Теоретические и прикладные аспекты статистического обучения

К. В. Воронцов

Конференция профессоров РАН
по Отделению математических наук РАН
14 июня 2016

Содержание

Задача статистического обучения с учителем

Этап *обучения* (training):
Дано: выборка $X = (x_i, y_i)_{i=1}^{\ell}$, $y_i = y(x_i)$, метод обучения μ
Найти: алгоритм $a = \mu(X)$, приближающий $y(x)$ вне X:

$$\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_1(x_\ell) \end{pmatrix} \xrightarrow{y} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix} \xrightarrow{\mu} a$$

Этап *применения* (testing):
Дано: тестовая выборка $X^k = (x'_i)_{i=1}^{k}$, алгоритм a
Найти: ответы на новых объектах x'_i:

$$\begin{pmatrix} f_1(x'_1) & \dots & f_n(x'_1) \\ \dots & \dots & \dots \\ f_1(x'_k) & \dots & f_n(x'_k) \end{pmatrix} \xrightarrow{a} \begin{pmatrix} a(x'_1) \\ \dots \\ a(x'_k) \end{pmatrix}$$

Модель алгоритмов и метод обучения

Модель алгоритмов — параметрическое семейство отображений

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

где g — заданная функция, θ — вектор параметров модели.

Эмпирический риск относительно функции потерь $\mathscr{L}(a, y)$:

$$Q(a, X) = \frac{1}{|X|} \sum_{x_i \in X} \mathscr{L}\big(a(x_i), y_i\big),$$

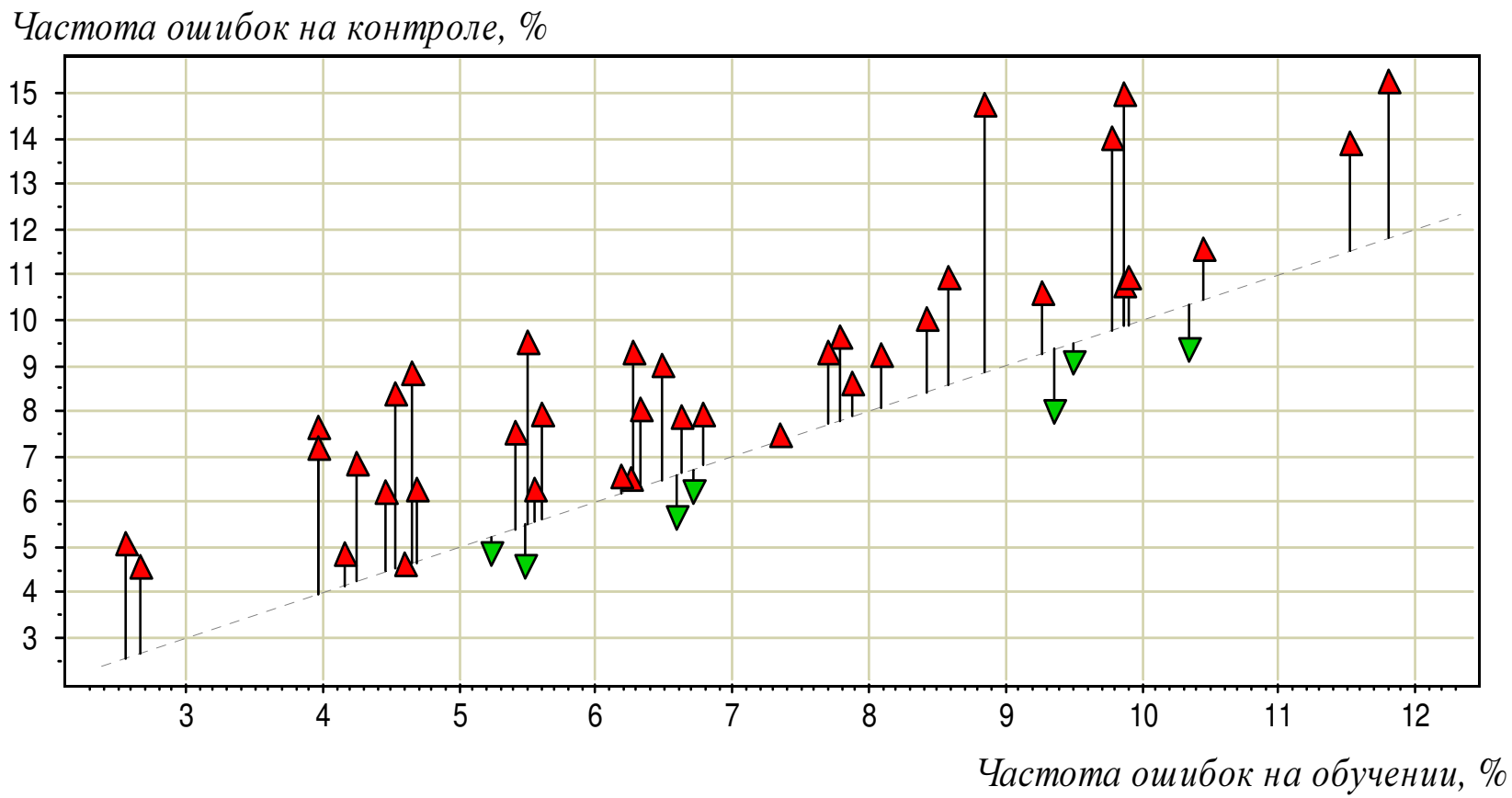
Минимизация эмпирического риска — метод обучения вида

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

Проблема переобучения: алгоритм $a = \mu(X)$ может плохо восстанавливать «закон природы» $y(x)$, допуская много ошибок на *контрольной выборке* $X^k = (x'_i, y'_i)_{i=1}^k$, $y'_i = y(x'_i)$

Пример переобучения. Реальная задача классификации

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



Бинарная функция потерь. Матрица ошибок

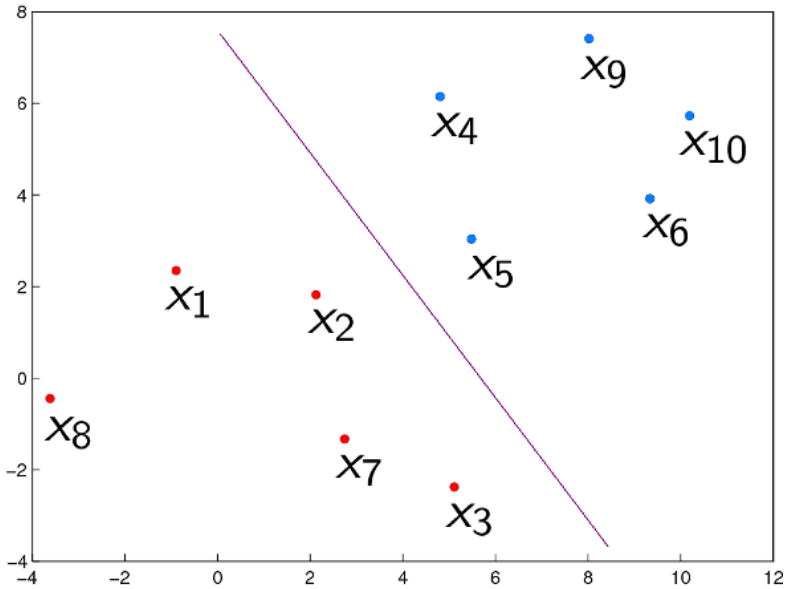
$X^L = \{x_1, \ldots, x_L\}$ — конечное *генеральное множество* объектов;
$A = \{a_1, \ldots, a_D\}$ — конечное семейство *алгоритмов*;
$I(a, x) = [$алгоритм a ошибается на объекте $x]$;

$L \times D$-*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^ℓ — наблюдаемая
$\ldots$	0	0	0	**0**	1	1	$\cdots$	1	(обучающая) выборка
x_ℓ	0	0	1	**0**	0	0	$\cdots$	0	объёма ℓ
$x_{\ell+1}$	0	0	0	**1**	1	1	$\cdots$	0	X^k — скрытая
$\ldots$	0	0	0	**1**	0	0	$\cdots$	1	(контрольная) выборка
x_L	0	1	1	**1**	1	1	$\cdots$	0	объёма $k = L - \ell$

$n(a, X) = \sum\limits_{x \in X} I(a, x)$ — *число ошибок* $a \in A$ на выборке $X \subset X^L$;
$\nu(a, X) = n(a, X)/|X|$ — *частота ошибок* a на выборке X;

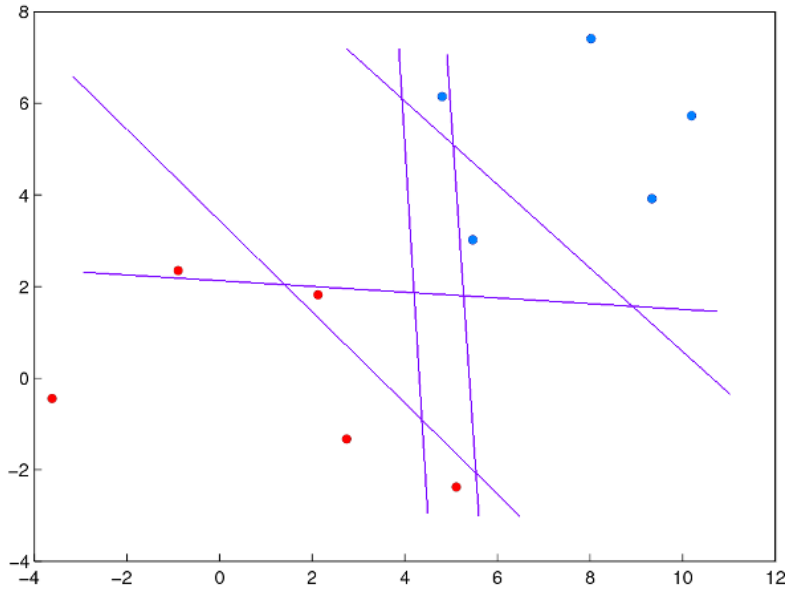
Пример. Матрица ошибок линейных классификаторов



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

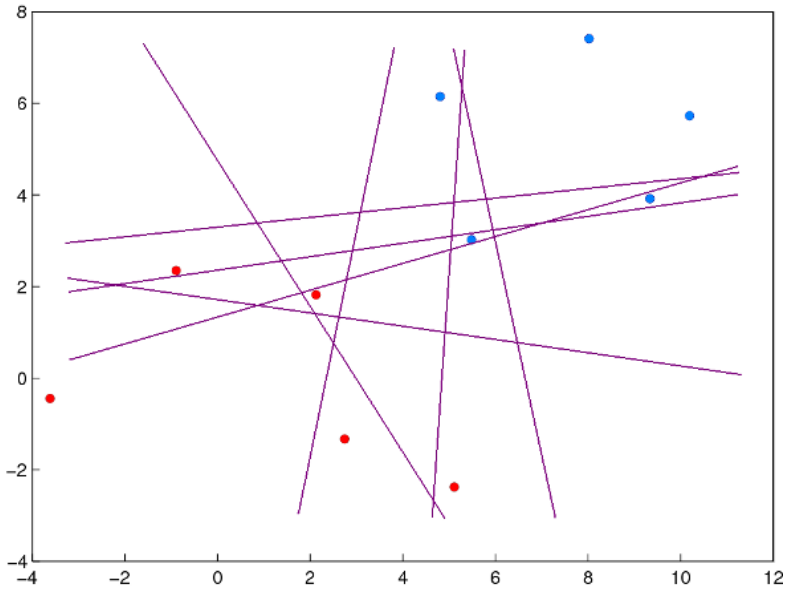


1 вектор с 0 ошибками

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	...

Основное вероятностное предположение

Все C_L^ℓ разбиений $X^\ell \sqcup X^k = X^L$ равновероятны.

В этом случае $\mathsf{P} \equiv \mathsf{E} \equiv \frac{1}{C_L^\ell} \sum_{X^\ell \subset X^L}$ — доля разбиений выборки.

Функционалы обобщающей способности

- ожидаемая частота ошибок на контроле:
$$\mathrm{CCV}(\mu, X^L) = \mathsf{E}\,\nu\big(\mu(X^\ell), X^k\big).$$
- вероятность большой частоты ошибок на контроле:
$$R_\varepsilon(\mu, X^L) = \mathsf{P}\big[\nu\big(\mu(X^\ell), X^k\big) \geqslant \varepsilon\big].$$
- вероятность переобучения:
$$Q_\varepsilon(\mu, X^L) = \mathsf{P}\big[\delta(\mu(X^\ell), X^\ell, X^k) \geqslant \varepsilon\big],$$
$\delta(a, X^\ell, X^k) = \nu(a, X^k) - \nu(a, X^\ell)$ — *переобученность* a.

Отличия от классической постановки задачи

Функционалы обобщающей способности:

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

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

- Вероятность переобучения

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

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.

Пара 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.»*

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.»

Теория Вапника–Червоненкиса

Теорема (В. Н. Вапник, А. Я. Червоненкис, 1971)

Для любых X^L, A, μ и $\varepsilon \in [0,1]$, при $\ell = k$

$$S_\varepsilon(A, X^L) \leqslant |A| \cdot \tfrac{3}{2} \exp\bigl(-\varepsilon^2 \ell\bigr).$$

Проблема завышенности:

- эта оценка завышена в 10^8–10^{11} раз;
- что приводит к оценкам длины обучения $\ell = 10^6$–10^{10}, когда на самом деле достаточно $\ell = 10^2$–10^3.

Причина завышенности — это оценка «худшего случая»:

- она зависит только от размеров матрицы ошибок $L \times D$;
- не зависит от её содержимого $I(a,x)$, выборки X^L, метода μ.

Теория Вапника–Червоненкиса в комбинаторной постановке

Основные шаги доказательства:

Для любых X^L, A, μ и $\varepsilon \in [0,1]$

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

$$\leqslant |A| \cdot \max_a Q_\varepsilon(a, X^L) \leqslant$$

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

Почему эта оценка так сильно завышена?

1) uniform bound завышена, когда A *расслаивается* по $n(a, X^L)$
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]
- Oracle inequalities [Koltchinskii, 2011]

Эксперименты с модельными семействами алгоритмов

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

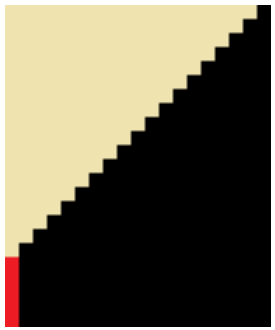
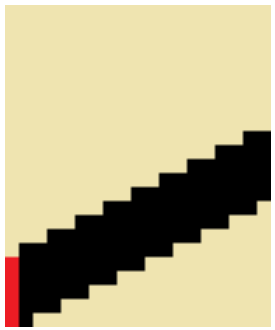
Выясним в экспериментах на модельных данных, как структура матрицы ошибок влияет на переобучение.

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

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

Эксперимент. Переобучение монотонных цепей

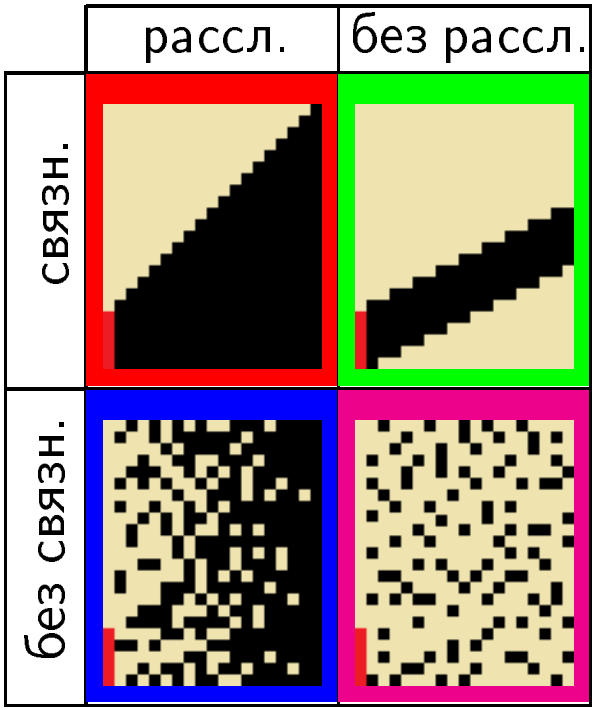
Четыре матрицы ошибок; лучший алгоритм у всех одинаков:

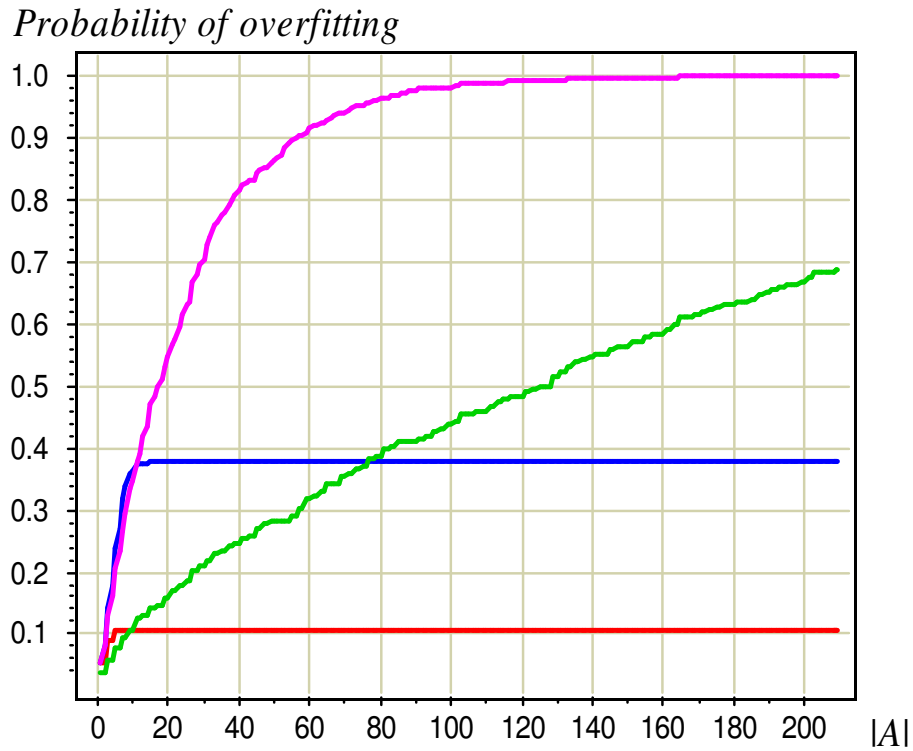
	с расслоением по числу ошибок	без расслоения по числу ошибок
каждая пара соседних алгоритмов отличается только на одном объекте (образуется *цепь*)		
соседние алгоритмы существенно различны, (*цепь* не образуется)		

Постепенно добавляя алгоритмы в $\{a_1, \ldots, a_D\}$, построим зависимости вероятности переобучения Q_ε от числа D.

Эксперимент. Переобучение монотонных цепей

$\ell = k = 100$, $\varepsilon = 0.05$, $N = 1000$ разбиений Монте-Карло.

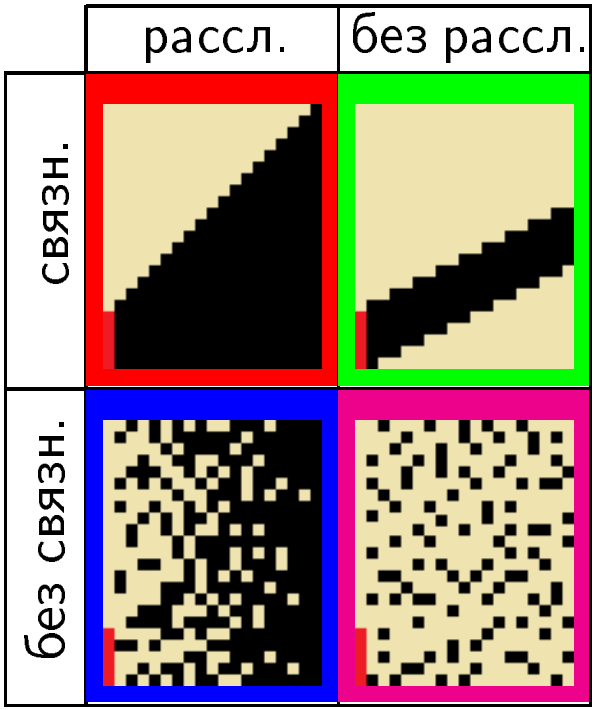


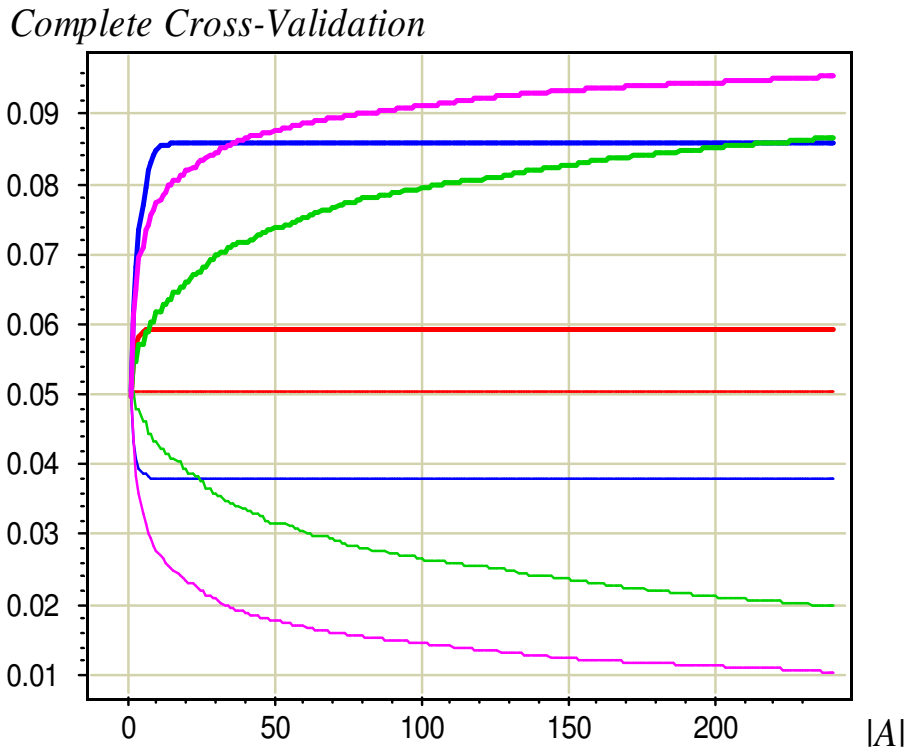


- Огромные семейства с P&C могут почти не переобучаться
- Без P&C даже 30 алгоритмов могут сильно переобучаться

Эксперимент. Переобучение монотонных цепей

$\ell = k = 100$, $\varepsilon = 0.05$, $N = 1000$ разбиений Монте-Карло.





$\mathsf{E}\,\nu\big(\mu(X^\ell), X^k\big)$ — CCV, жирные линии

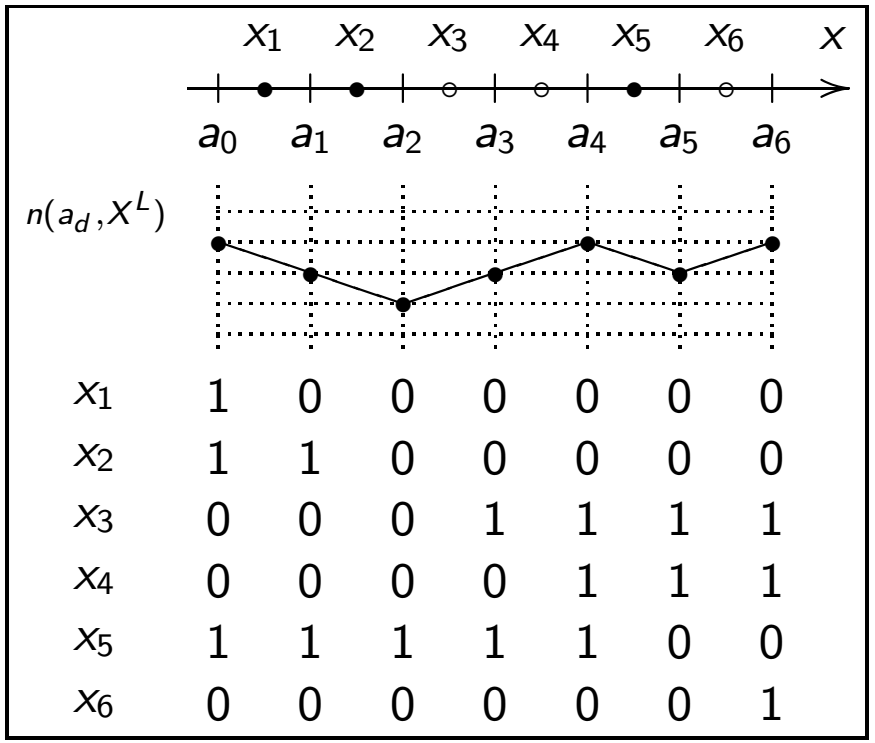
$\mathsf{E}\,\nu\big(\mu(X^\ell), X^\ell\big)$ — тонкие линии

Цепь, порождаемая семейством пороговых классификаторов

Пусть $f(x)$ — вещественный признак,

$$A = \{a(x) = [f(x) \geqslant \theta] : \theta \in \mathbb{R}\}.$$

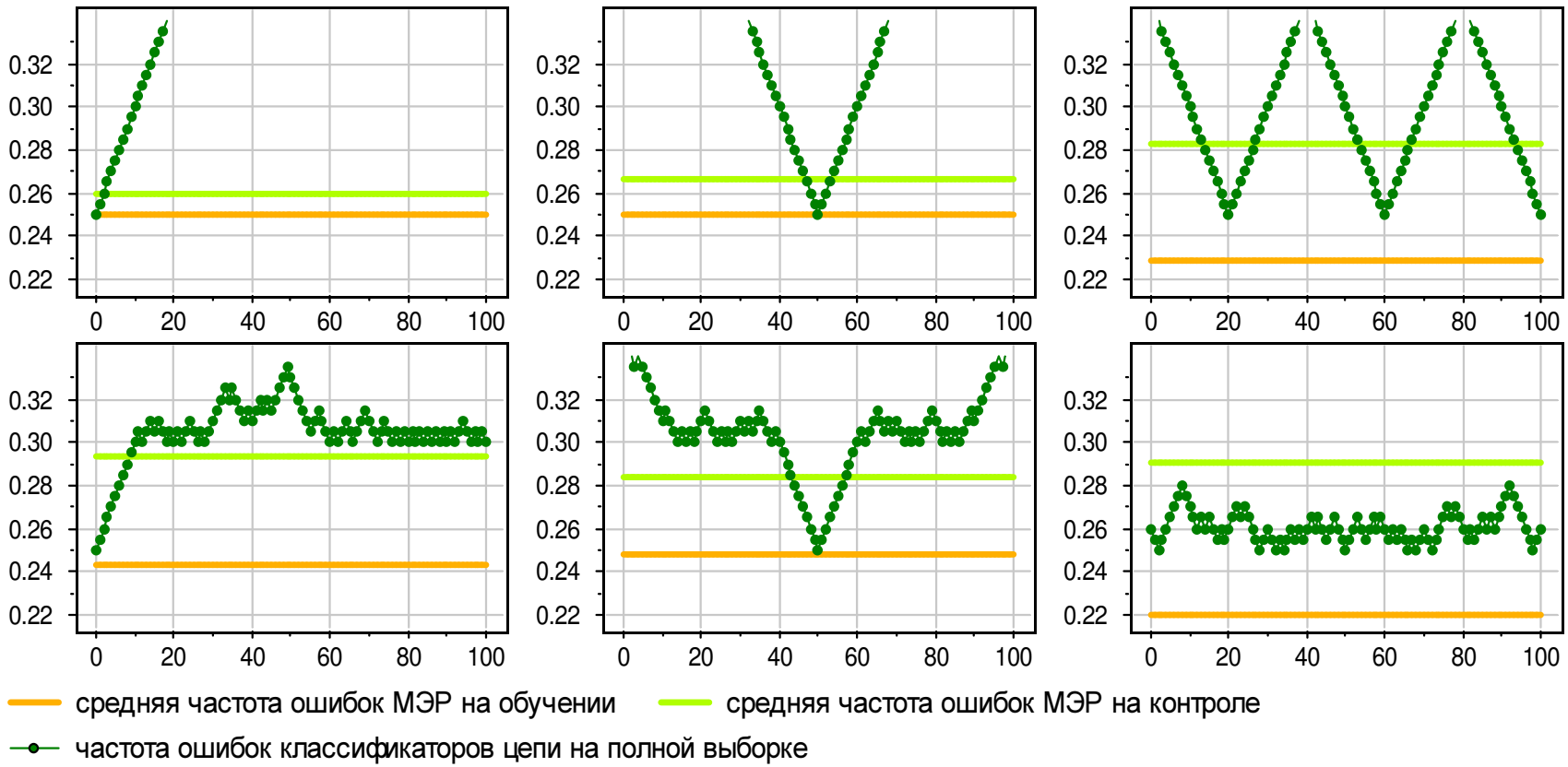
Пример:



x_1	1	0	0	0	0	0	0
x_2	1	1	0	0	0	0	0
x_3	0	0	0	1	1	1	1
x_4	0	0	0	0	1	1	1
x_5	1	1	1	1	1	0	0
x_6	0	0	0	0	0	0	1

Эксперимент. Переобучение цепей с различным расслоением

Условия эксперимента: $L = 100$, $\ell = 50$, $m = 25$, $\varepsilon = 0.05$, метод Монте-Карло по $N = 100000$ случайных разбиений.



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 $x \in X^L$;

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

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

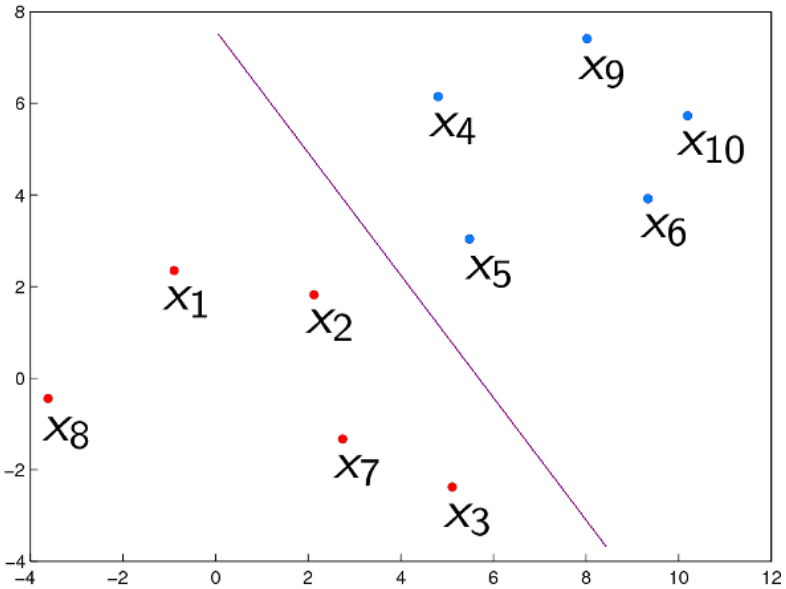
A — множество попарно различных векторов ошибок;

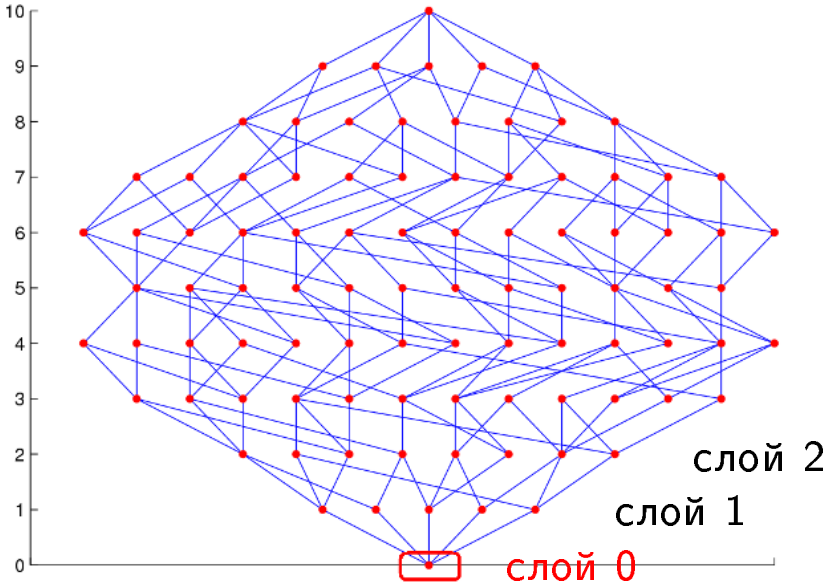
$E = \{(a,b)\colon a \prec b\}$.

Свойства графа расслоения–связности:

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

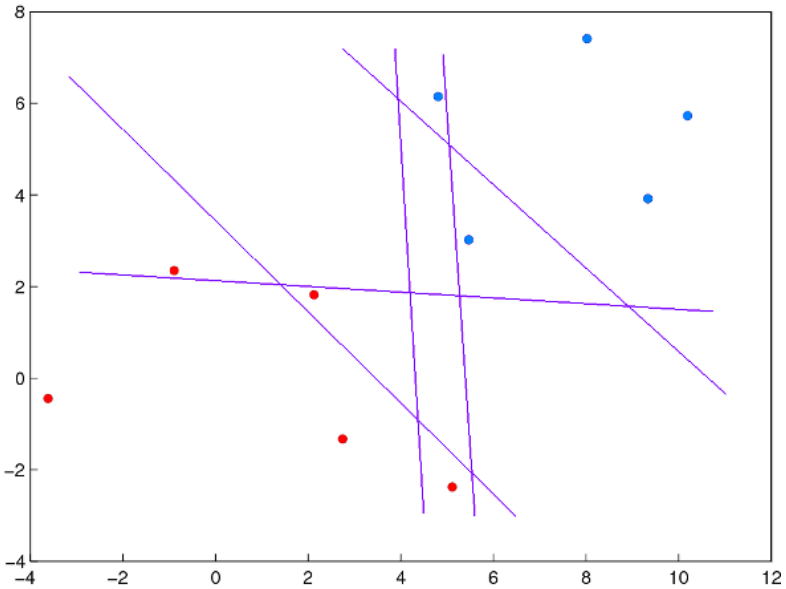
Пример 1. Семейство линейных алгоритмов классификации

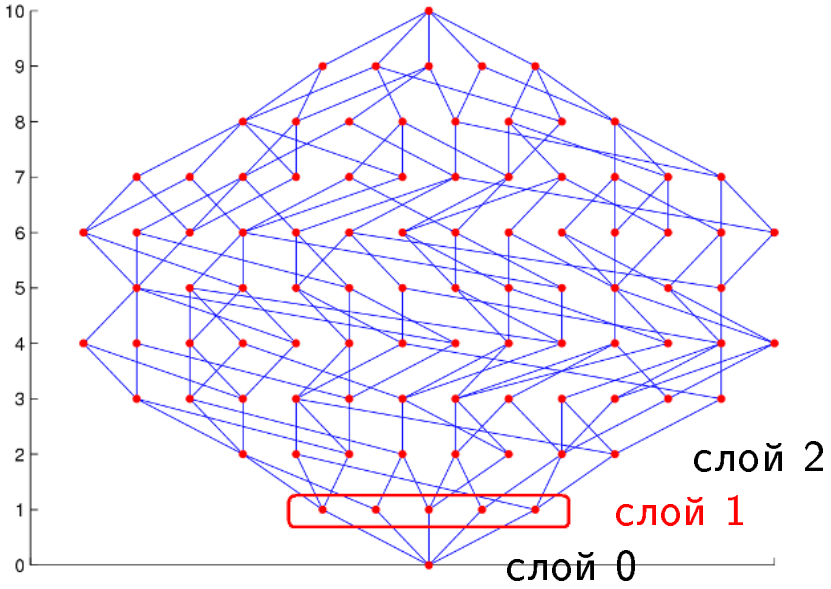




	слой 0
x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0

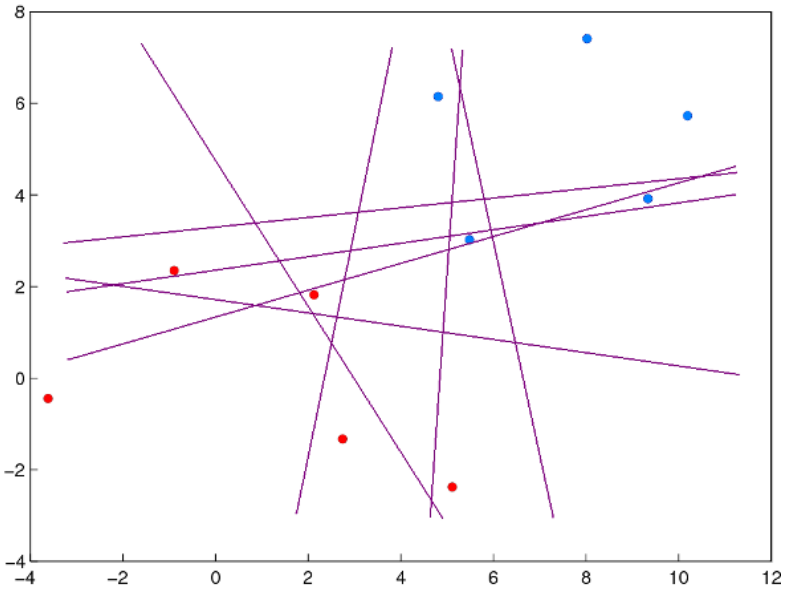
Пример 1. Семейство линейных алгоритмов классификации

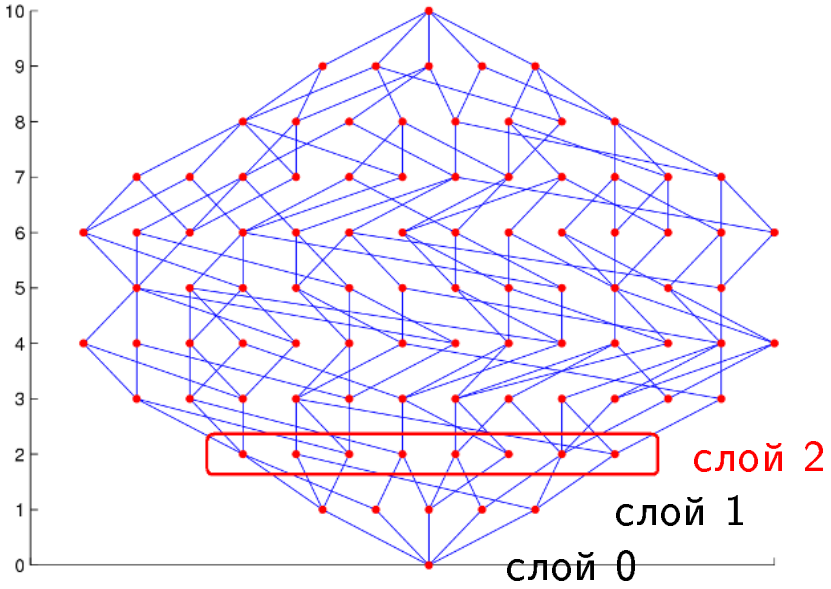




	слой 0	слой 1				
x_1	0	1	0	0	0	0
x_2	0	0	1	0	0	0
x_3	0	0	0	1	0	0
x_4	0	0	0	0	1	0
x_5	0	0	0	0	0	1
x_6	0	0	0	0	0	0
x_7	0	0	0	0	0	0
x_8	0	0	0	0	0	0
x_9	0	0	0	0	0	0
x_{10}	0	0	0	0	0	0

Пример 1. Семейство линейных алгоритмов классификации





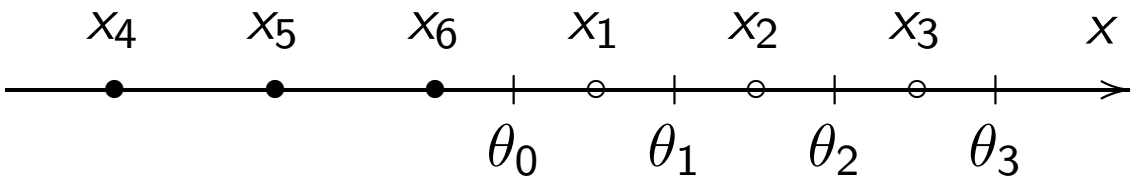
	слой 0	слой 1					слой 2								
x_1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	...
x_2	0	0	1	0	0	0	1	1	0	0	0	0	0	0	...
x_3	0	0	0	1	0	0	0	1	1	0	0	0	0	1	...
x_4	0	0	0	0	1	0	0	0	1	1	0	0	0	0	...
x_5	0	0	0	0	0	1	0	0	0	1	1	1	0	0	...
x_6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	...
x_7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	...
x_8	0	0	0	0	0	0	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	...

Пример 2. Монотонная цепь

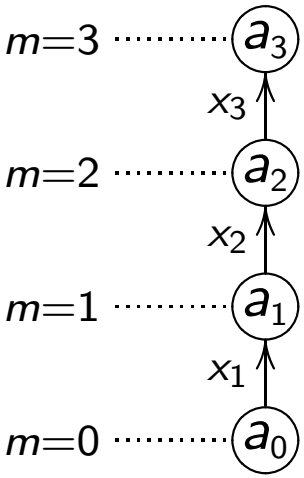
Опр. *Монотонная цепь* алгоритмов: $a_0 \prec a_1 \prec \cdots \prec a_D$.

Пример: пороговый классификатор $a_d(x) = \big[f(x) - \theta_d\big]$;

2 класса $\{\bullet, \circ\}$
6 объектов



Граф семейства:



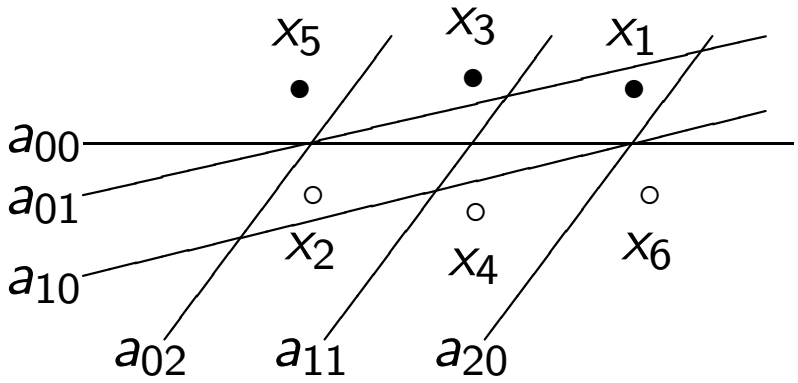
Матрица ошибок:

	a_0	a_1	a_2	a_3
x_1	0	1	1	1
x_2	0	0	1	1
x_3	0	0	0	1
x_4	0	0	0	0
x_5	0	0	0	0
x_6	0	0	0	0

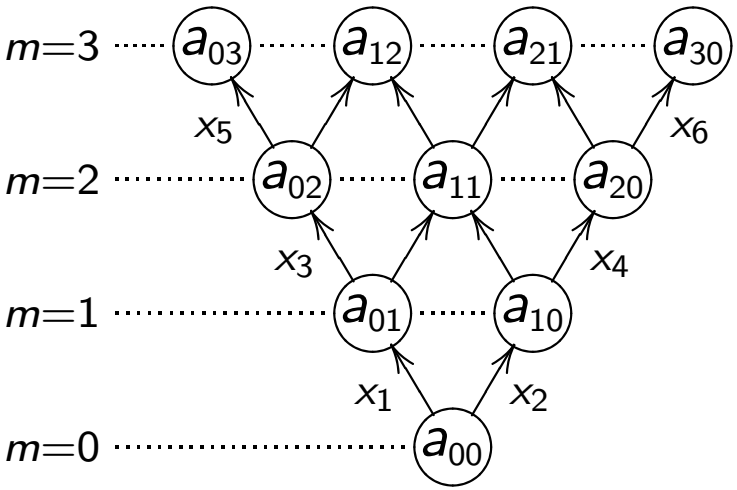
Пример 3. Двумерная сеть классификаторов

Пример:

2D линейный классификатор,
2 класса $\{\bullet, \circ\}$,
6 объектов



Граф семейства:



Матрица ошибок:

	a_{00}	a_{01}	a_{10}	a_{02}	a_{11}	a_{20}	a_{03}	a_{12}	a_{21}	a_{30}
x_1	0	1	0	1	1	0	1	1	1	0
x_2	0	0	1	0	1	1	0	1	1	1
x_3	0	0	0	1	0	0	1	1	0	0
x_4	0	0	0	0	0	1	0	0	1	1
x_5	0	0	0	0	0	0	1	0	0	0
x_6	0	0	0	0	0	0	0	0	0	1

Характеристики расслоения и связности алгоритма $a \in A$

Определение

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

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

X_a называется *порождающим множеством* алгоритма a.

Определение

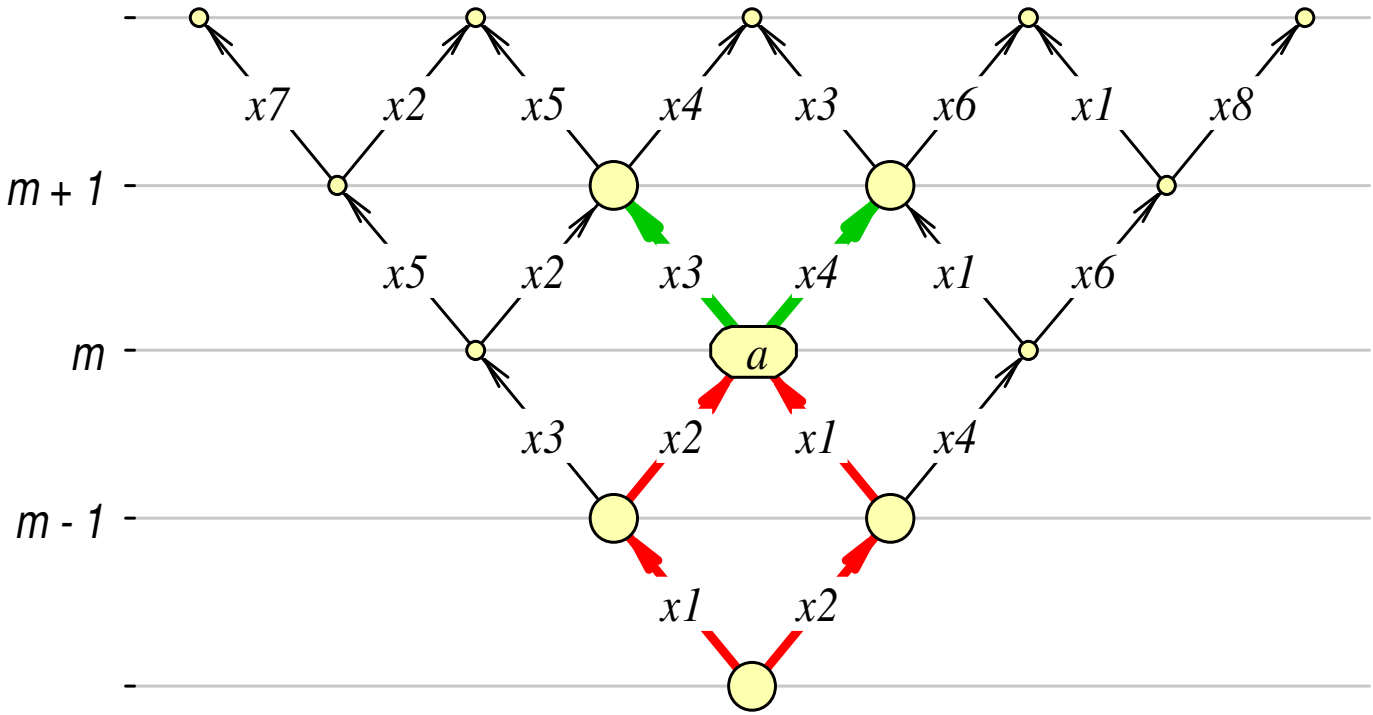
Неполноценность $q(a)$ алгоритма a — это мощность множества объектов, соответствующих всем рёбрам на путях, ведущих в a:

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

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 $X_a \subseteq X^\ell$ и $X'_a \subseteq X^k$.

Основная оценка расслоения–связности

Теорема (Воронцов, Ивахненко, Решетняк, 2010)

Для любого монотонного метода μ, любых X^L, A и $\varepsilon \in (0,1)$

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

где $u = |X_a|$ — верхняя связность алгоритма a,
$q = |X'_a|$ — неполноценность алгоритма a,
$m = n(a, X^L)$ — 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:

Идея доказательства

Опр. Метод обучения μ называется *монотонным*, если

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

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

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

Опр. Метод μ называется *пессимистичным*, если

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

Основная лемма

Если метод обучения μ монотонный и пессимистичный, то

$$\big[\mu X = a\big] \leqslant \big[X_a \subseteq X^\ell\big]\big[X'_a \subseteq X^k\big].$$

Идея доказательства

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

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

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

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

4. По формуле полной вероятности:

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

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

1. Вклад алгоритма $a \in A$ убывает экспоненциально
   по $u(a) \Rightarrow$ **связные семейства меньше переобучаются**;
   по $q(a) \Rightarrow$ **только нижние слои вносят вклад в Q_ε.**
2. Оценка обращается в равенство в случае многомерных монотонных сетей алгоритмов.
3. Вероятность получить алгоритм в результате обучения
$$\mathsf{P}\left[\mu(X^\ell) = a\right] \;\leqslant\; P_a = \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell}.$$
4. Если $q(a) > k$, то $P_a = 0$ и вклад алгоритма a равен 0
   $\Rightarrow$ при малых k оценка вырождается.
5. $\sum_{a\in A} P_a$ — оценка степени завышенности.

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 $|A|=1$ это вероятность большого отклонения частот в двух выборках:

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

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

$$Q_\varepsilon \;\leqslant\; \sum_{a\in A} \mathcal{H}_L^{\ell,\,m}\left(\tfrac{\ell}{L}\left(m-\varepsilon k\right)\right) \;\leqslant\; |A|\cdot\tfrac{3}{2}\exp(-\varepsilon\ell^2).$$

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

$$\widetilde{Q}_\varepsilon(A, X^L) = \mathsf{P}\Big[\sup_{a\in A}\big(\nu(a, X^k) - \nu(a, X^\ell)\big) \geqslant \varepsilon\Big],$$

который учитывает связность, но не учитывает расслоение.

Обобщения метода порождающих и запрещающих множеств

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

$$\big[\mu X^\ell = a\big] \leqslant \big[X_a \subseteq X^\ell\big]\big[X'_a \subseteq X^k\big], \quad \forall X.$$

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

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

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

$$\big[x_i \in X^k\big] I(\mu X, x_i) \leqslant \sum_{v \in V_i} c_{iv}\big[X_{iv} \subseteq X^\ell\big]\big[X'_{iv} \subseteq X^k\big], \quad \forall X.$$

Второе обобщение очень удобно для оценивания CCV.

Оценивание CCV методом ПЗМ по объектам

$$\mathrm{CCV} = \mathsf{E}\nu(\mu X, X^k) = \mathsf{E}\frac{1}{k}\sum_{i=1}^{L}\big[x_i \in X^k\big] I(\mu X, x_i) =$$

$$= \sum_{i=1}^{L}\sum_{v\in V_i}\frac{c_{iv}}{k}\,\mathsf{P}\big[X_{iv} \subseteq X^\ell\big]\big[X'_{iv} \subseteq X^k\big] =$$

$$= \sum_{i=1}^{L}\sum_{v\in V_i}\frac{c_{iv}}{k}\,\frac{C^{\ell-u}_{L-u-q}}{C^\ell_L}.$$

где $u = |X_{iv}|$, $q = |X'_{iv}|$.

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

Развитие комбинаторной теории переобучения

1. Оценки Q_ε и критерии отбора признаков для пороговых логических закономерностей [А.Ивахненко 2010]
2. Точные оценки Q_ε и CCV для многомерных сетей алгоритмов, методы обучения деревьев решений [П.Ботов 2011]
3. Точные оценки CCV для метода k ближайших соседей, отбор эталонных объектов [М.Иванов 2012, А.Зухба 2010]
4. Верхние оценки CCV для монотонных классификаторов, монотонные композиции классификаторов [И.Гуз 2012, Г.Махина 2012]
5. Монотонной корректор для поискового ранжирования [Н.Спирин 2011]
6. Монотонный классификатор для категоризации текстов [М.Иванов 2012]
7. Критерий точности оценок расслоения–связности [Н.Животовский 2011]
8. Теоретико-групповой подход [А.Фрей, И.Толстихин 2013]
9. Оценки расслоения по симметричным кластерам [А.Фрей 2013]
10. Оценки случайного блуждания по нижним слоям графа [Е.Соколов 2012]
11. Точные оценки для пороговых решающих правил [Ш.Ишкина 2015]

Резюме

- Свойства расслоения и связности уменьшают переобучение.
- На практике семейства, как правило, ими обладают.
  Иначе вероятность переобучения была бы близка к 1
  уже при $|A|$ порядка нескольких десятков.
- Практическое применение комбинаторных оценок:
  1) оценить $\eta = Q_\varepsilon(\mu, X^L)$ по нескольким нижним слоям;
  2) выразить $\varepsilon(\eta)$ из $\eta(\varepsilon)$ с помощью обращения;
  3) использовать оценку $\nu(X^L) + \varepsilon(\eta)$ как внешний критерий
  для выбора модели, метода или отбора признаков.
- Информацию о нижних слоях можно получать
  в ходе перебора порогов [Ивахненко 2010, Ботов 2011],
  случайным блужданием по нижним слоям [Соколов 2013],
  в ходе итераций метода обучения.