Комбинаторная теория переобучения

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Теория надёжности обучения по прецедентам
(курс лекций, К. В. Воронцов)»

25 марта 2015 • 1 апреля 2015

Содержание

Бинарная функция потерь. Матрица ошибок

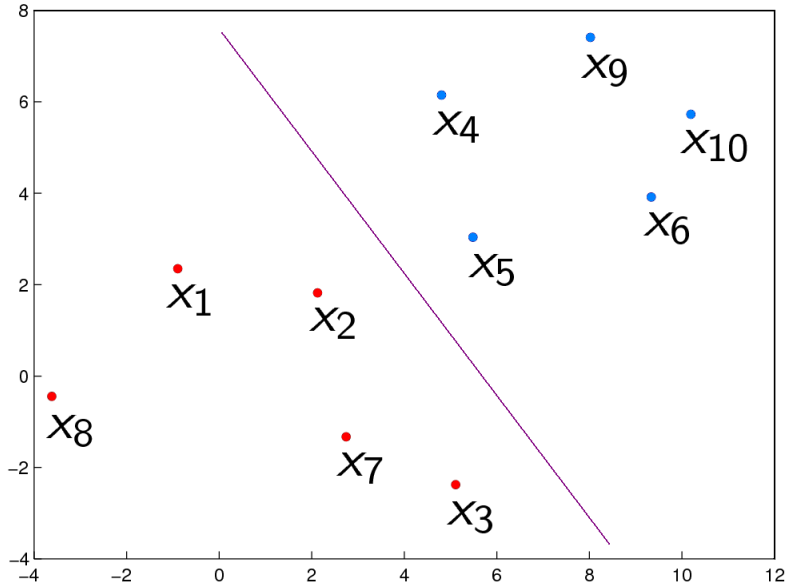
$X^L = \{x_1, \ldots, x_L\}$ — конечное *генеральное множество* объектов;
$A = \{a_1, \ldots, a_D\}$ — конечное семейство *алгоритмов*;
$I(a, x) = [\text{алгоритм } a \text{ ошибается на объекте } x]$;

$L \times D$-*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^ℓ — 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	X^k — скрытая
$\ldots$	0	0	0	**1**	0	0	$\cdots$	1	(контрольная) выборка
x_L	0	1	1	**1**	1	1	$\cdots$	0	длины $k = L - \ell$

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

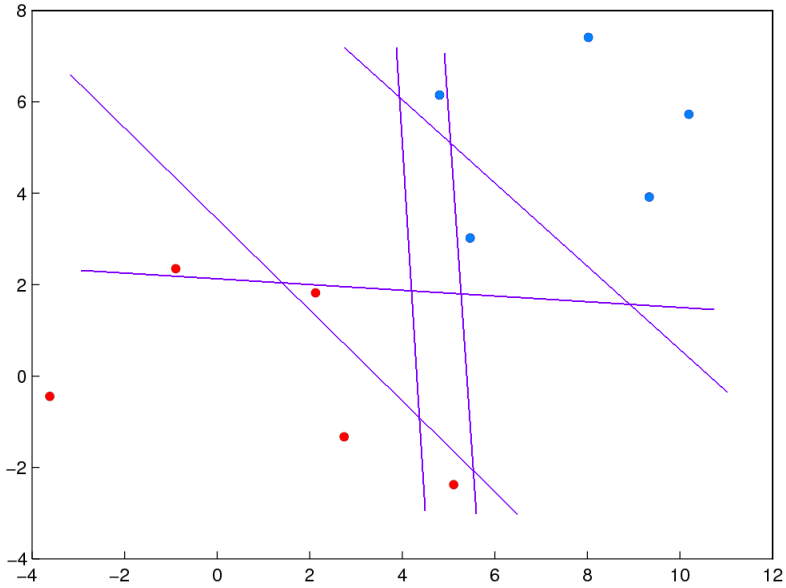
Пример. Матрица ошибок линейных классификаторов



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

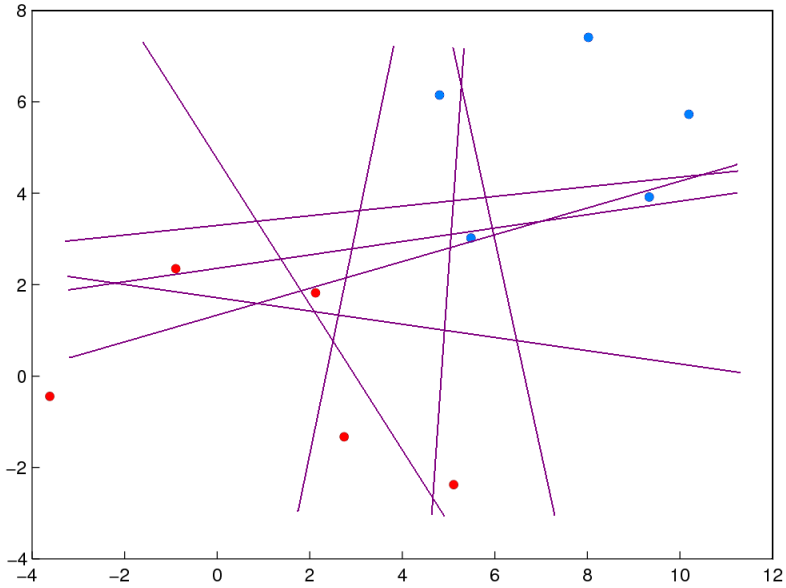


1 вектор с 0 ошибками

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	...

Задача обучения по прецедентам

$\mu\colon X \mapsto a$ — *метод обучения*, произвольной выборке $X \subset X^L$ ставит в соответствие некоторый алгоритм $a \in A$.

$\mu(X) = \arg\min\limits_{a \in A} \nu(a, X)$ — *минимизация эмпирического риска*

$\delta(\mu, X^\ell) = \nu\big(\mu(X^\ell), X^k\big) - \nu\big(\mu(X^\ell), X^\ell\big)$ — *переобученность*

$\delta(\mu, X^\ell) \geqslant \varepsilon$ — событие *переобучения*

Основное вероятностное предположение:
все разбиения $X^\ell \sqcup X^k = X^L$ равновероятны
(слабый вариант гипотезы независимости выборки X^L).

Основная задача — оценить *вероятность переобучения*:

$$Q_\varepsilon(\mu, X^L) = \mathsf{P}\big[\delta(\mu, X^\ell) \geqslant \varepsilon\big] = \frac{1}{C_L^\ell} \sum_{X^\ell \subset X^L} \big[\delta(\mu, X^\ell) \geqslant \varepsilon\big].$$

Обобщающая способность метода обучения

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

$\delta(\mu, X^\ell) = \nu\big(\mu(X^\ell), X^k\big) - \nu\big(\mu(X^\ell), X^\ell\big)$ — *переобученность*.

Функционалы обобщающей способности:

- Полный скользящий контроль (Complete Cross-Validation):
$$\mathrm{CCV}(\mu, X^L) = \mathsf{E}\,\nu\big(\mu(X^\ell), X^k\big).$$
- Ожидаемая переобученность (Expected OverFitting):
$$\mathrm{EOF}(\mu, X^L) = \mathsf{E}\,\delta(\mu, X^\ell).$$
- Вероятность большой частоты ошибок на контроле:
$$R_\varepsilon(\mu, X^L) = \mathsf{P}\big[\nu(\mu(X^\ell), X^k) \geqslant \varepsilon\big].$$
- Вероятность переобучения:
$$Q_\varepsilon(\mu, X^L) = \mathsf{P}\big[\delta(\mu, X^\ell) \geqslant \varepsilon\big].$$

Отличия от постановки задачи, принятой в SLT

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

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

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.

Связь с кросс-валидацией

Полный скользящий контроль и вероятность переобучения:

$$\mathrm{CCV}(\mu, X^L) = \frac{1}{C_L^\ell} \sum_{X^\ell \subset X^L} \nu\big(\mu(X^\ell), X^k\big);$$

$$Q_\varepsilon(\mu, X^L) = \frac{1}{C_L^\ell} \sum_{X^\ell \subset X^L} \big[\nu(\mu(X^\ell), X^k) - \nu(\mu(X^\ell), X^\ell) \geqslant \varepsilon\big].$$

Оценки вероятности методом Монте–Карло —
как доли разбиений выборки из случайного подмножества
N разбиений, $|N|$ порядка 10^3–10^4:

$$\widehat{\mathrm{CCV}}(\mu, X^L) = \frac{1}{|N|} \sum_{(X^k, X^\ell) \in N} \nu\big(\mu(X^\ell), X^k\big);$$

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

Связь с рademахеровской сложностью

Рademахеровская сложность множества A на X^L

$$\mathrm{RC}(A, X^L) = \mathsf{E}_\sigma \sup_{a \in A} \frac{2}{L} \sum_{i=1}^{L} \sigma_i I(a, x_i), \qquad \sigma_i = \begin{cases} +1, & \text{вер. } \frac{1}{2} \\ -1, & \text{вер. } \frac{1}{2} \end{cases}$$

$\sigma_1, \ldots, \sigma_L$ — независимые рademахеровские случ. величины

Ожидаемая переобученность в случае $\ell = k$:

$$\mathrm{EOF}(\mu, X^L) = \mathsf{E} \sup_{a \in A} \frac{2}{L} \sum_{i=1}^{L} \sigma_i I(a, x_i), \qquad \sigma_i = \begin{cases} +1, & x_i \in X^k \\ -1, & x_i \in X^\ell \end{cases}$$

если μ — метод *максимизации переобученности*:

$$\mu(X) = \arg\max_{a \in A} \Big(\nu\big(a, X^k\big) - \nu\big(a, X^\ell\big) \Big)$$

Переход от комбинаторных оценок к SLT

Обычные вероятностные предположения:
X^L — i.i.d. выборка
из вероятностного пространства $\langle \mathscr{X}, \sigma, \mathsf{P} \rangle$ над бесконечным $\mathscr{X}$

Переход от комбинаторной оценки к вероятностной:

1. Пусть имеется комбинаторная оценка:

$$\mathsf{P}_{X \sim X^L} \left[\delta(\mu, X) \geqslant \varepsilon \right] = Q_\varepsilon(\mu, X^L) \leqslant \eta(\varepsilon, X^L)$$

2. Возьмём матожидание от обеих частей неравенства по X^L:

$$\mathsf{P}_{\substack{X \sim \mathscr{X}^\ell \\ X^k \sim \mathscr{X}^k}} \left[\delta(\mu, X) \geqslant \varepsilon \right] = \mathsf{E}_{X^L}\, Q_\varepsilon(\mu, X^L) \leqslant \mathsf{E}_{X^L}\, \eta(\varepsilon, X^L).$$

Если оценка $\eta(\varepsilon, X^L)$ не зависит от выборки X^L,
то она переносится из КТП в SLT непосредственно.

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*.»

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 $X^\ell \sqcup X^k = X^L$ равновероятны.
2. Получение комбинаторных формул для эффективного вычисления функционалов полного скользящего контроля Q_ε, CCV и др., основанных на усреднении *по всем* разбиениям заданной конечной выборки X^L на обучающую X^ℓ и контрольную X^k подвыборки, *без явного применения метода обучения* μ.

Простейший, но важный частный случай

Пусть $A = \{a\}$ — одноэлементное множество, $m = n(a, X^L)$.

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

$$Q_\varepsilon(a, X^L) = \mathsf{P}\left[\nu\left(a, X^k\right) - \nu\left(a, X^\ell\right) \geqslant \varepsilon\right].$$

Теорема

Для любого X^L, любого $\varepsilon \in [0, 1]$

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

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.

Доказательство

1. Обозначим $s = n(a, X^\ell)$.

2. «Школьная» задача по теории вероятностей:
в урне L шаров, m из них чёрные; извлекаем ℓ шаров наугад.
Какова вероятность того, что s из них чёрные?

$$\mathsf{P}\big[n(a, X^\ell) = s\big] = C_m^s C_{L-m}^{\ell-s} / C_L^\ell.$$

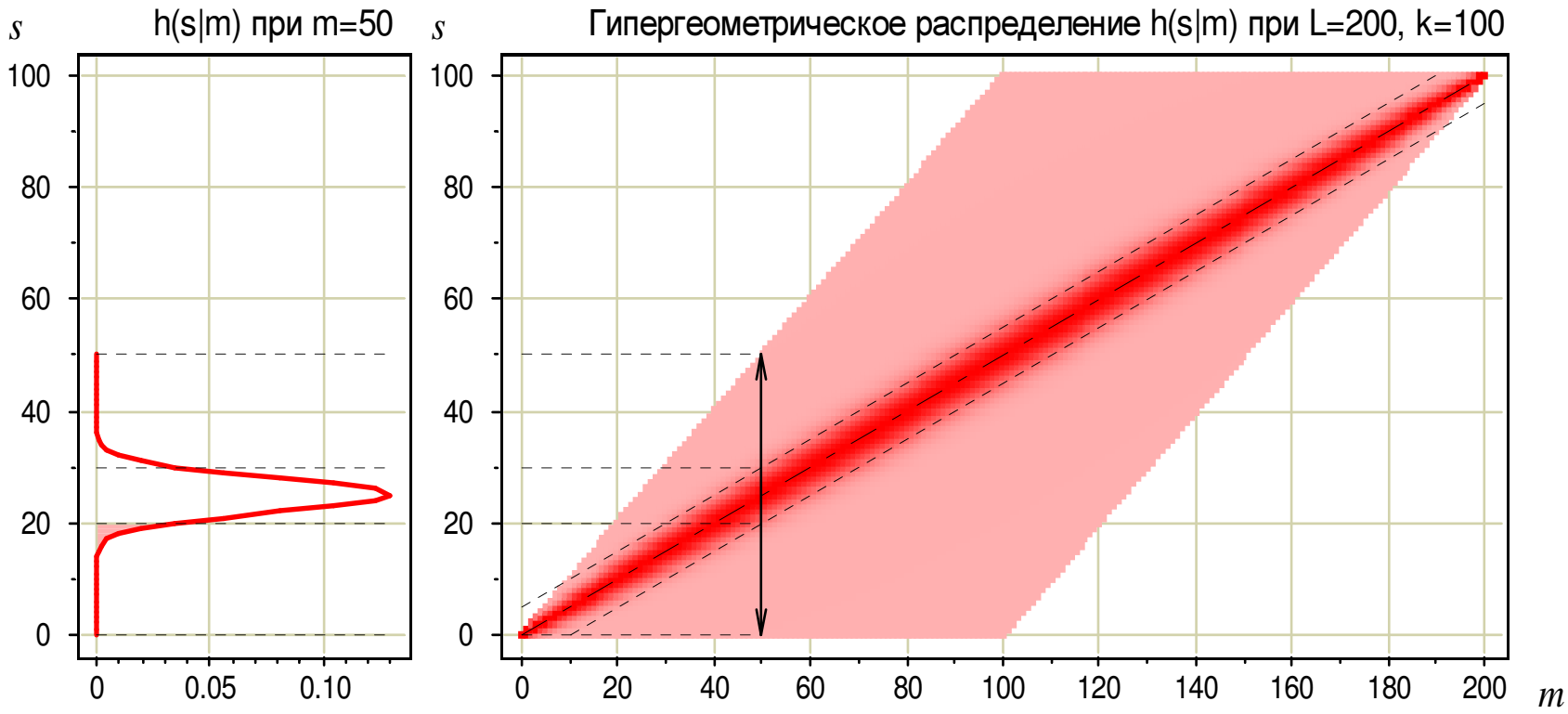
3. Распишем Q_ε, подставив $\nu(a, X^k) = \frac{m-s}{k}$, $\nu(a, X^\ell) = \frac{s}{\ell}$:

$$Q_\varepsilon(a, X^L) = \mathsf{P}\big[\nu\big(a, X^k\big) - \nu\big(a, X^\ell\big) \geqslant \varepsilon\big] =$$

$$= \sum_{s=0}^{\ell} \Big[\underbrace{\tfrac{m-s}{k} - \tfrac{s}{\ell} \geqslant \varepsilon}_{s \leqslant \frac{\ell}{L}(m-\varepsilon k)}\Big] \underbrace{\mathsf{P}\big[n(a, X^\ell) = s\big]}_{C_m^s C_{L-m}^{\ell-s} / C_L^\ell} =$$

$$= \mathcal{H}_L^{\ell,\, m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right). \qquad \blacksquare$$

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



Предсказание числа $m = n(a, X^L)$ по числу $s = n(a, X^\ell)$ возможно благодаря узости гипергеометрического пика, причём при $\ell, k \to \infty$ он сужается, и $\nu(a, X^\ell) \to \nu(a, X^k)$ (явление *концентрации вероятности*, закон больших чисел).

Теория Вапника–Червоненкиса

Рассмотрим общий случай — A произвольное, конечное.

1. Вероятность переобучения оценим сверху вероятностью большого *равномерного отклонения* частот: для любых X^L, μ

$$Q_\varepsilon(\mu, X^L) = \mathsf{P}\big[\delta(\mu, X^\ell) \geqslant \varepsilon\big] \leqslant$$
$$\leqslant \mathsf{P}\Big[\max_{a\in A} \delta(a, X^\ell) \geqslant \varepsilon\Big] = \widetilde{Q}_\varepsilon(A, X^L).$$

2. Оценим вероятность объединения событий суммой их вероятностей (неравенство Буля, union bound):

$$\widetilde{Q}_\varepsilon(A, X^L) = \mathsf{P} \max_{a\in A} \big[\delta(a, X^\ell) \geqslant \varepsilon\big] \leqslant$$
$$\leqslant \mathsf{P} \sum_{a\in A} \big[\delta(a, X^\ell) \geqslant \varepsilon\big] = \sum_{a\in A} \underbrace{\mathsf{P}\big[\delta(a, X^\ell) \geqslant \varepsilon\big]}_{Q_\varepsilon(a, X^L)}.$$

Теория Вапника–Червоненкиса

Таким образом, доказали важную теорему:

Теорема

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

$$\widetilde{Q}_\varepsilon(A, X^L) \leqslant \sum_{a \in A} \mathcal{H}_L^{\ell,\, m}\left(\tfrac{\ell}{L}\big(m - \varepsilon k\big)\right),$$

где $m = n(a, X^L)$.

Следствие (Вапник и Червоненкис, 1968)

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

$$\widetilde{Q}_\varepsilon(A, X^L) \leqslant |A| \cdot \max_m \mathcal{H}_L^{\ell,\, m}\left(\tfrac{\ell}{L}\big(m - \varepsilon k\big)\right) \leqslant$$
$$\leqslant |A| \cdot \tfrac{3}{2} \exp\big(-\varepsilon^2 \ell\big), \quad \text{при } \ell = k.$$

Обобщение на случай бесконечных семейств A

Функция роста $\Delta^A(L)$ семейства A — это максимальное по X^L число различных векторов ошибок $\mathbf{a} = \big(I(a, x_1), \ldots, I(a, x_L)\big)$.
В оценке надо заменить $|A|$ на функцию роста $\Delta^A(L)$.

Ёмкость (размерность Вапника-Червоненкиса) семейства A — это максимальная длина выборки h, для которой $\Delta^A(h) = 2^h$.

Теорема

Если такое h существует, то $\Delta^A(L) \leqslant C_L^0 + \cdots + C_L^h \leqslant \frac{3}{2}\frac{L^h}{h!}$.

Теорема

Ёмкость семейства линейных классификаторов на два класса

$$a(x) = \operatorname{sign}(w_1 x^1 + \cdots + w_n x^n), \quad x = (x^1, \ldots, x^n) \in X.$$

равна размерности пространства параметров, $\operatorname{VCdim}(A) = n$.

Обращение оценки Вапника-Червоненкиса (при $\ell = k$)

1. Оценка: $\mathsf{P}\Big[\max\limits_{a\in A}\big(\nu(a,X^k)-\nu(a,X^\ell)\big)\geqslant\varepsilon\Big]\leqslant\Delta\,\frac{3}{2}\exp\big(-\ell\varepsilon^2\big)$.

Тогда для любого $a\in A$ с вероятностью не менее $(1-\eta)$

$$\nu(a,X^k)\leqslant \underbrace{\nu(a,X^\ell)}_{\substack{\text{эмпирический}\\ \text{риск}}} + \underbrace{\sqrt{\frac{1}{\ell}\ln\Delta+\frac{1}{\ell}\ln\frac{3}{2\eta}}}_{\text{штраф за сложность}}.$$

2. Оценка: $\mathsf{P}\Big[\max\limits_{a\in A}\big(\nu(a,X^k)-\nu(a,X^\ell)\big)\geqslant\varepsilon\Big]\leqslant\frac{3}{2}\frac{L^h}{h!}\cdot\frac{3}{2}\exp\big(-\ell\varepsilon^2\big)$.

Тогда для любого $a\in A$ с вероятностью не менее $(1-\eta)$

$$\nu(a,X^k)\leqslant \underbrace{\nu(a,X^\ell)}_{\substack{\text{эмпирический}\\ \text{риск}}} + \underbrace{\sqrt{\frac{h}{\ell}\ln\frac{2e\ell}{h}+\frac{1}{\ell}\ln\frac{9}{4\eta}}}_{\text{штраф за сложность}}.$$

Метод структурной минимизации риска (СМР)

Дано: система вложенных подсемейств возрастающей ёмкости

$$A_0 \subset A_1 \subset \cdots \subset A_h \subset \cdots$$

Найти: оптимальную ёмкость h^*, такую, что

$$\nu(a, X^k) \leqslant \underbrace{\min_{a \in A_h} \nu(a, X^\ell)}_{\text{минимизация эмпирического риска}} + \underbrace{\sqrt{\frac{h}{\ell} \ln \frac{2e\ell}{h} + \frac{1}{\ell} \ln \frac{9}{4\eta}}}_{\text{штраф за сложность}} \to \min_h$$

Недостатки СМР:

- h^* может оказаться заниженной из-за завышенности Q_ε.
- На практике эмпирический CV предпочтительнее этих оценок.

Причины завышенности оценок Вапника-Червоненкиса

- **Оценка равномерного отклонения** сильно завышена, когда бóльшая часть алгоритмов имеет исчезающе малую вероятность быть результатом обучения.

  На практике распределение

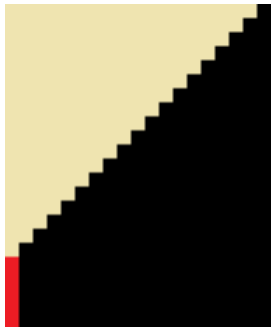
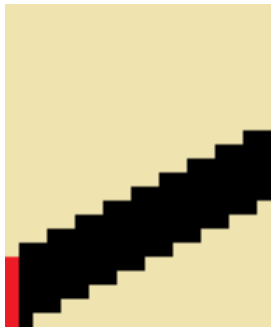
  $$p(a) = \mathsf{P}\big[\mu(X^\ell) = a\big], \quad a \in A$$

  как правило, существенно неравномерно!

  Будем называть это эффектом расслоения семейства A.
- **Неравенство Буля** сильно завышено, когда среди бинарных векторов ошибок есть много похожих.

  Будем называть это эффектом сходства алгоритмов.

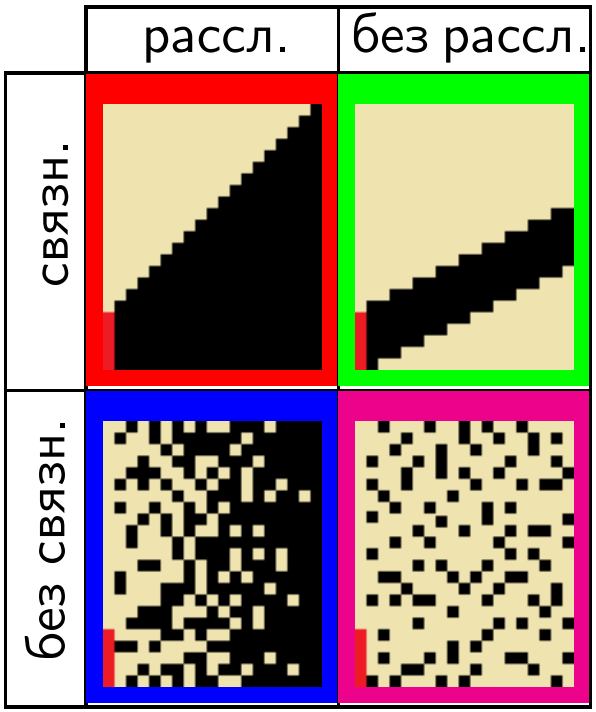
Эксперимент: четыре семейства алгоритмов, заданных матрицами ошибок; лучший алгоритм у всех одинаков

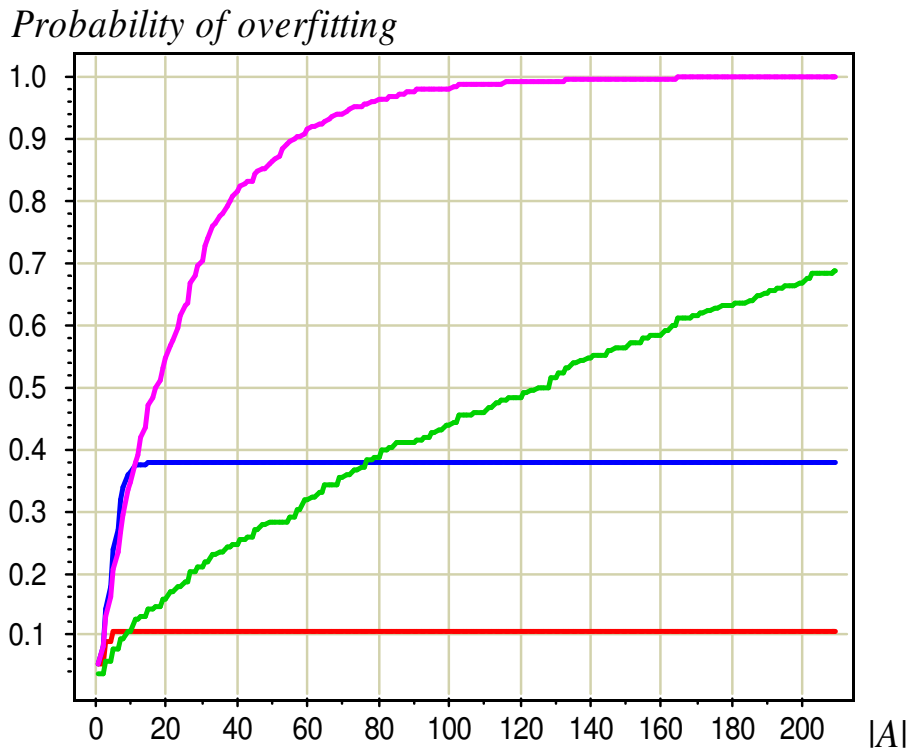
	с расслоением по числу ошибок	без расслоения по числу ошибок
каждая пара соседних алгоритмов отличается только на одном объекте (образуется *цепь*)		
соседние алгоритмы существенно различны, (*цепь* не образуется)		

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

Эксперимент с «трижды испорченной» монотонной цепью

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

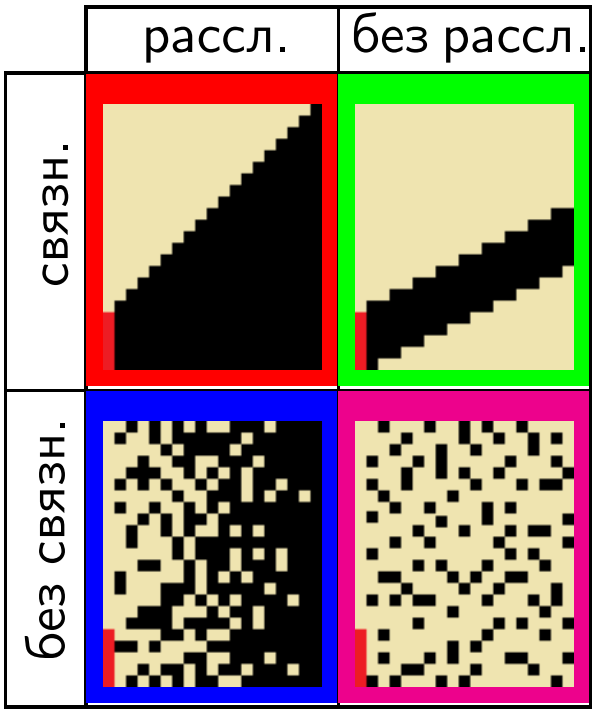


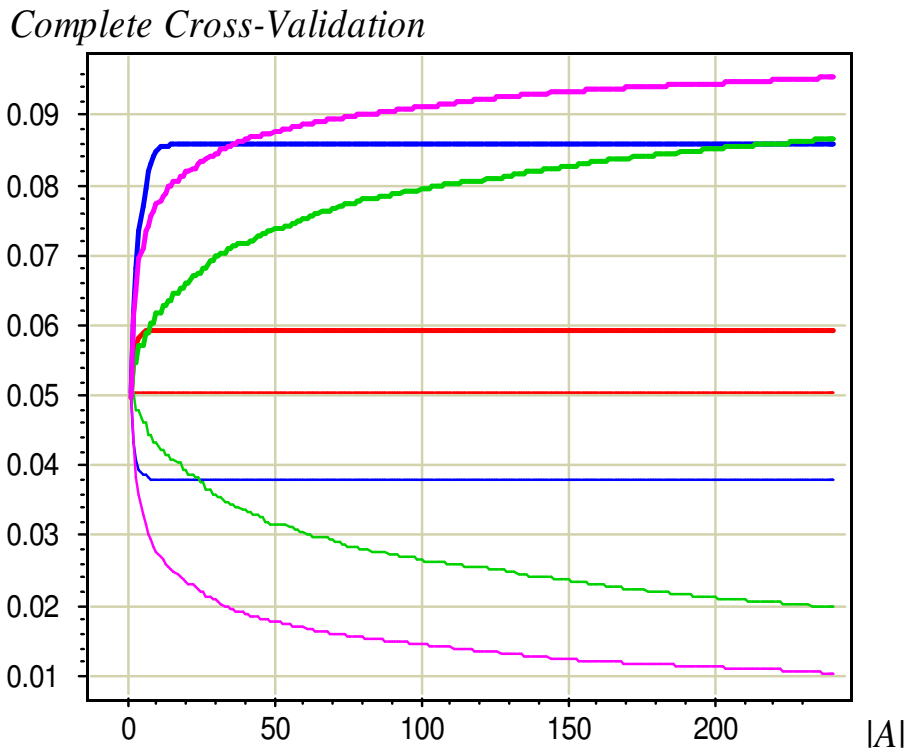


- Огромные семейства с P&C могут почти не переобучаться
- Без P&C даже 30 алгоритмов могут сильно переобучаться

Эксперимент с «трижды испорченной» монотонной цепью

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





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

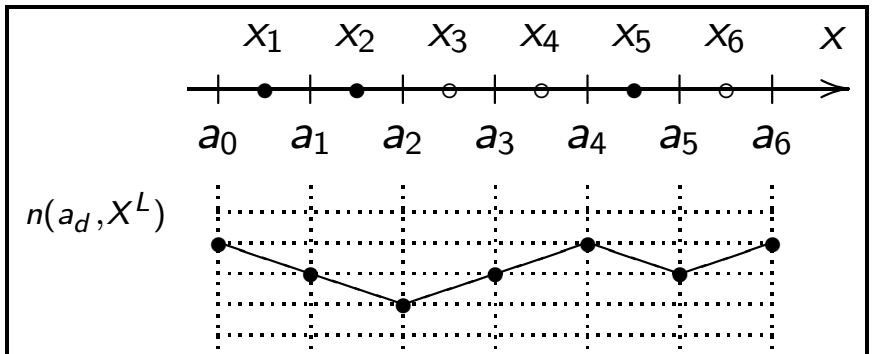
$\mathsf{E}\,\nu\big(\mu(X^\ell), X^\ell\big)$ — тонкие линии

Цепь, порождаемая семейством пороговых классификаторов

Пусть $f(x)$ — вещественный признак,

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

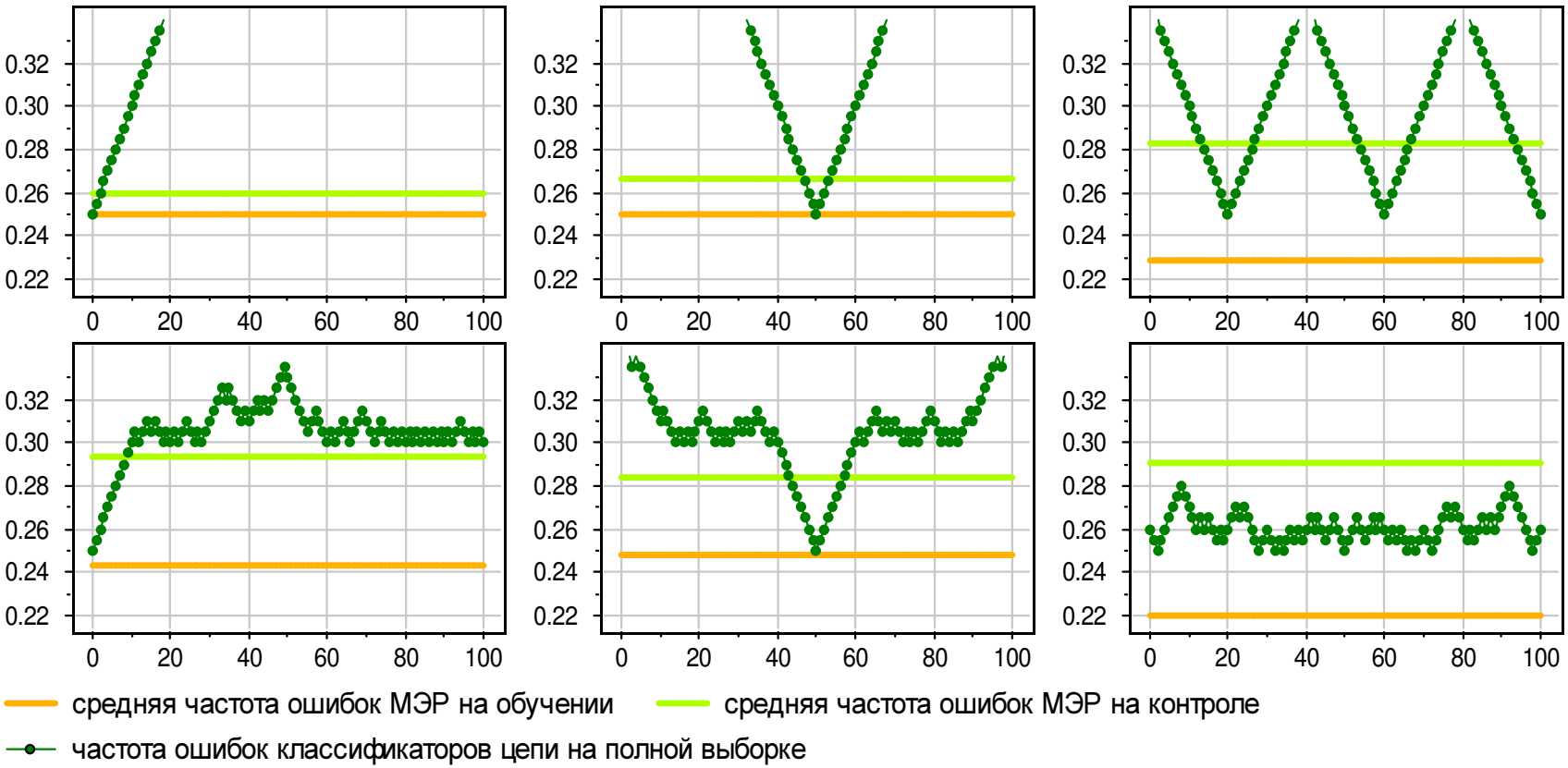
Пример:



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* $\langle A, E \rangle$:

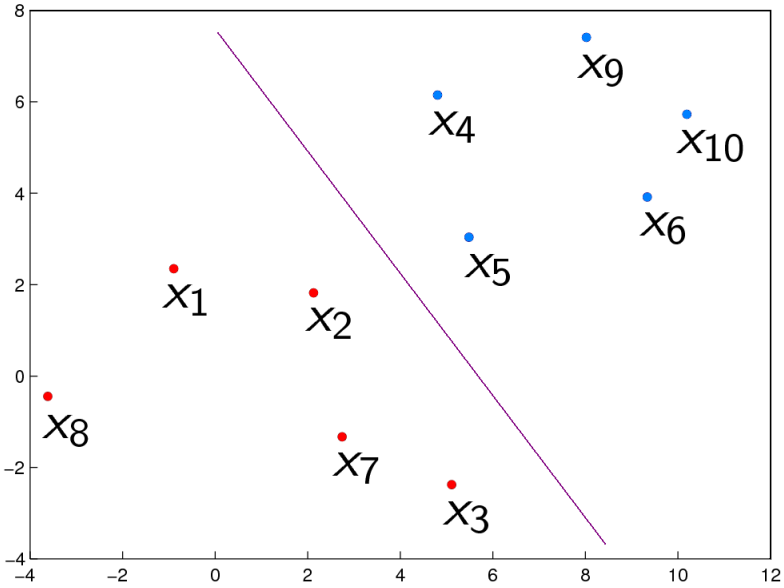
A — множество попарно различных векторов ошибок;

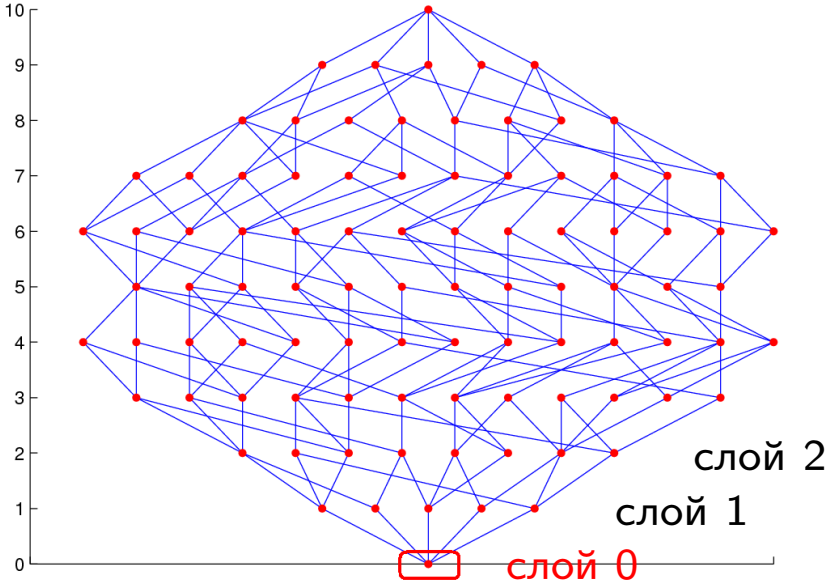
$E = \{(a,b)\colon a \prec b\}$.

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 $A_m = \{a \in A\colon n(a,X^L) = m\}, \quad m = 0,\ldots,L$;

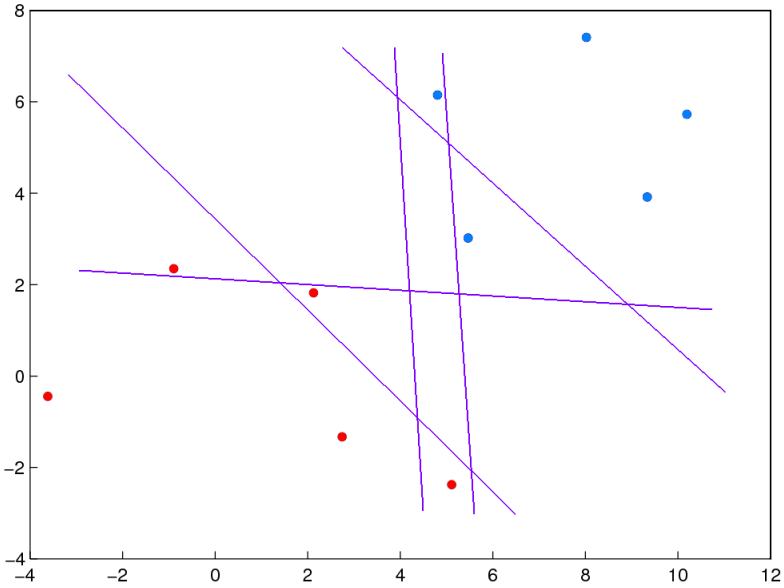
Пример 1. Семейство линейных алгоритмов классификации

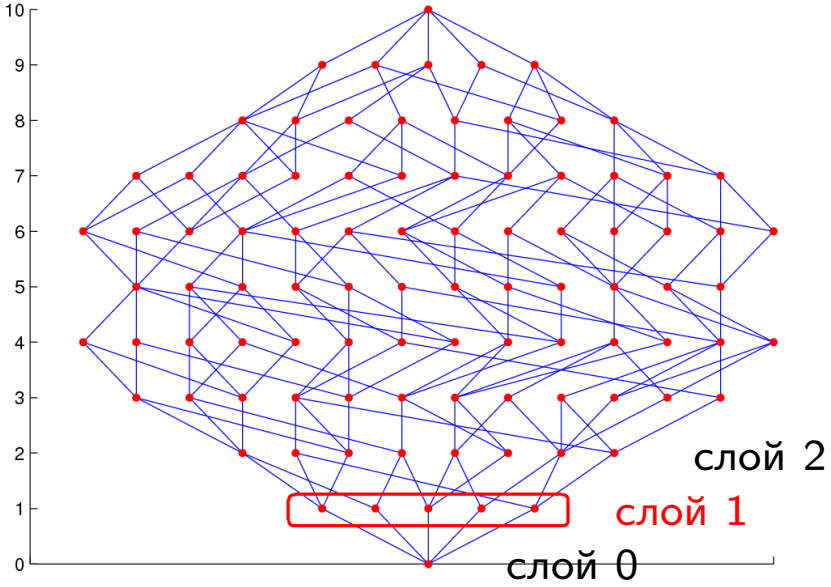




	слой 0
x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0

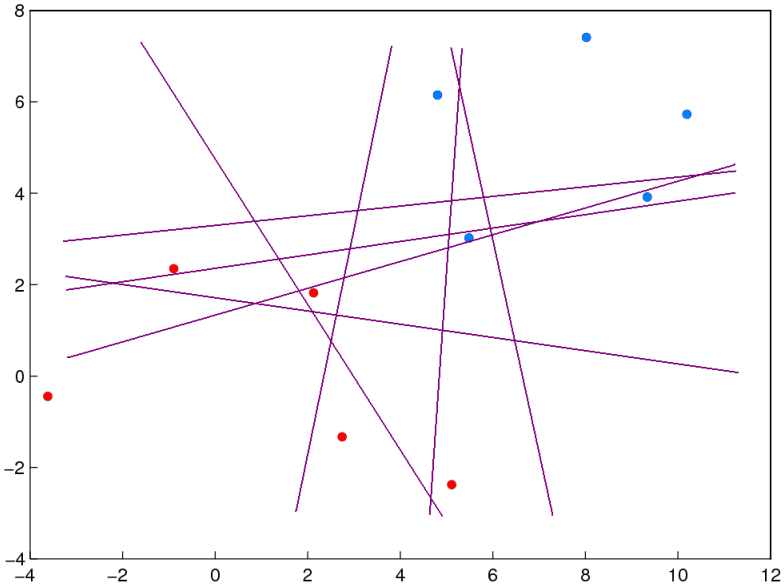
Пример 1. Семейство линейных алгоритмов классификации

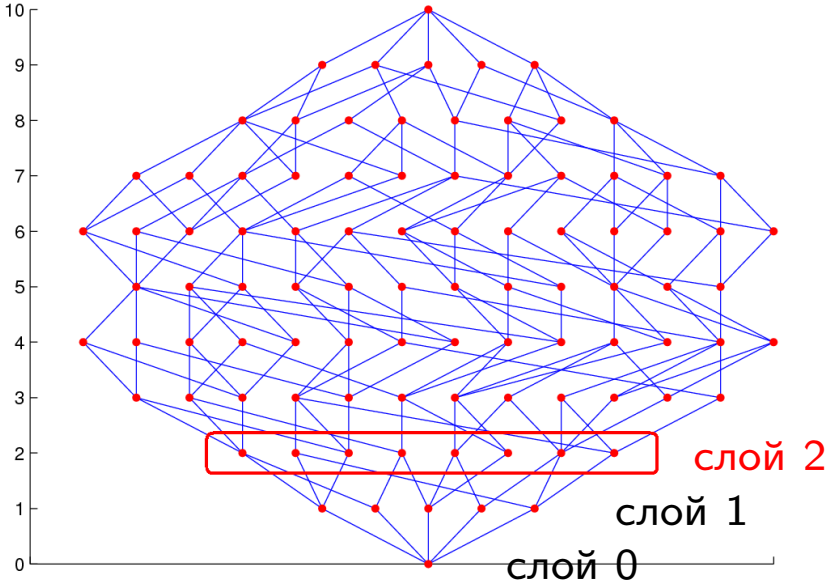




	слой 0	слой 1				
x_1	0	1	0	0	0	0
x_2	0	0	1	0	0	0
x_3	0	0	0	1	0	0
x_4	0	0	0	0	1	0
x_5	0	0	0	0	0	1
x_6	0	0	0	0	0	0
x_7	0	0	0	0	0	0
x_8	0	0	0	0	0	0
x_9	0	0	0	0	0	0
x_{10}	0	0	0	0	0	0

Пример 1. Семейство линейных алгоритмов классификации





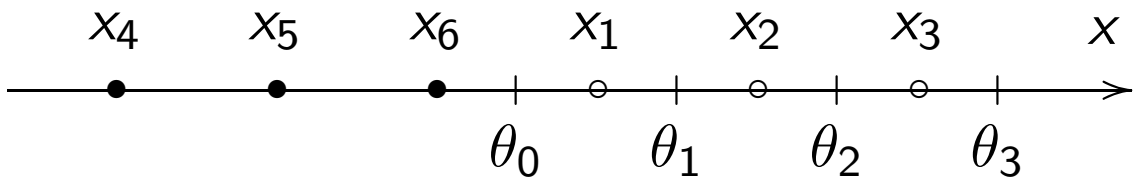
	слой 0	слой 1					слой 2								
x_1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	...
x_2	0	0	1	0	0	0	1	1	0	0	0	0	0	0	...
x_3	0	0	0	1	0	0	0	1	1	0	0	0	0	1	...
x_4	0	0	0	0	1	0	0	0	1	1	0	0	0	0	...
x_5	0	0	0	0	0	1	0	0	0	1	1	1	0	0	...
x_6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	...
x_7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	...
x_8	0	0	0	0	0	0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

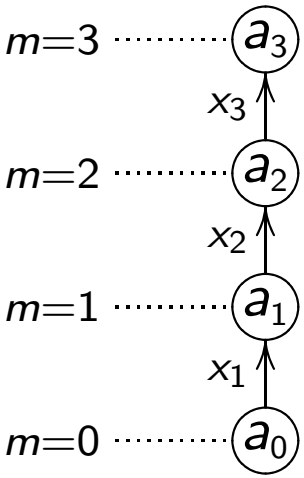
Опр. *Монотонная цепь* алгоритмов: $a_0 \prec a_1 \prec \cdots \prec a_D$.

Пример: 1D пороговый классификатор $a_d(x) = [x - \theta_d]$;

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



Граф семейства:



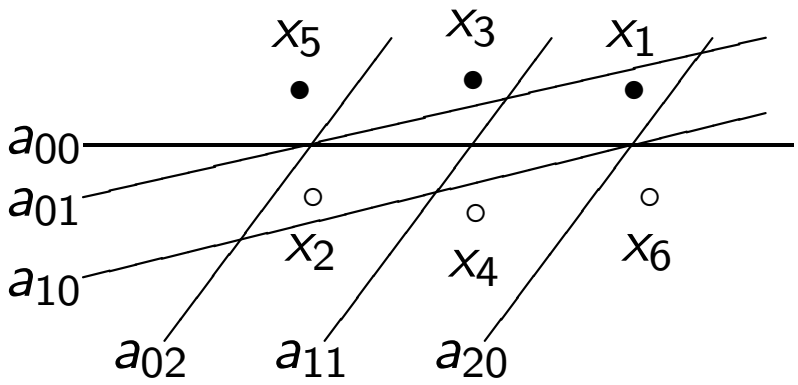
Матрица ошибок:

	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

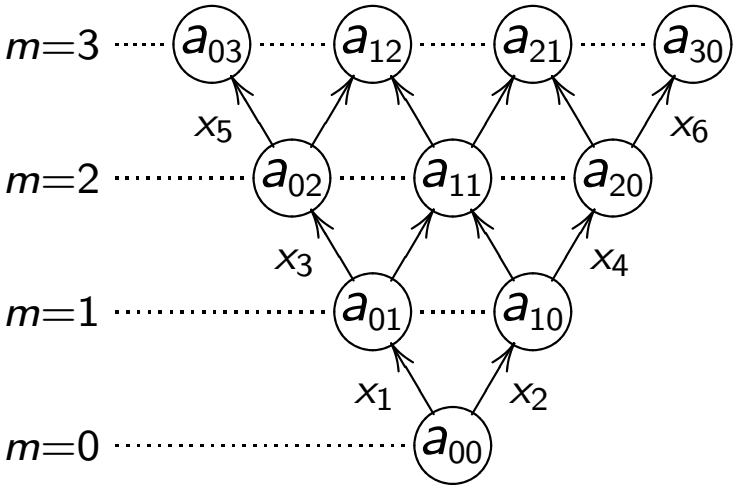
Пример 3. Двумерная сеть классификаторов

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

Порождающее и запрещающее множества алгоритмов

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 число различных 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.

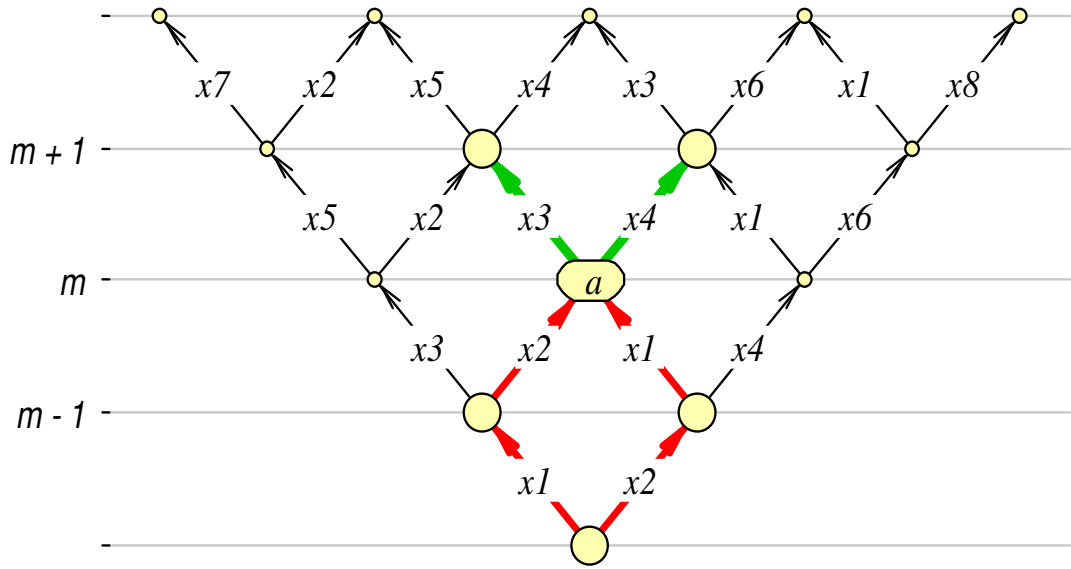
Характеристики расслоения и связности алгоритма

Верхняя связность $u(a) = \#\{x_{ab} \in X^L \mid a \prec b\}$
Нижняя связность $d(a) = \#\{x_{ba} \in X^L \mid b \prec a\}$

Неполноценность $q(a) = \#\{x \in X^L \mid \exists b \in A\colon\ b < a,\ I(b,x) < I(a,x)\}$
Число ошибок $m(a) = n(a, X^L)$.

Утв.
$d(a) \leqslant q(a) \leqslant m(a)$

Пример: двумерная сеть алгоритмов
$u(a) = \#\{x3, x4\} = 2$
$d(a) = \#\{x1, x2\} = 2$
$q(a) = \#\{x1, x2\} = 2$



Теорема о порождающих и запрещающих множествах

Если $\big[\mu(X^\ell) = a\big] = \big[X_a \subseteq X^\ell\big]\big[X'_a \subseteq X^k\big]$, то

$$Q_\varepsilon(\mu, X^L) = \sum_{a\in A} P_a \mathcal{H}_{L_a}^{\ell_a, m_a}(s_a(\varepsilon)), \quad P_a = \mathsf{P}\big[\mu(X^\ell){=}a\big] = \frac{C_{L_a}^{\ell_a}}{C_L^\ell},$$

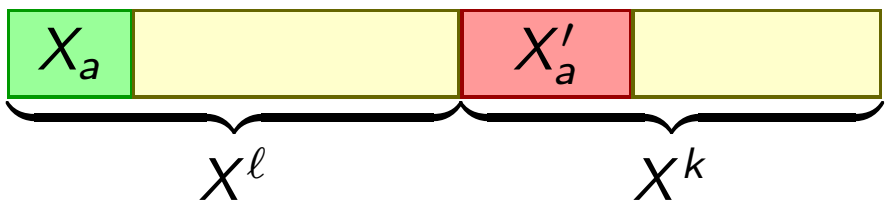
$$L_a = L - |X_a| - |X'_a|,$$

$$\ell_a = \ell - |X_a|,$$

$$m_a = n(a, X^L \backslash X_a \backslash X'_a);$$

$$s_a(\varepsilon) = \tfrac{\ell}{L}\big(n(a, X^L) - \varepsilon k\big) - n(a, X_a).$$

$\mu(X^\ell) = a$:



Монотонные методы обучения

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

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

Лемма

Если метод обучения μ монотонный и пессимистичный, то

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

Верхняя оценка расслоения–связности

Теорема (Воронцов, Ивахненко, Решетняк, 2010)

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

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

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

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.

Верхние оценки средней частоты ошибок на контроле

Теорема

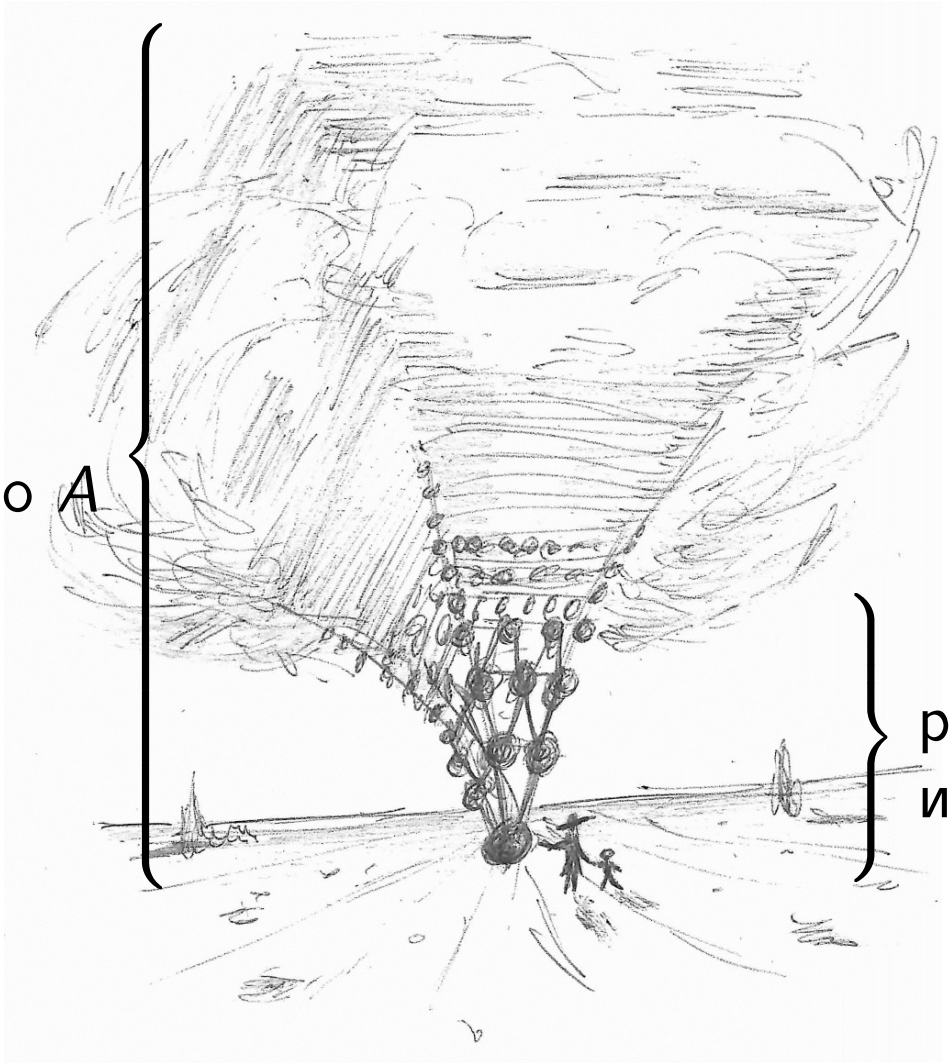
Для любого монотонного метода μ, любых X^L и A

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

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

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

Идея использования оценок расслоения–связности

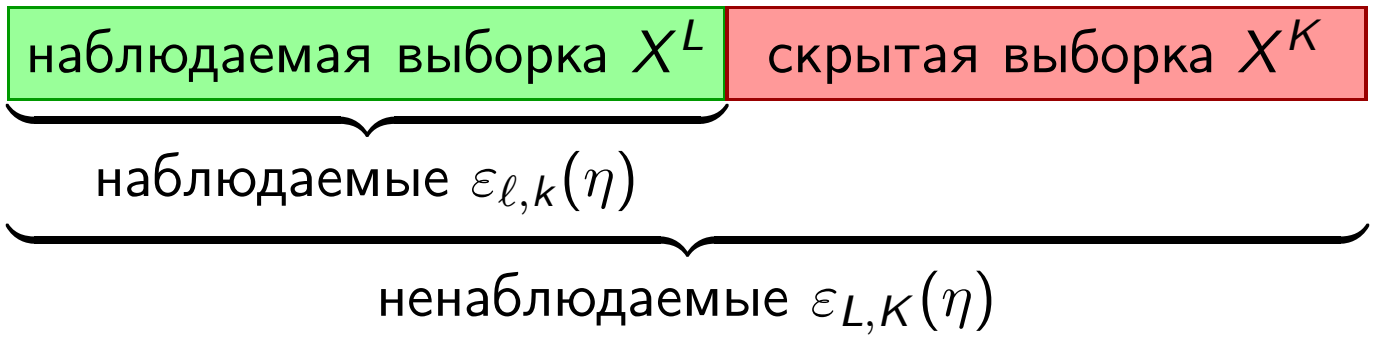


Резюме

- 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) применив обращение, оценить ε через η;
  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 A [Соколов 2013],
  в ходе итераций метода обучения,...

Нерешённые проблемы комбинаторной теории переобучения

- **Проблема вычислительной неэффективности.**
  На больших выборках оценки расслоения–связности избыточно детализируют структуру семейства алгоритмов.
- **Обоснование переноса оценки на скрытую выборку.**
  Оценки вычисляются для разбиения $X^L = X^\ell \sqcup X^k$,
  но применяются к разбиению $X^L \sqcup X^K$ со скрытой X^K.



В экспериментах это не приводит к переобучению,
оценки устойчивы к изменению длины выборки при $L > 60$.
Но теоретического обоснования пока нет.