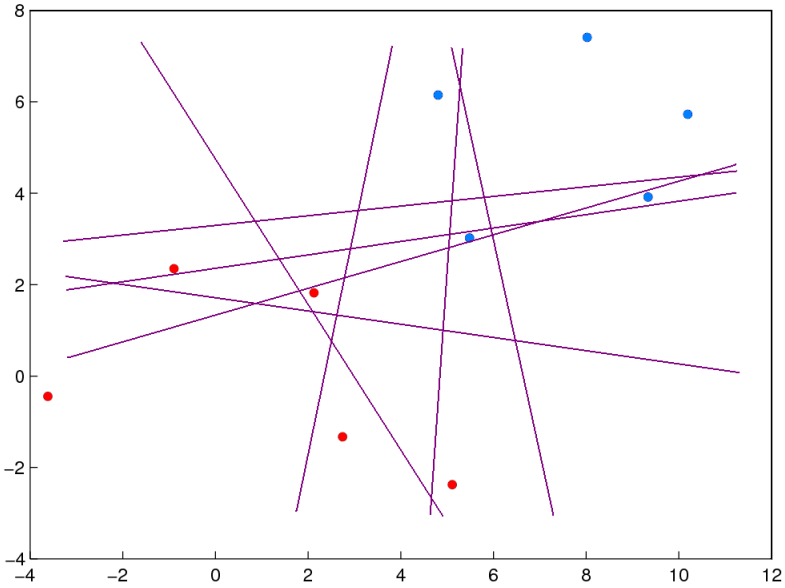
Пример. Семейство линейных алгоритмов классификации

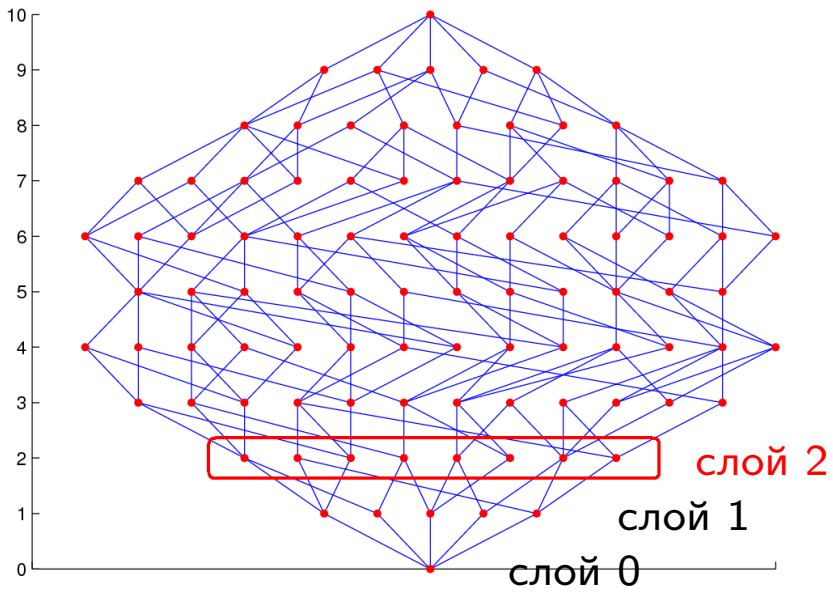




	слой 0	слой 1				
x_1	0	1	0	0	0	0
x_2	0	0	1	0	0	0
x_3	0	0	0	1	0	0
x_4	0	0	0	0	1	0
x_5	0	0	0	0	0	1
x_6	0	0	0	0	0	0
x_7	0	0	0	0	0	0
x_8	0	0	0	0	0	0
x_9	0	0	0	0	0	0
x_{10}	0	0	0	0	0	0

Пример. Семейство линейных алгоритмов классификации





	слой 0	слой 1					слой 2								
x_1	0	1	0	0	0	0	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	...

Порождающее и запрещающее множества алгоритма

Определение

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

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

X_a называется *порождающим множеством* алгоритма a.

Определение

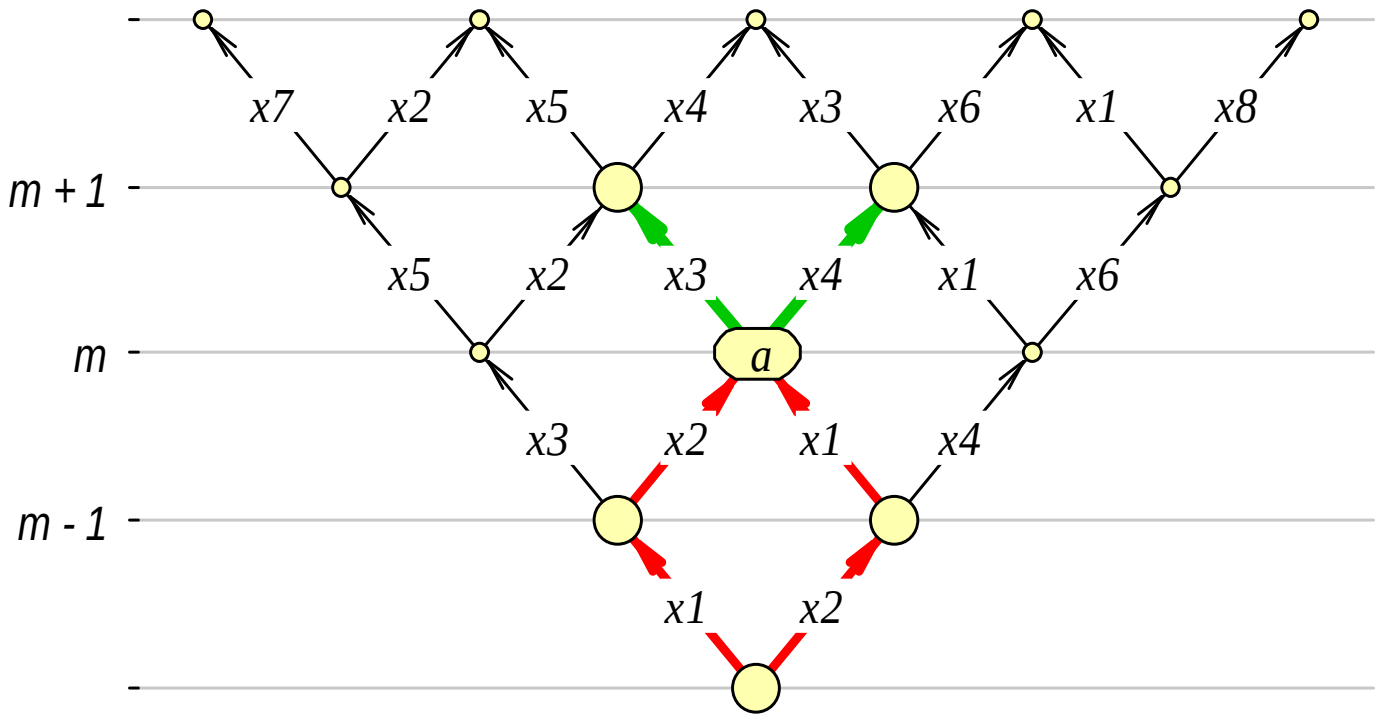
Неполноценность $q(a)$ алгоритма a — это число различных объектов, соответствующих всем рёбрам на путях, ведущих в a:

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

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 $\varepsilon \in (0,1)$

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

где $u = |X_a|$ — верхняя связность алгоритма a,
$q = |X'_a|$ — неполноценность алгоритма a,
$m = m(a, \mathbb{X})$ — число ошибок алгоритма a,

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

— функция гипергеометрического распределения:

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

Идея доказательства

Опр. Метод обучения μ называется *монотонным*, если

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

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

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

Опр. Метод μ называется *пессимистичным*, если

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

Основная лемма

Если метод обучения μ монотонный и пессимистичный, то

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

Идея доказательства

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

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

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

4. По формуле полной вероятности:

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

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

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):
$$Q_\varepsilon = \mathcal{H}_L^{\ell,\,m}\left(\frac{\ell}{L}\left(m-\varepsilon k\right)\right) \to 0 \quad \text{при} \quad \ell, k \to \infty.$$
3. При $q = u = 0$ и $\ell = k$ это оценка Вапника-Червоненкиса:
$$Q_\varepsilon \;\leqslant\; \sum_{a\in A} \mathcal{H}_L^{\ell,\,m}\left(\frac{\ell}{L}\left(m-\varepsilon k\right)\right) \;\leqslant\; |A| \cdot \exp(-\varepsilon\ell^2).$$
4. Оценка обращается в равенство в случае многомерных монотонных сетей алгоритмов [Павел Ботов].
5. Имеется критерий точности оценки [Никита Животовский]

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

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

$$S_\varepsilon(A,\mathbb{X}) = \mathsf{P}\Big[\sup_{a\in A}\big(\nu(a,\bar{X}) - \nu(a,X)\big) \geqslant \varepsilon\Big],$$

которая учитывает связность, но не учитывает расслоение.

7. Вероятность получить алгоритм в результате обучения

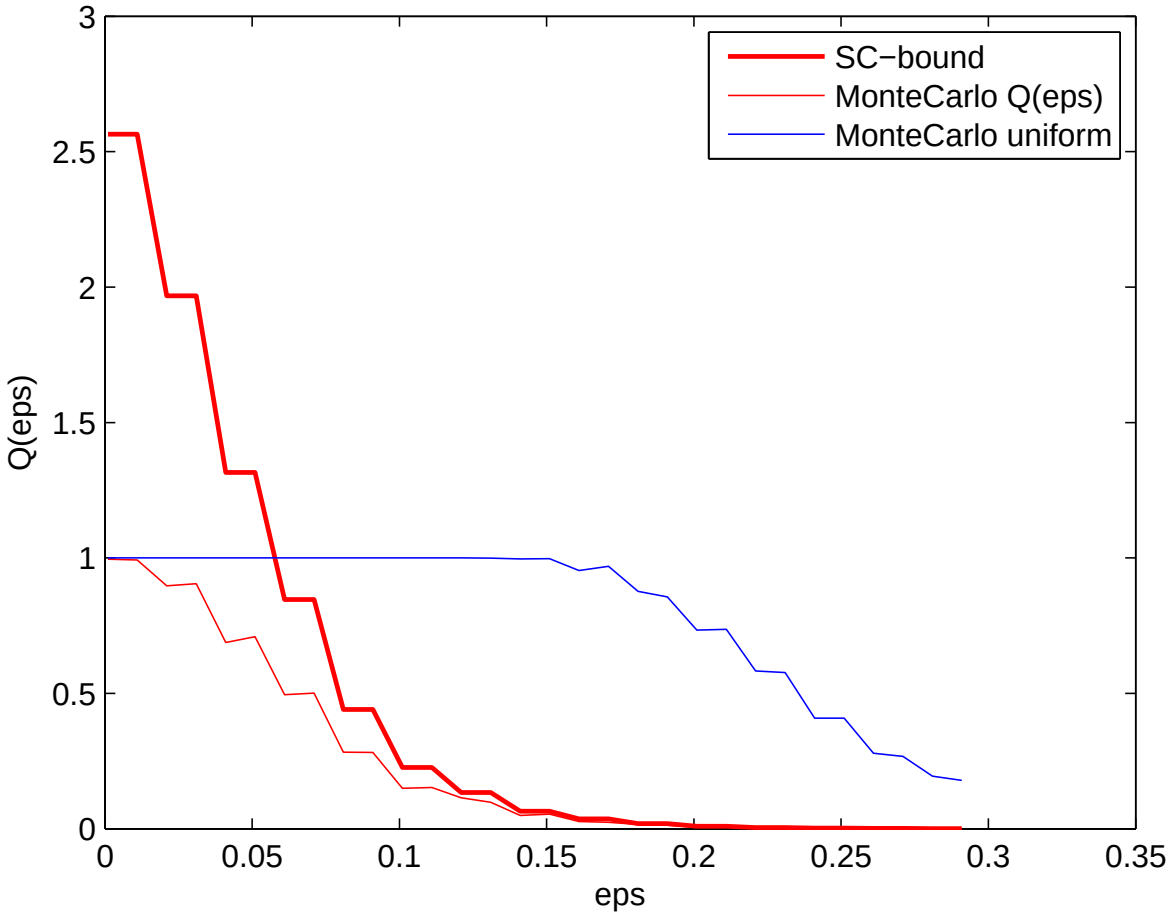
$$\mathsf{P}\big[\mu X = a\big] \;\leqslant\; P_a = C_{L-u-q}^{\ell-u} / C_L^\ell.$$

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.

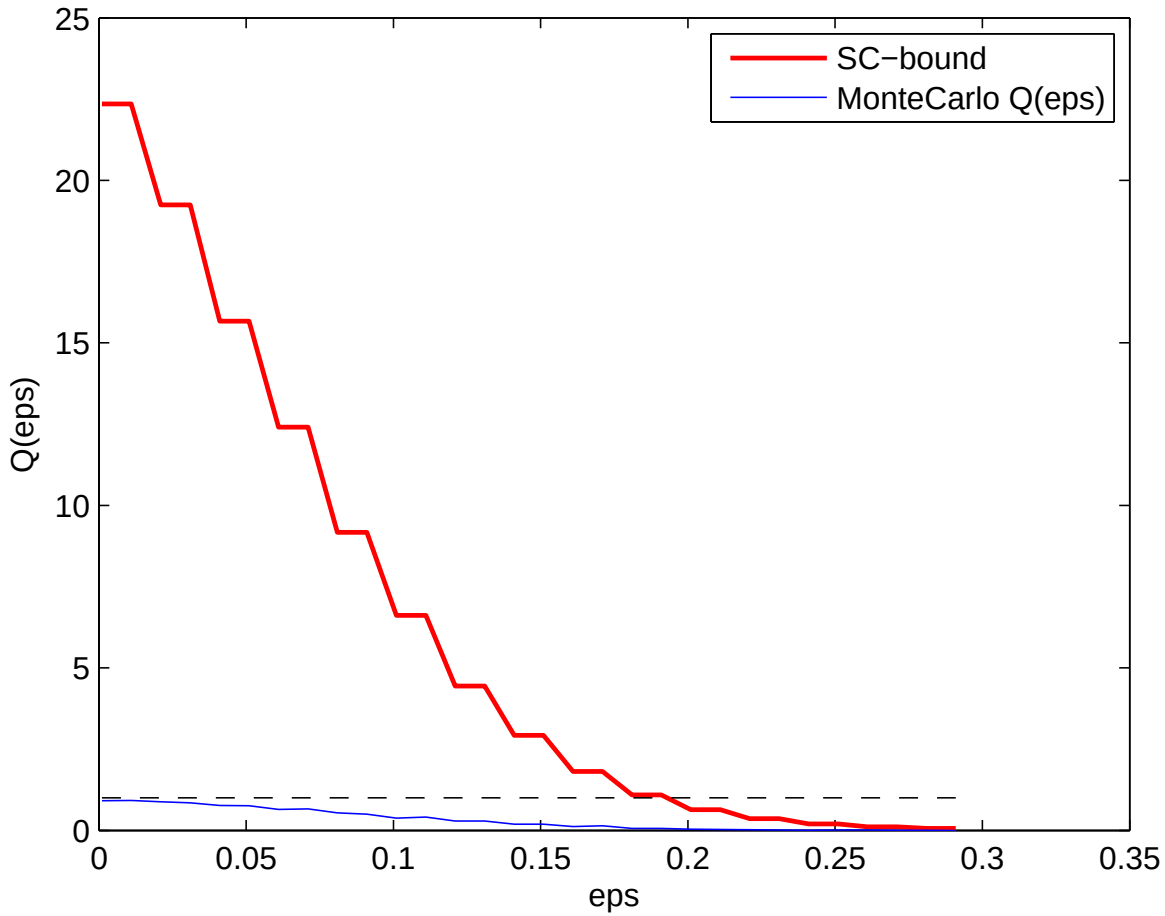
Эксперимент на модельных данных

Двумерная задача, $L = 100$, два разделимых класса (шум 0%).



Эксперимент на модельных данных

Двумерная задача, $L = 100$, шум 5%.



Верхние оценки средней частоты ошибок на контроле

Теорема

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

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

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.

Понятие логической закономерности

Задача классификации: $y_i = y^*(x_i)$ — класс объекта x_i.
Закономерность класса y — это предикат $r\colon \mathbb{X} \to \{0,1\}$,
выделяющий $(r(x) = 1)$ как можно больше объектов класса y:

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

и как можно меньше объектов всех остальных классов:

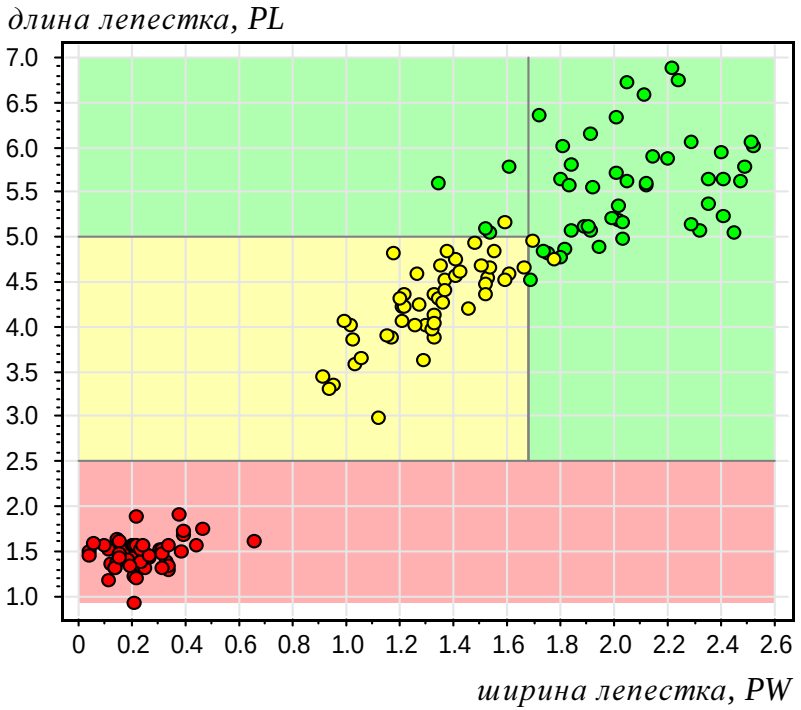
$$n(r, X) = \sum_{x_i \in X} r(x_i)\big[y_i \neq y\big] \;\to\; \min.$$

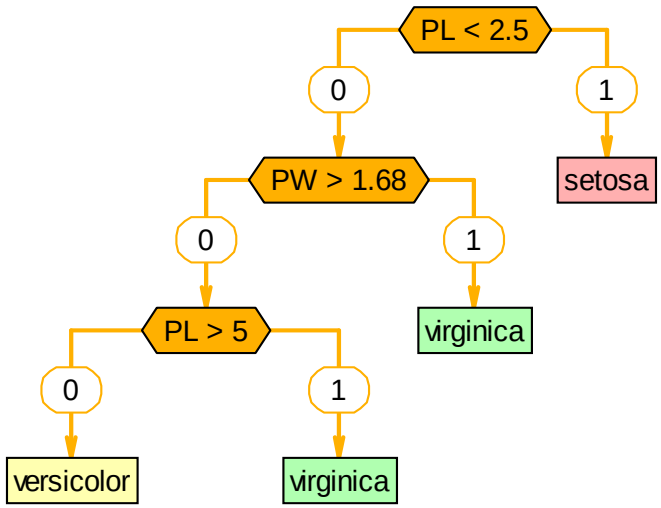
Логическая закономерность — конъюнкция пороговых условий:

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

где $f_j(x)$ — числовые признаки, θ_j — пороги, $j = 1, \ldots, n$;
$J \subseteq \{1, \ldots, n\}$ — подмножество признаков, обычно $|J| = 1..5$.

Пример. Закономерности в задаче с ирисами Фишера





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

Критерии информативности для поиска закономерностей

$\mathscr{H}(p,n) \to \max$

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

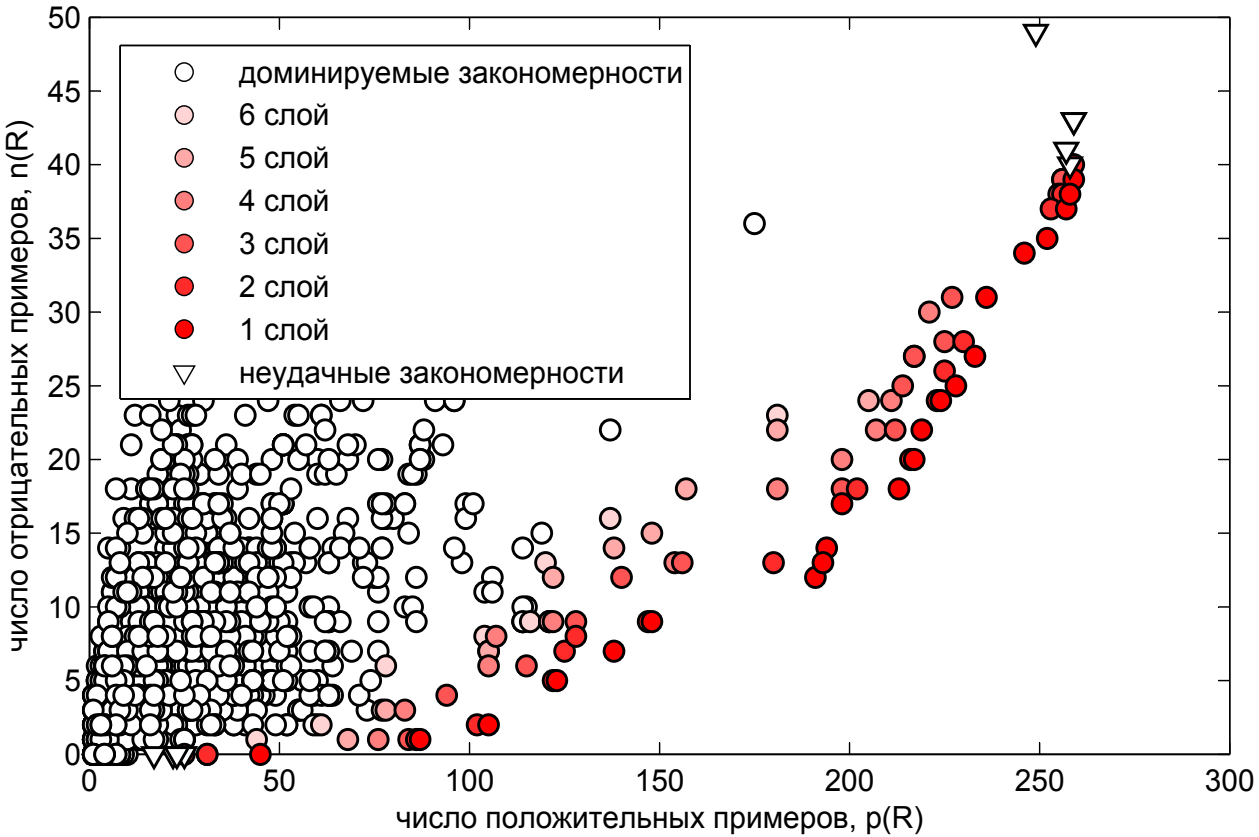
$$\mathscr{H}(p,n) = h\left(\frac{P}{\ell}\right) - \frac{p+n}{\ell}\, h\left(\frac{p}{p+n}\right) - \frac{\ell-p-n}{\ell}\, h\left(\frac{P-p}{\ell-p-n}\right),$$

  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: $\mathscr{H}(p,n) = -\log C_P^p C_N^n / C_{P+N}^{p+n}$;
- критерий бустинга: $\mathscr{H}(p,n) = \sqrt{p} - \sqrt{n}$;

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

Двухкритериальная оптимизация для поиска закономерностей

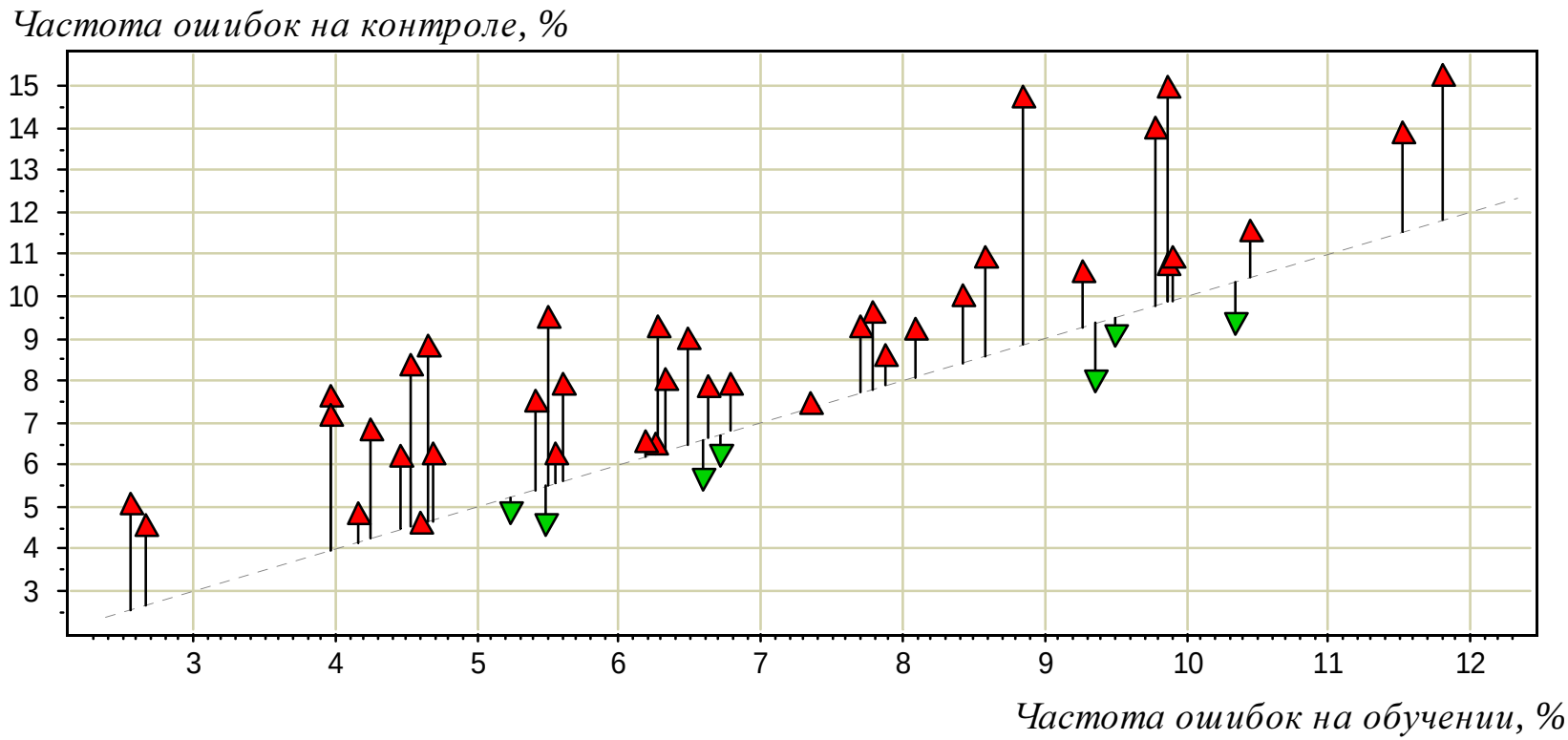
Парето-фронт — множество недоминируемых закономерностей (точка (p, n) недоминируема, если правее и ниже точек нет)



Задача кредитного скоринга german из репозитория UCI.

Пример. Переобучение закономерностей

Как отбросить переобученные закономерности на этапе обучения?



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,

$$\mathsf{P}\left[\tfrac{1}{k}n(r,\bar{X}) - \tfrac{1}{\ell}n(r,X) \geqslant \varepsilon\right] \leqslant \eta_n(\varepsilon);$$

$$\mathsf{P}\left[\tfrac{1}{\ell}p(r,X) - \tfrac{1}{k}p(r,\bar{X}) \geqslant \varepsilon\right] \leqslant \eta_p(\varepsilon);$$

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

$$n(r,\bar{X}) \leqslant \hat{n}(r,X) \equiv \tfrac{k}{\ell}n(r,X) + k\varepsilon_n(\eta);$$
$$p(r,\bar{X}) \geqslant \hat{p}(r,X) \equiv \tfrac{k}{\ell}p(r,X) - k\varepsilon_p(\eta).$$

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

Три проблемы

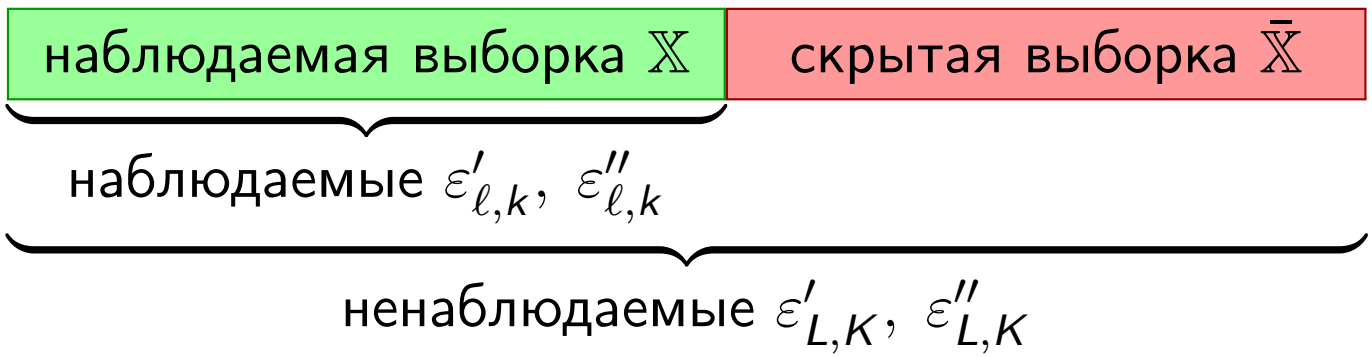
1. Как получить оценки для ошибок I и II рода?

   I рода: $I'(r, x_i) = [r(x_i) = 0]\,[y_i = y]$, $m' = P - p$, ν';

   II рода: $I''(r, x_i) = [r(x_i) = 1]\,[y_i \neq y]$, $m'' = n$, ν''.

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 μ: $\mathscr{H}(p,n) \to \max$, где $\mathscr{H}$ строго возрастает по p и строго убывает по n; любых $\mathbb{X}$, R, $\varepsilon \in (0,1)$

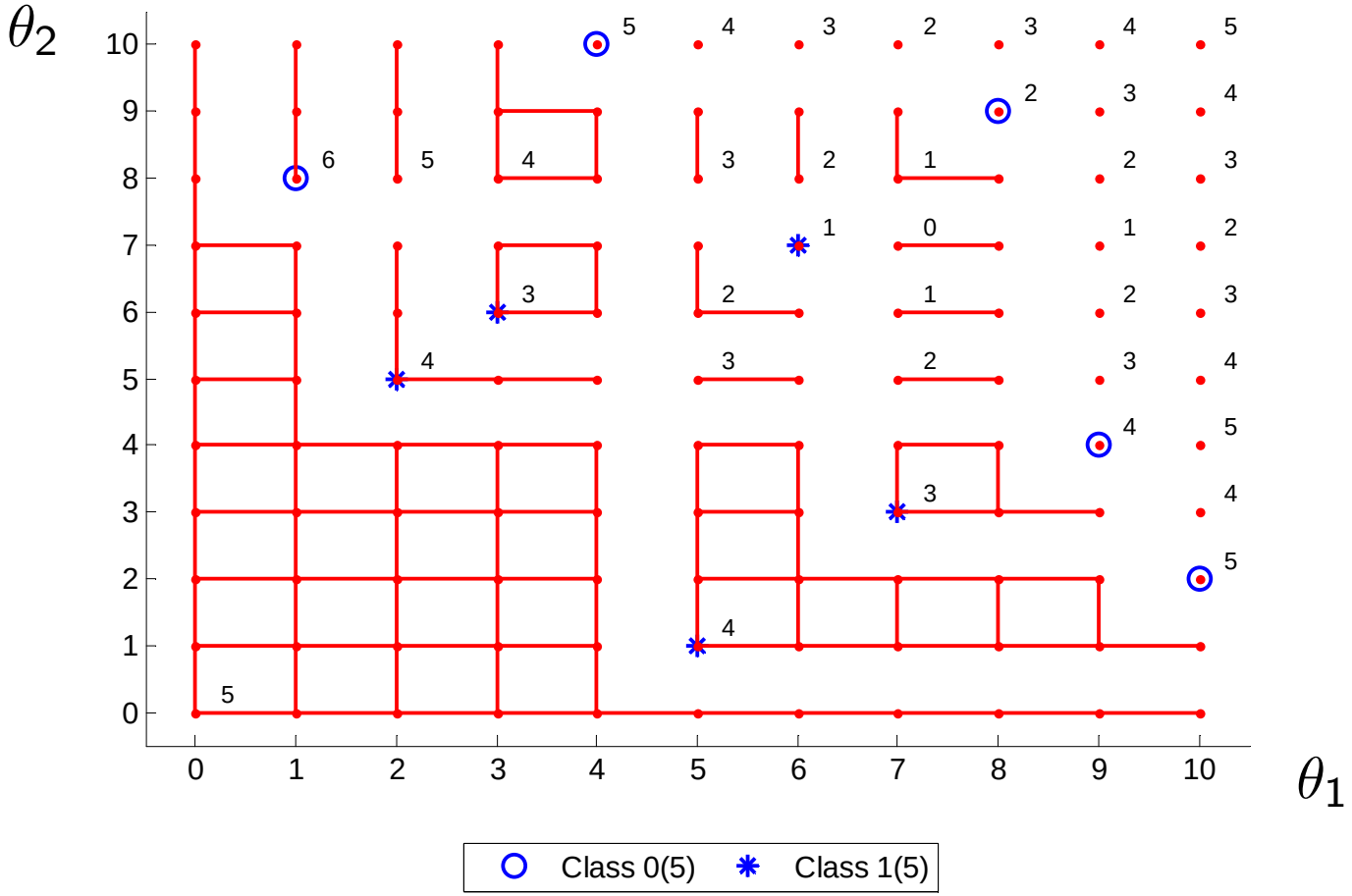
$$Q'_\varepsilon \leqslant \sum_{r\in R} \frac{C^{\ell-u}_{L-u-q}}{C^\ell_L} \mathcal{H}^{\ell-u,\,m'-q'}_{L-u-q}\left(\tfrac{\ell}{L}(m'-\varepsilon k)\right);$$

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

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''.

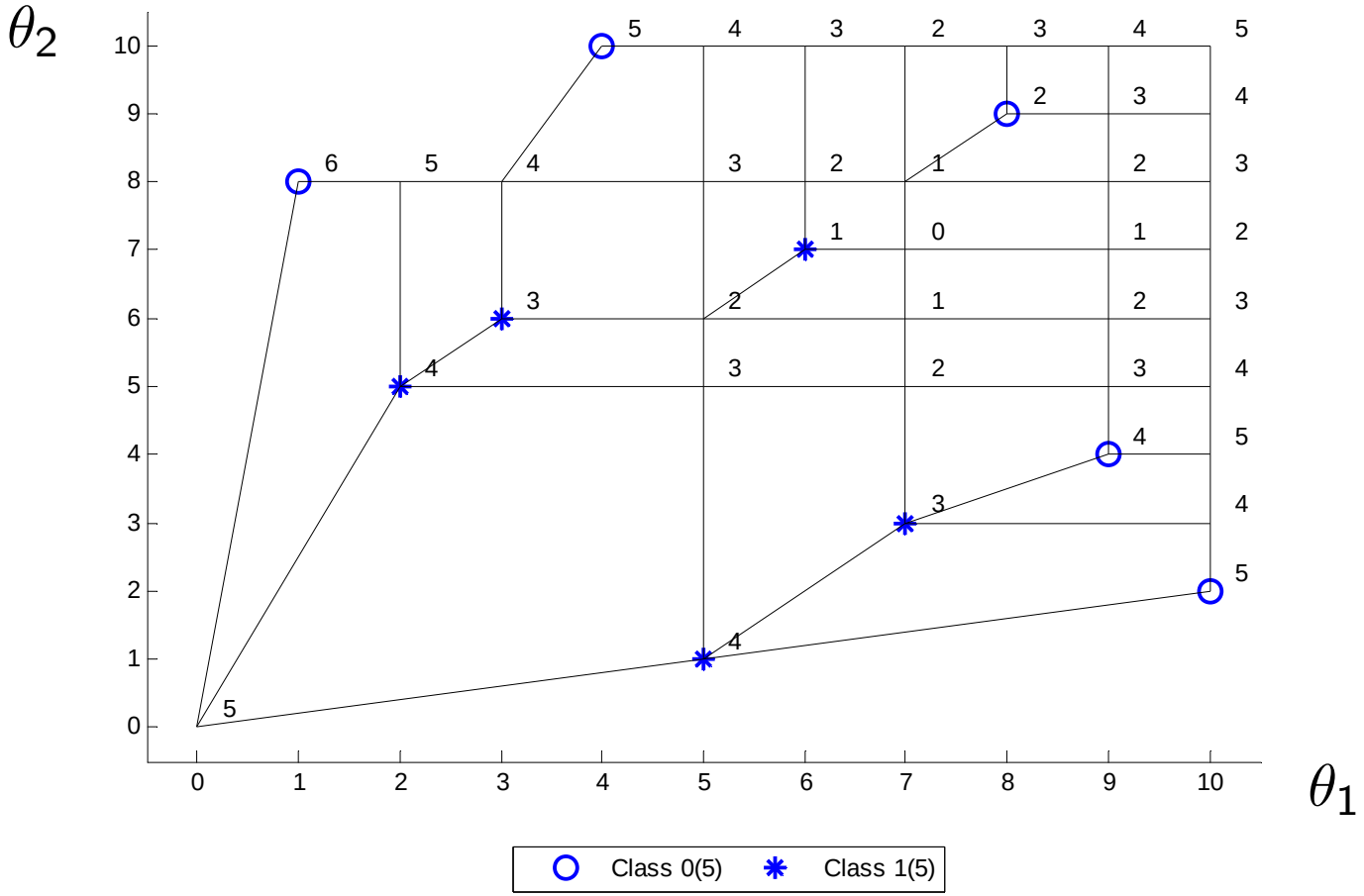
Проблема 2. Перебор конъюнктивных закономерностей

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



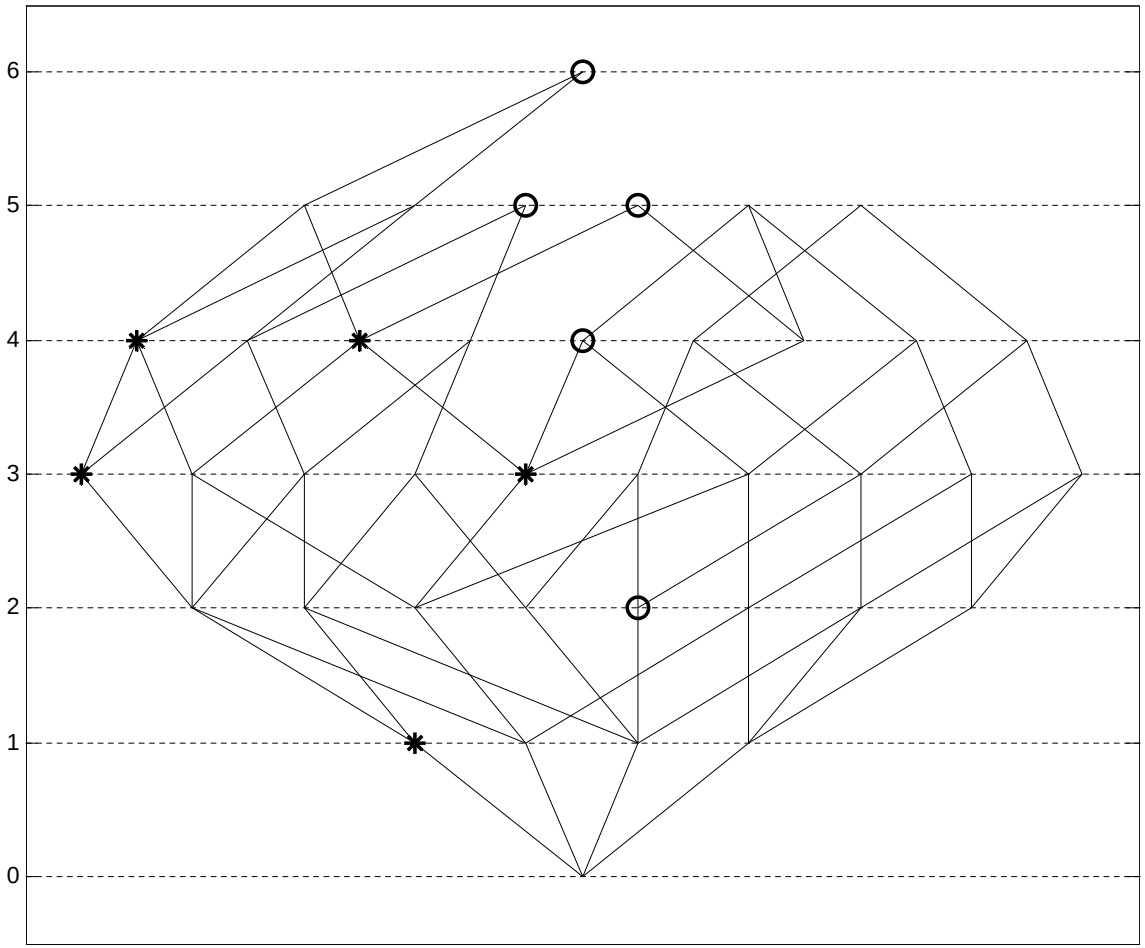
Проблема 2. Классы эквивалентности закономерностей

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



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:*

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

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

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

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

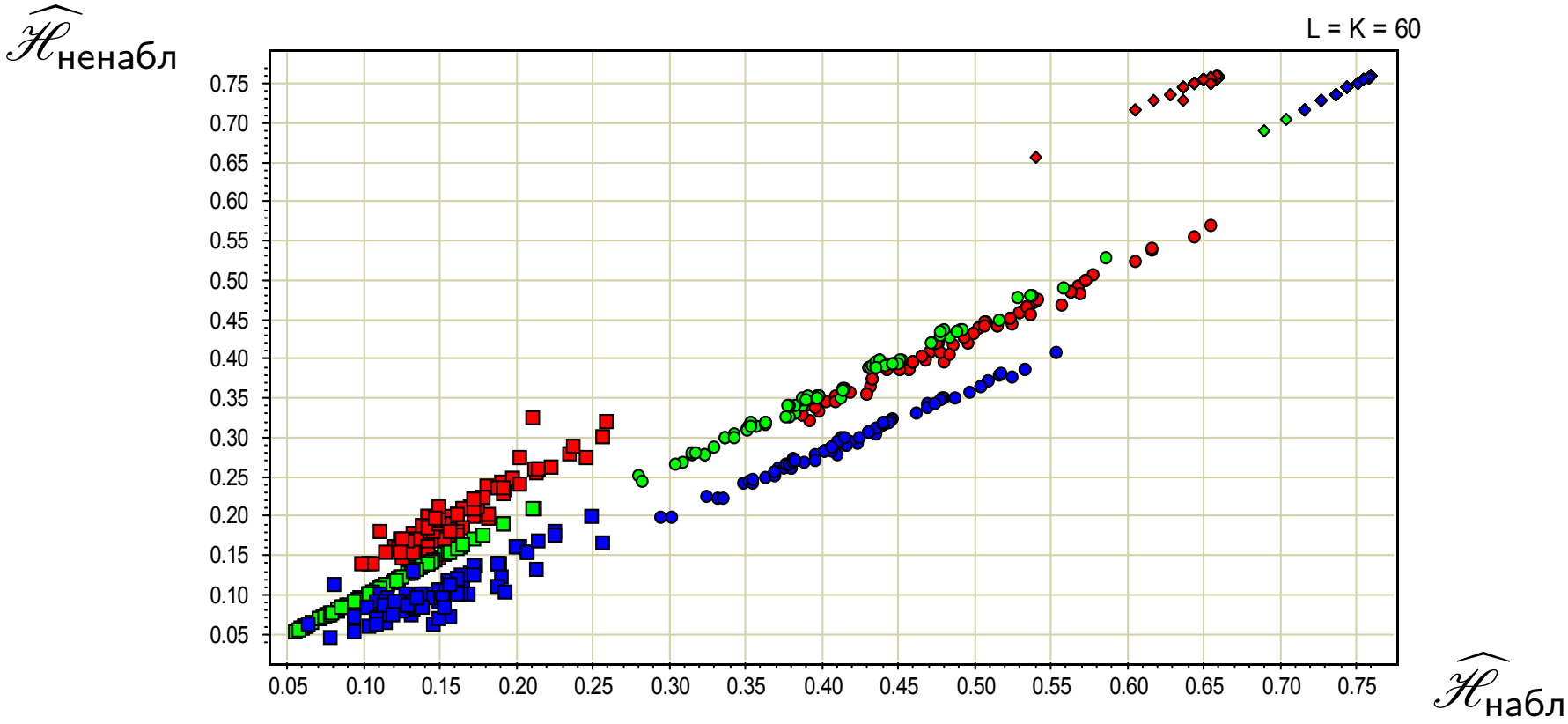
Эвристика: заменив $\varepsilon'_{L,K}$ на $\varepsilon'_{\ell,k}$ и $\varepsilon''_{L,K}$ на $\varepsilon''_{\ell,k}$, получим *наблюдаемый критерий предсказанной информативности:*

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

Почему мы имеем право на такую замену?

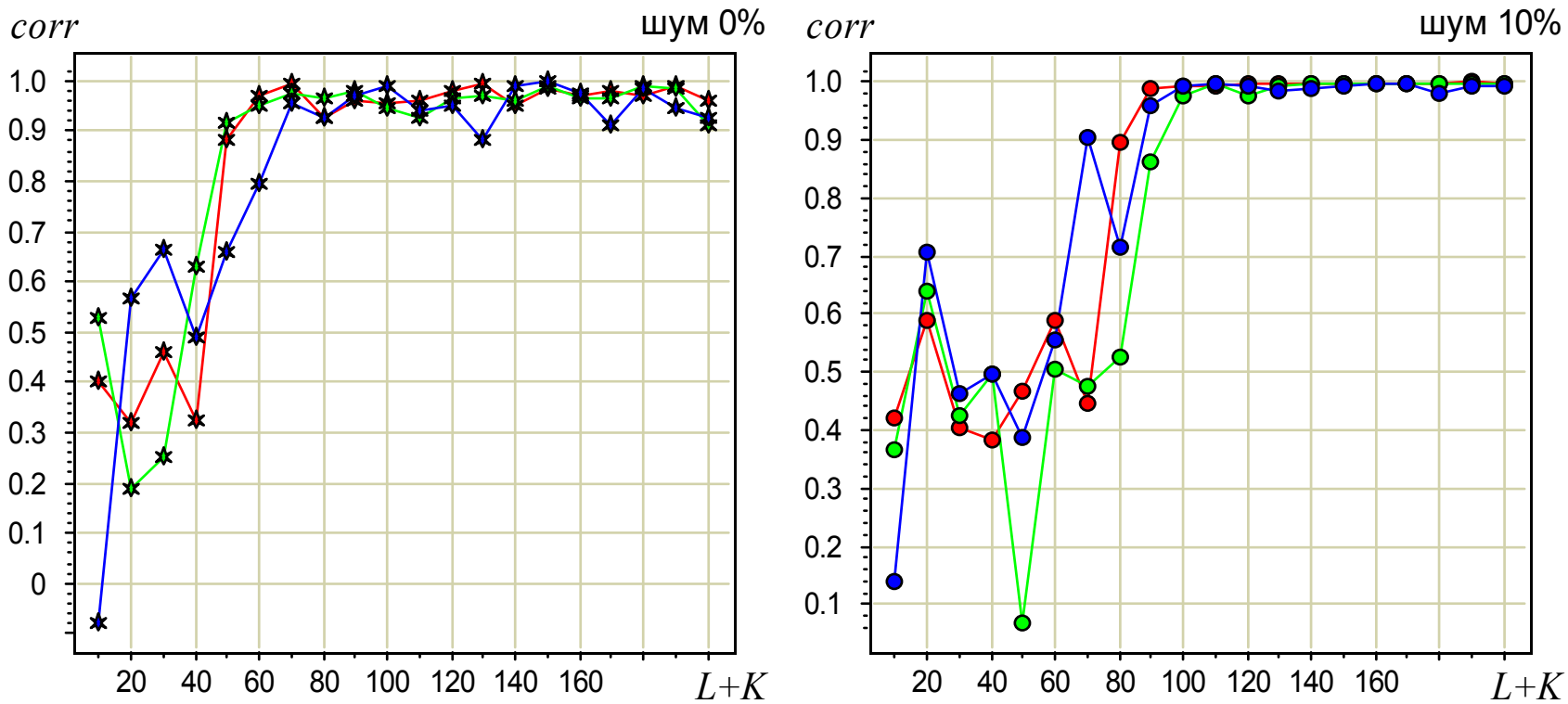
Эксперимент: зависимость $\widehat{\mathscr{H}}_{\text{ненабл}}$ от $\widehat{\mathscr{H}}_{\text{набл}}$

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 $\mathbb{X} \sqcup \bar{\mathbb{X}}$.



Эксперимент: корреляция $\widehat{\mathscr{H}}_{\text{ненабл}}$ и $\widehat{\mathscr{H}}_{\text{набл}}$

Зависимость корреляции $\mathscr{H}_{\text{ненабл}}$ и $\mathscr{H}_{\text{набл}}$ 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.

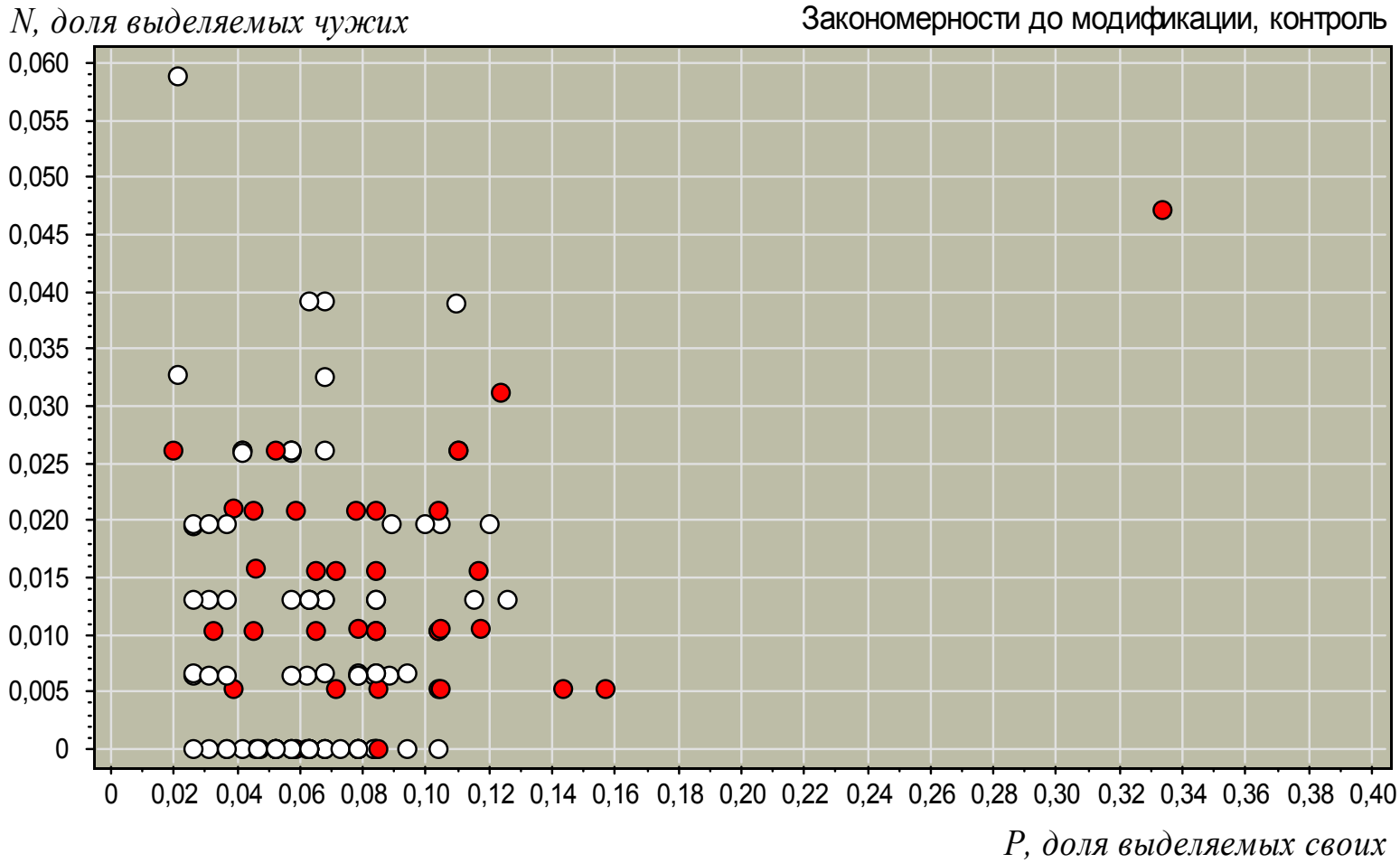
Результаты эксперимента на реальных данных

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.

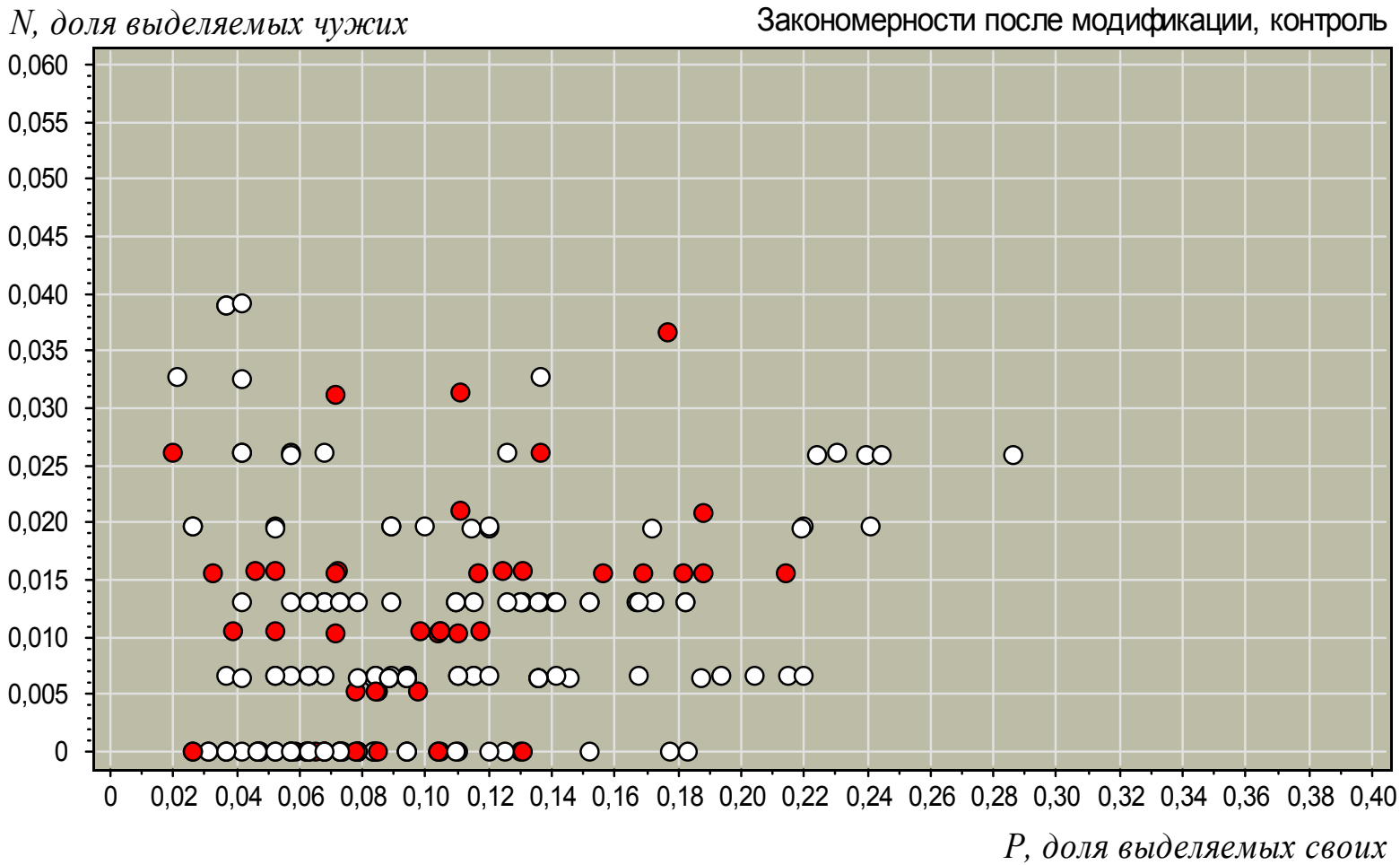
Закономерности в (p, n)-пространстве

Задача UCI:australian



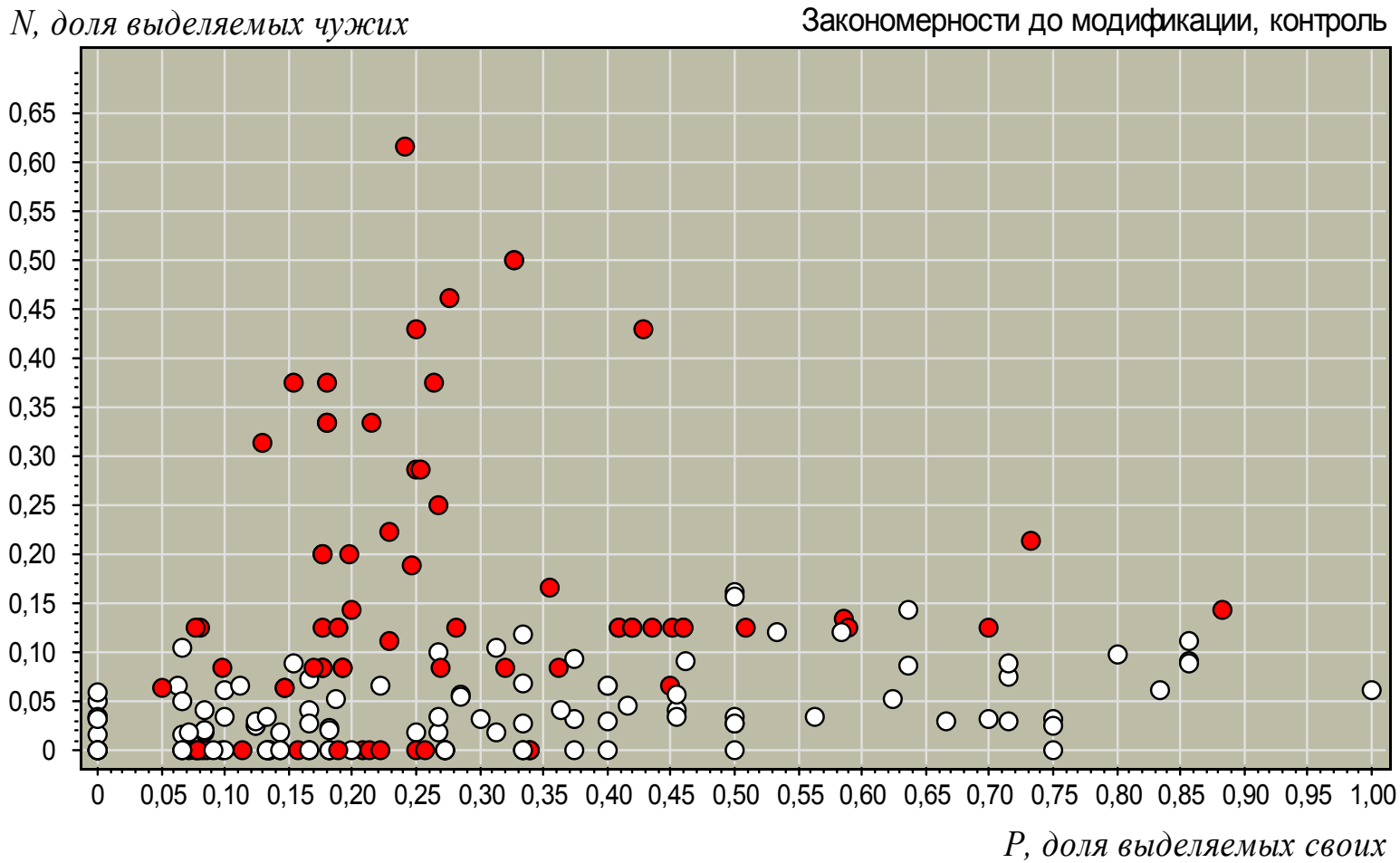
Закономерности в (p, n)-пространстве

Задача UCI:australian



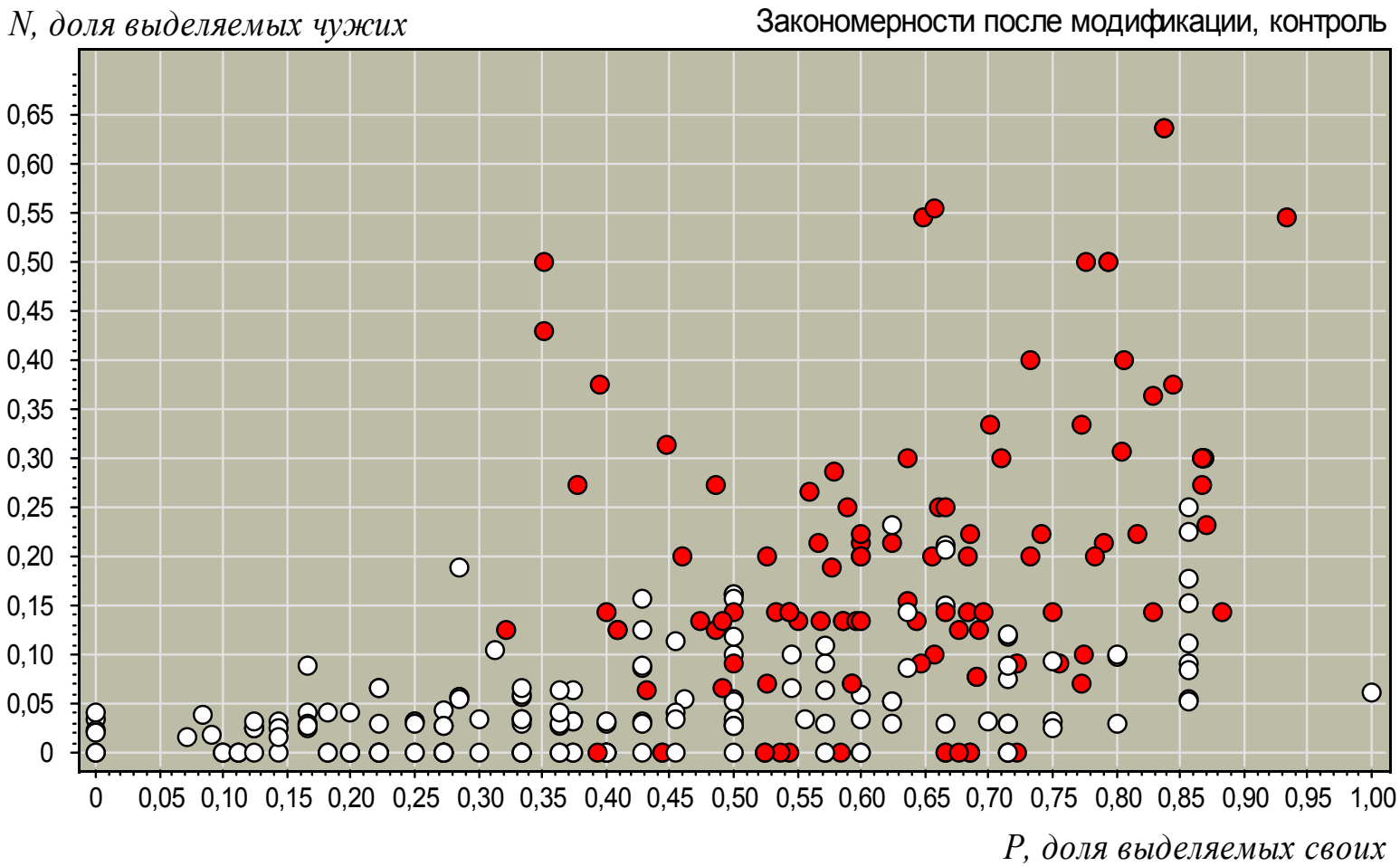
Закономерности в (p, n)-пространстве

Задача UCI:hepatitis



Закономерности в (p, n)-пространстве

Задача UCI:hepatitis



Обобщения метода порождающих и запрещающих множеств

Предыдущие оценки были основаны на том, что
$\forall\, a \in A \quad \exists\, (X_a, X'_a)$:

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

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

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

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

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

Второе обобщение очень удобно для оценивания CCV.

Классификатор ближайшего соседа (NN, nearest neighbor)

Пусть $\rho(x, x')$ — функция расстояния на множестве $\mathbb{X}$.

$$\mu\colon X \mapsto a, \qquad a(x) = y^*\big(\arg\min_{x' \in X} \rho(x, x')\big).$$

Определение (*профиль компактности* выборки $\mathbb{X}$)

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

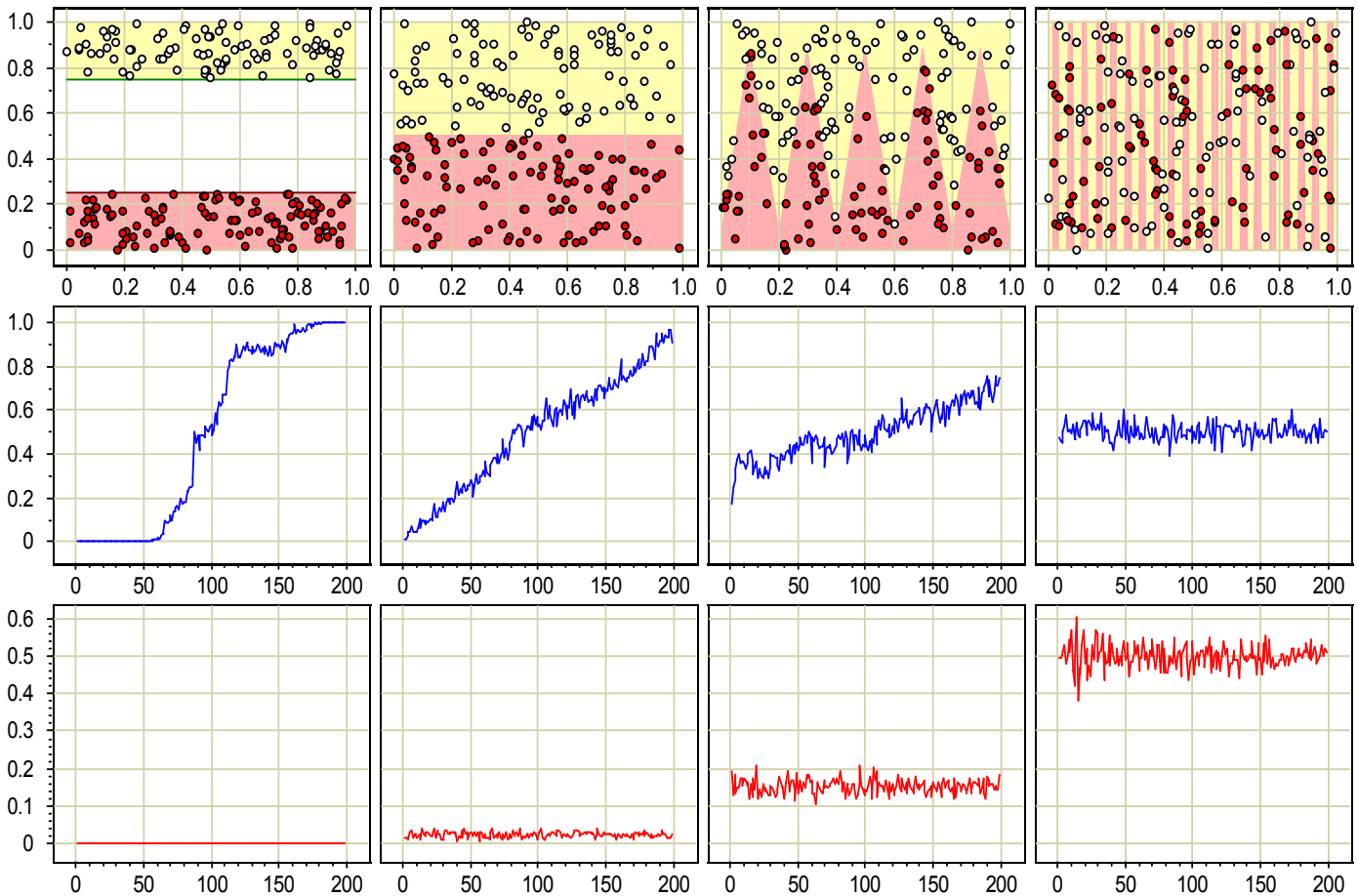
$$K(m, \mathbb{X}) = \frac{1}{L} \sum_{i=1}^{L} \big[y^*(x_{im}) \neq y_i\big]; \quad m = 1, \ldots, L-1,$$

Теорема (точная оценка для метода ближайшего соседа)

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

Профили компактности для серии модельных задач

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



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

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

Определение (*профиль компактности* относительно Ω)

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

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

Теорема (точное выражение CCV относительно Ω)

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

Жадные алгоритмы отбора эталонов

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

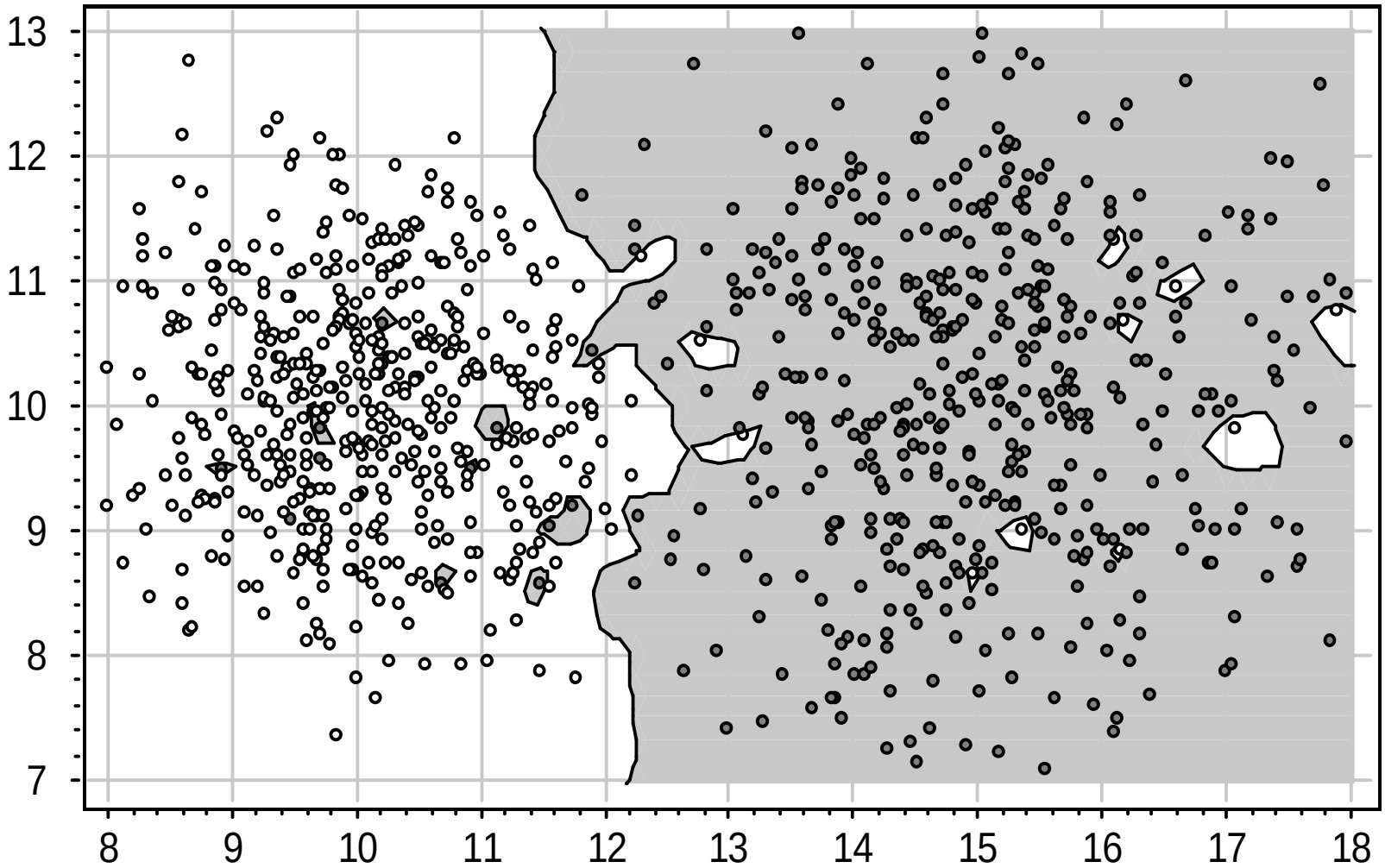
Жадный алгоритм удаления не-эталонов

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

Жадный алгоритм добавления эталонов

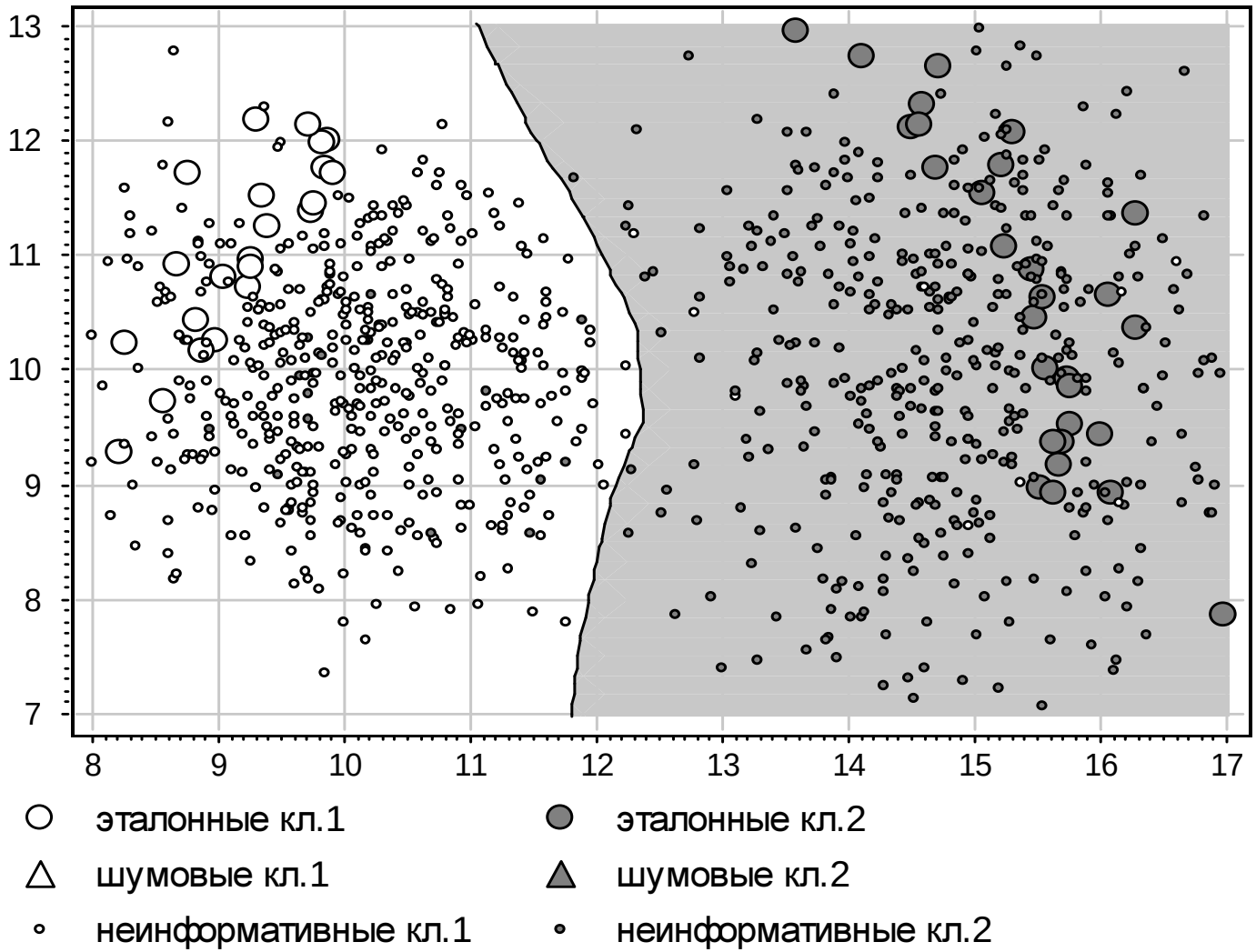
1: $\Omega := \{$по одному объекту от каждого класса$\}$;
2: **повторять**
3: найти $x \in \mathbb{X}\setminus\Omega$: $\mathrm{CCV}(\mu_{\Omega\cup\{x\}}, \mathbb{X}) \to \min$;
4: $\Omega := \Omega \cup \{x\}$;
5: **пока** CCV уменьшается;

Модельные данные

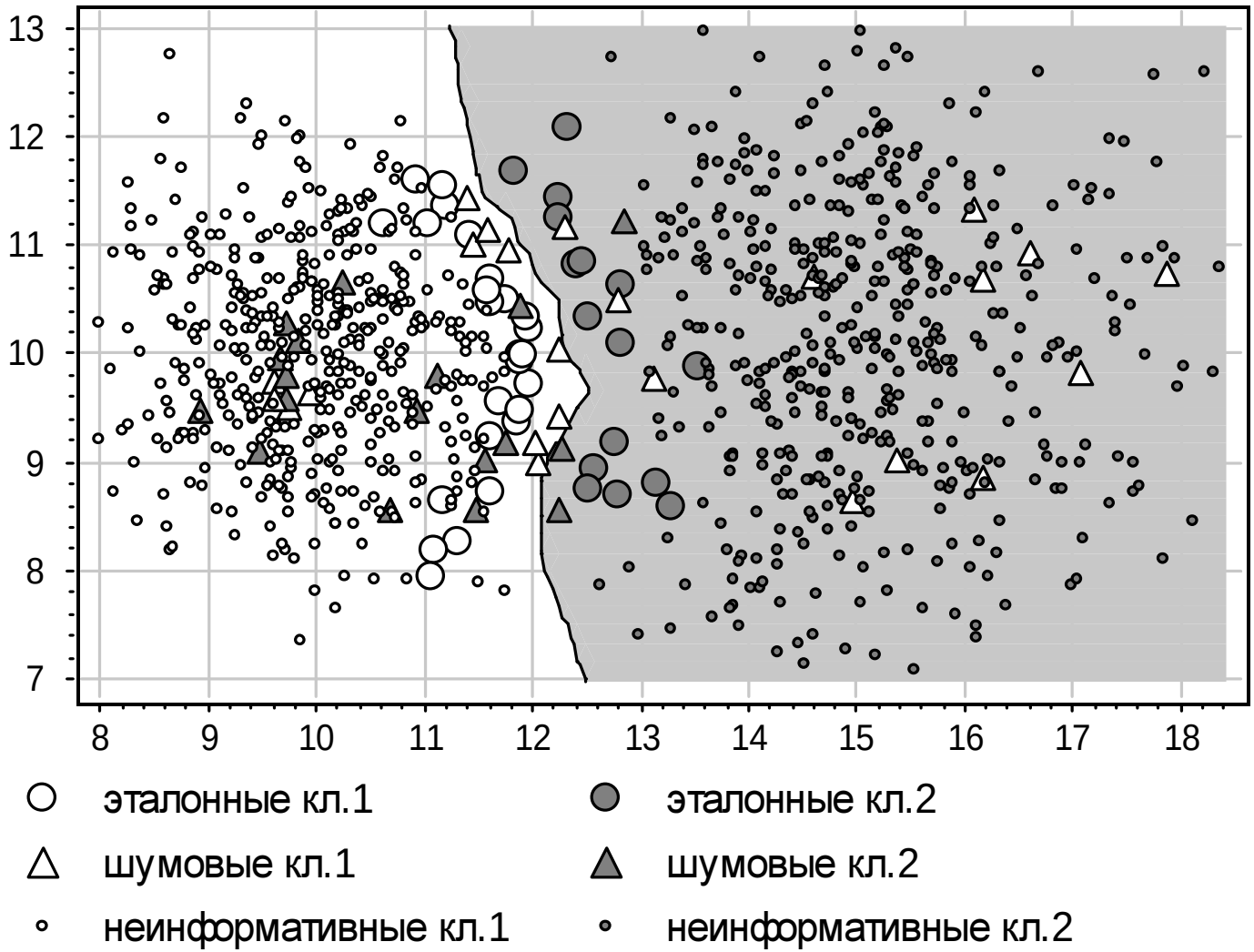


Модельная задача классификации: 1000 объектов, метод NN.

Жадное добавление эталонных объектов

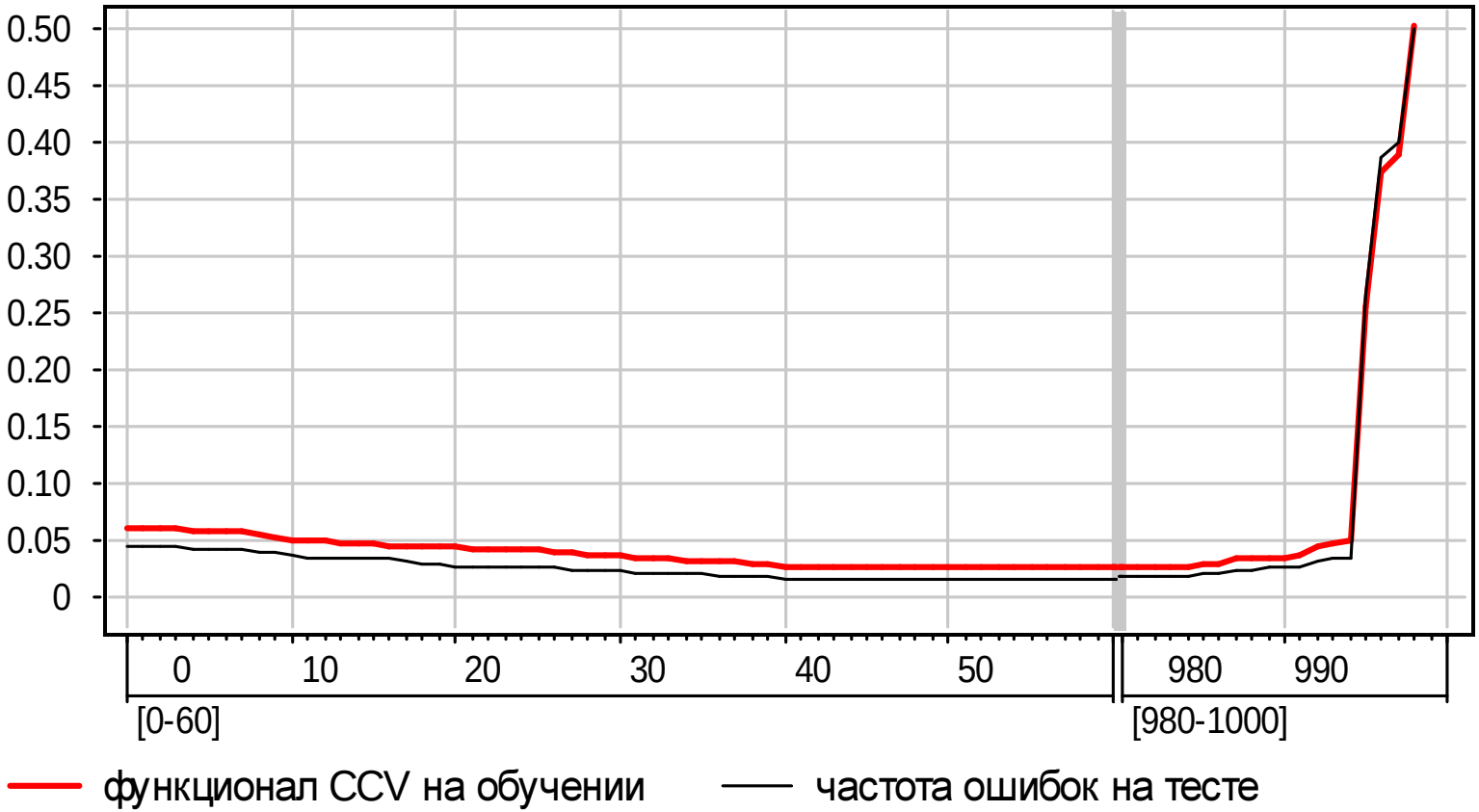


Жадное удаление не-эталонных объектов



Жадное удаление не-эталонных объектов

Зависимость CCV от числа удалённых неэталонных объектов.

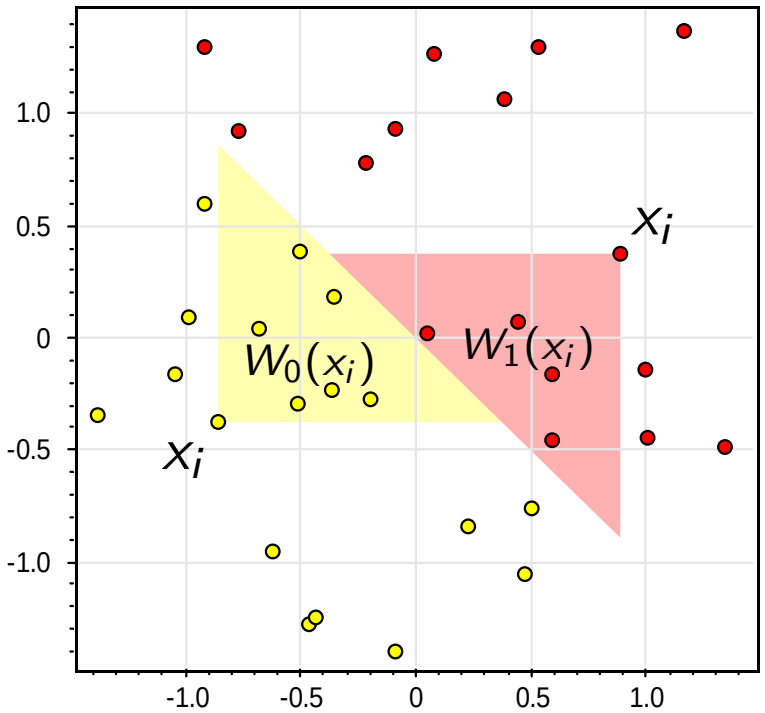


Чудо: при отборе эталонов переобучения вообще нет!

Монотонные алгоритмы классификации: определения

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

Верхний клин объекта x_i: $W_0(x_i) = \{x \in \mathbb{X}: x_i < x \text{ и } y(x) = 0\}$.
Нижний клин объекта x_i: $W_1(x_i) = \{x \in \mathbb{X}: x < x_i \text{ и } y(x) = 1\}$.



Профиль монотонности выборки

Определение (*профиль монотонности* выборки $\mathbb{X}$)

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

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

Теорема

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

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

Свойства профиля монотонности и оценки CCV
Выводы

- Невырожденность: $\mathrm{CCV}(\mu, \mathbb{X}) \leqslant 1$.
- Для минимизации CCV важен только начальный участок профиля, т. к. $\mathcal{H}_{L-1}^{\ell,\,m}(\theta L) \to 0$ 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).

Развитие комбинаторной теории переобучения

1. Оценки Q_ε и отбор признаков по критерию предсказанной информативности для пороговых логических правил [Андрей Ивахненко*]
2. Точные оценки Q_ε и CCV для многомерных сетей алгоритмов и их применение для обучения деревьев решений [Павел Ботов*]
3. Оценки расслоения–связности для линейных классификаторов [Денис Кочедыков*, Евгений Соколов]
4. Обобщение оценок расслоения–связности на случай произвольного графа Хассе [Илья Решетняк, Александр Фрей]
5. Уточнение оценок расслоения–связности с учётом конкуренции между несравнимыми алгоритмами [Евгений Соколов]
6. Комбинаторные оценки Радемахеровской сложности

Развитие комбинаторной теории переобучения

1. Оценки Q_ε для рандомизированных методов обучения и симметричных семейства на основе теории групп [Александр Фрей, Илья Толстихин]
2. Критерий точности оценок расслоения–связности [Никита Животовский]
3. Эмпирическое обоснование перехода от ненаблюдаемых оценок к наблюдаемым [Марина Дударенко]
4. Точные оценки CCV для метода k ближайших соседей и новые методы отбора эталонных объектов [Максим Иванов, Анастасия Зухба]
5. Верхние оценки CCV для монотонных классификаторов и новые методы обучения композиций классификаторов [Иван Гуз*, Галина Махина]
6. Обучение монотонного классификатора ближайшего соседа по критерию CCV в задачах каталогизации текстов

Открытые проблемы

1. **Обосновать**, что оценку расслоения–связности можно вычислять не по генеральному множеству $\mathbb{X}$, а по случайной наблюдаемой подвыборке X.
2. **Найти** способы быстрого пересчёта оценок при добавлении в выборку ещё одного объекта, ещё одного признака.
3. **Уточнять** оценки расслоения–связности с учётом конкуренции между алгоритмами с хэмминговым расстоянием, большим 1.
4. **Исследовать** эффекты расслоения и связности в случае произвольных (не бинарных) функций потерь.
5. **Разрабатывать** методы обучения с контролируемым переобучением.

Задача текущего этапа — переход от теории к алгоритмам, технологиям и приложениям.

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