# Теоретико-групповой подход в комбинаторной теории переобучения

Фрей Александр Ильич
sashafrey@gmail.com

Научнай руководитель: Константин Вячеславович Воронцов

Семинар «Стохастический анализ в задачах» НМУ–МФТИ

26 октября 2013

# Содержание

# Проблема переобучения

$X = \{x_1, \ldots, x_\ell\}$ — конечное множество объектов,
$A$ — семейство алгоритмов классификации,

$a = \arg\min_{a \in A} Err(a, X)$ — минимизация эмпирического риска,
или, в более общем случае,

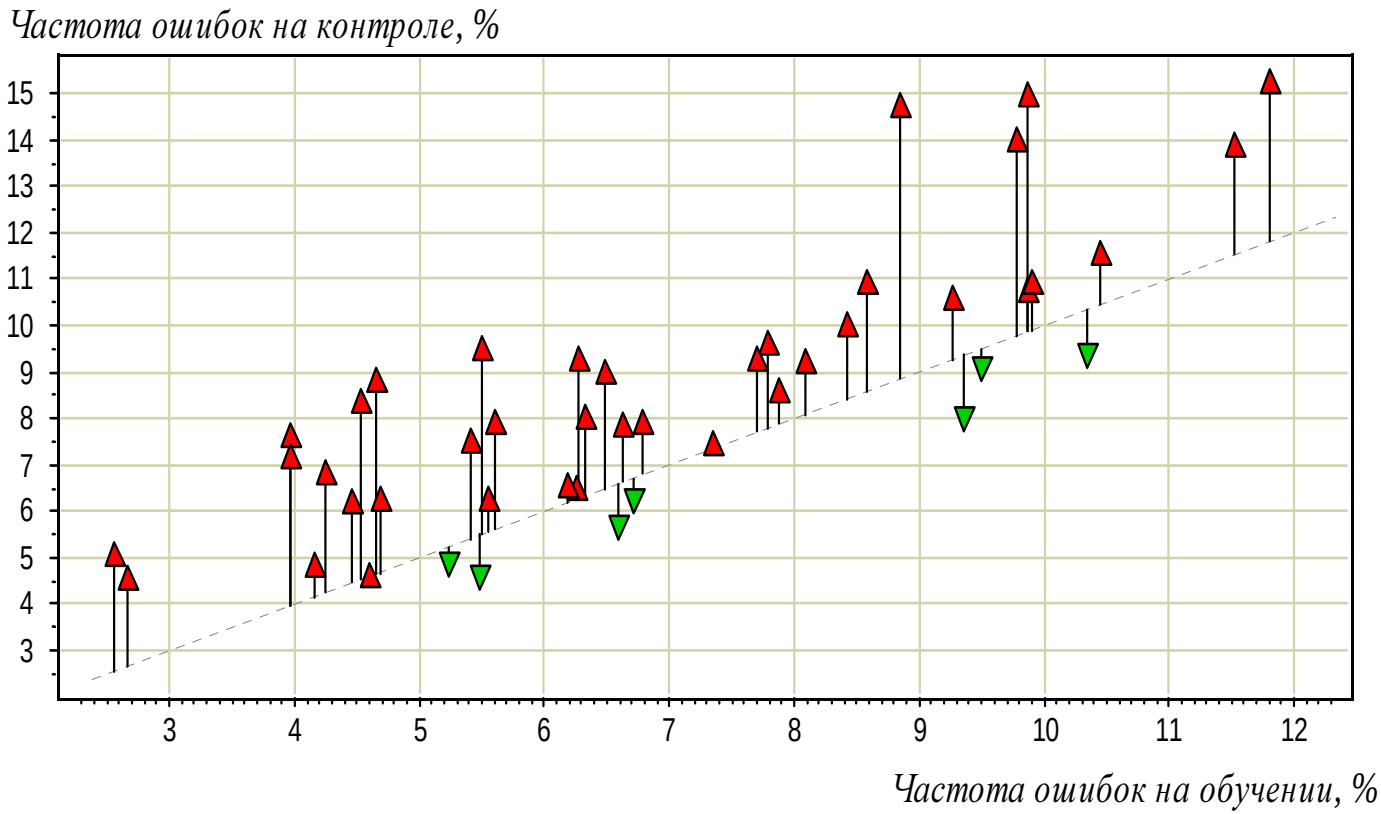
$a = \mu(X)$ — *метод обучения* $\mu$ настраивает алгоритм $a$ по выборке $X$.

**Оценивание обобщающей способности:**

1. Как ограничить ошибку $Err(a, \bar{X})$, где $\bar{X} = \{x'_1, \ldots, x'_k\}$ — независимая контрольная выборка?
2. Как строить методы обучения с высокой обобщающей способностью? (т.е. с низкой ошибкой $Err(a, \bar{X})$).

# Пример переобучения. Реальная задача классификации

Задача предсказания отдалённого результата хирургического лечения астеросклероза. Точки — различные алгоритмы.



# Принцип равномерной сходимости

Классический подход — принцип равномерной сходимости:

$$\mathsf{P}_{\bar{X}}\left( \sup_{a \in A} \left| P(a) - Err(a, X) \right| \geq \varepsilon \right) \leq \text{GenBound}(\ell, k, A, \varepsilon)$$

где $P(a) = \mathsf{E}_{\bar{X}} Err(a, \bar{X})$ [Вапник, Червоненкис, 1971].

**Ключевая проблема:**

- Такие оценки могут быть завышены в $\sim 10^5..10^9$ раз [1].

**Подход к решению проблемы:**

1. изменить постановку задачи (левая часть);
2. комбинаторный подход к выводу оценок (правая часть).

[1] Vorontsov K. V. Combinatorial probability and the tightness of generalization bounds // PRIA. — 2008. — V. 18, no. 2. — P. 243–259.

# Бинарная матрица ошибок

$\mathbb{X} = \{x_1, \ldots, x_L\}$ — конечное генеральное множество объектов

$\mathbb{A} = \{a_1, \ldots, a_D\}$ — конечное множество алгоритмов

$I(a, x) = [\text{алгоритм } a \text{ ошибается на } x]$ — индикатор ошибки

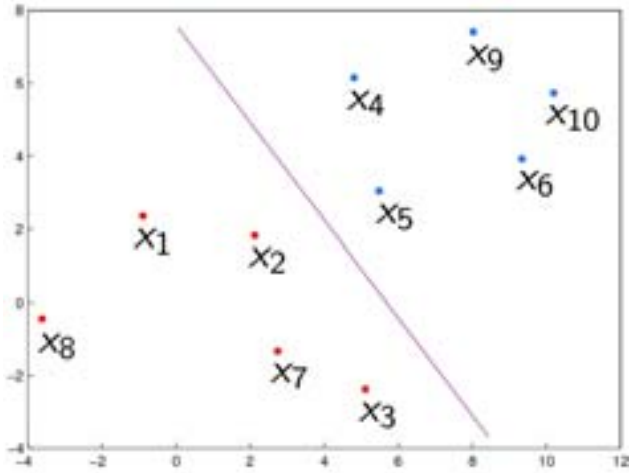
*Матрица ошибок* размера $L \times D$, все столбцы различны:

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $\cdots$ | $a_D$ | |
|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 0 | **0** | 0 | 1 | $\cdots$ | 1 | $X$ — наблюдаемая |
| $\ldots$ | 0 | 0 | 0 | **0** | 1 | 1 | $\cdots$ | 1 | обучающая выборка |
| $x_\ell$ | 0 | 0 | 1 | **0** | 0 | 0 | $\cdots$ | 0 | размера $\ell$ |
| $x_{\ell+1}$ | 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 | размера $k = L - \ell$ |

$a \mapsto \big(I(a, x_1), \ldots, I(a, x_L)\big)$ — *вектор ошибок* алгоритма $a$

$\nu(a, X) = \frac{1}{|X|} \sum_{x \in X} I(a, x)$ — *частота ошибок* на выборке $X \subset \mathbb{X}$

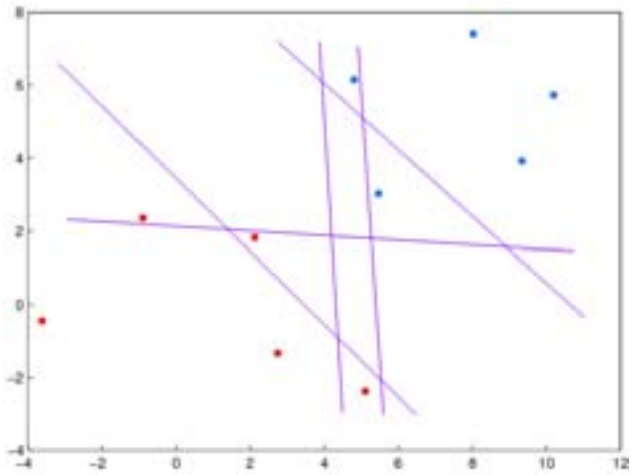
# Пример. Матрица ошибок линейных классификаторов.




1 вектор — 0 ошибок

| | нет ошибок |
|---|---|
| $x_1$ | 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.

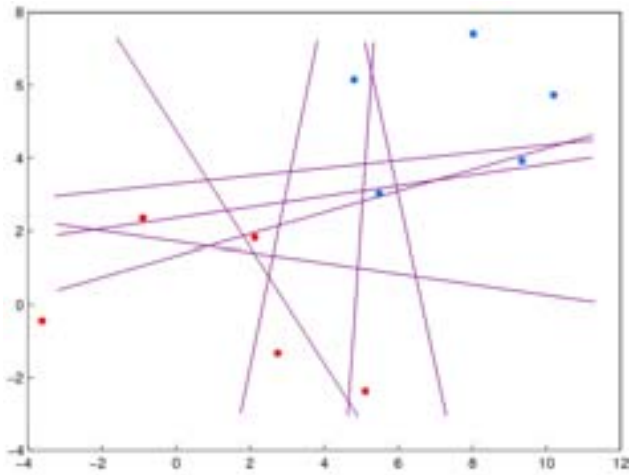


1 вектор — 0 ошибок

5 векторов — 1 ошибка

| | нет ошибок | 1 ошибка | | | | |
|---|---|---|---|---|---|---|
| $x_1$ | 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 | 0 |
| $x_6$ | 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 — 0 ошибок

5 векторов — 1 ошибка

8 векторов — 2 ошибки

| | нет ошибок | 1 ошибка | | | | | 2 ошибки | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 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 | ... |

# Вероятность больших уклонений

$\mu\colon X \mapsto a$ — *метод обучения*
$\nu\big(\mu X, X\big)$ — *частота ошибок на обучении*
$\nu\big(\mu X, \bar{X}\big)$ — *частота ошибок на контроле*
$\delta(\mu, X) \equiv \nu\big(\mu X, \bar{X}\big) - \nu\big(\mu X, X\big)$ — *переобучение* $\mu$ на $X$ и $\bar{X}$

## Аксиома (Более слабый вариант i.i.d. гипотезы)

$\mathbb{X}$ *фиксировано; разбиения* $\mathbb{X} = X \sqcup \bar{X}$ *— равновероятны,*
$X$ *— наблюдаемая обучающая выборка размера* $\ell$*,*
$\bar{X}$ *— скрытая контрольная выборка размера* $k$*,* $L = \ell + k$

## Определение. *Вероятность переобучения*

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

# Среднее переобучение (EOF)

$\mu\colon X \mapsto a$ — *метод обучения*
$\nu\left(\mu X, X\right)$ — *частота ошибок на обучении*
$\nu\left(\mu X, \bar{X}\right)$ — *частота ошибок на контроле*
$\delta(\mu, X) \equiv \nu\left(\mu X, \bar{X}\right) - \nu\left(\mu X, X\right)$ — *переобучение* $\mu$ на $X$ и $\bar{X}$

## Аксиома (Более слабый вариант i.i.d. гипотезы)

$\mathbb{X}$ *фиксировано; разбиения* $\mathbb{X} = X \sqcup \bar{X}$ *— равновероятны,*
$X$ *— наблюдаемая обучающая выборка размера* $\ell$*,*
$\bar{X}$ *— скрытая контрольная выборка размера* $k$*,* $L = \ell + k$

## Определение. *Среднее переобучение (EOF)*

$$EOF(\mu, \mathbb{X}) = \mathsf{E}\delta(\mu, X) = \frac{1}{C_L^\ell} \sum_{X \subset \mathbb{X}} \delta(\mu, X)$$

# Некоторые классические оценки

## Теорема (Оценка для одного алгоритма)

*Для* $A = \{a\}$, *любого* $\mathbb{X}$ *и любого* $\varepsilon \in (0,1)$

$$Q_\varepsilon(a, \mathbb{X}) = H_L^{\ell, n(a)}\left(s_a(\varepsilon)\right), \quad s_a(\varepsilon) = \frac{\ell}{L}\left(n(a) - \varepsilon k\right),$$

$H_L^{\ell,m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ — *гипергеометрическое распределение.*

## Теорема (Вапник и Червоненкис, 1971)

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

$$Q_\varepsilon(\mu, \mathbb{X}) \ll \sum_{a \in A} H_L^{\ell, n(a)}\left(s_a(\varepsilon)\right) \leq |A| \cdot \max_m H_L^{\ell,m}\left(\frac{\ell}{L}\left(m - \varepsilon k\right)\right).$$

# Методы вывода комбинаторных оценок

1. Метод производящих и запрещающих объектов
   - Монотонная цепочка и сетка
2. Блочная оценка
   - Пара алгоритмов
3. Рекуррентное вычисление вероятности переобучения по заданной матрице ошибок
   - Теоретический инструмент для доказательства универсальных оценок
4. Гипотеза $t$-слоев и метод $\beta$-многочленов
   - Точные оценки для унимодальных цепочек
   - Приближенные оценки для унимодальных сеток
5. Метод разбиения множества алгоритмов на орбиты
   - Пучок монотонных цепочек
   - Полный слой, полный куб алгоритмов
   - Шар алгоритмов
   - Точные оценки для монотонных и унимодальных сеток

# Рандомизированный метод обучения

- $\mu\colon \{X\} \to \mathbb{A}$ — детерминированный метод обучения
- $\mu X \in A(X) \equiv \operatorname*{Argmin}_{a\in\mathbb{A}} n(a, X)$ — детерминированный МЭР:
  - Пессимистический МЭР: $\mu X \in \operatorname*{Argmax}_{a\in A(X)} n(a, \mathbb{X})$;
  - Оптимистический МЭР: $\mu X \in \operatorname*{Argmin}_{a\in A} n(a, \mathbb{X})$.
- Рандомизированный метод обучения

$$\mu\colon 2^{\mathbb{A}} \times [\mathbb{X}]^{\ell} \to \{f\colon \mathbb{A} \to \{0,1\}\},$$

где $f\colon \mathbb{A} \to \{0,1\}$ — нормированы и задают распределение вероятности на $\mathbb{A}$; множество $\mathbb{A} \equiv \{0,1\}^L$ — булев куб.

- Рандомизированная минимизация эмпирического риска:

$$\mu(A, X)(a) \equiv \frac{[a \in A(X)]}{|A(X)|};$$

# Новое определение вероятности переобучения

- Вероятность переобучения детерминированного метода:

$$Q_\varepsilon(A, \mathbb{X}) = \mathbf{P}\left[\delta(\mu X, X) \geq \varepsilon\right] = \mathbf{P} \sum_{a \in A} [\mu X = a][\delta(a, X) \geq \varepsilon].$$

- Вероятность переобучения для рандомизированного метода обучения:

$$Q_\varepsilon(A, \mathbb{X}) = \mathbf{E} \sum_{a \in A} \mu(A, X, a)[\delta(a, X) \geq \varepsilon],$$

где

$$\mathbf{E} \equiv \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell}, \quad \mu(A, X, a) \equiv \mu(A, X)(a).$$

# Численное сравнение ПМЭР и РМЭР

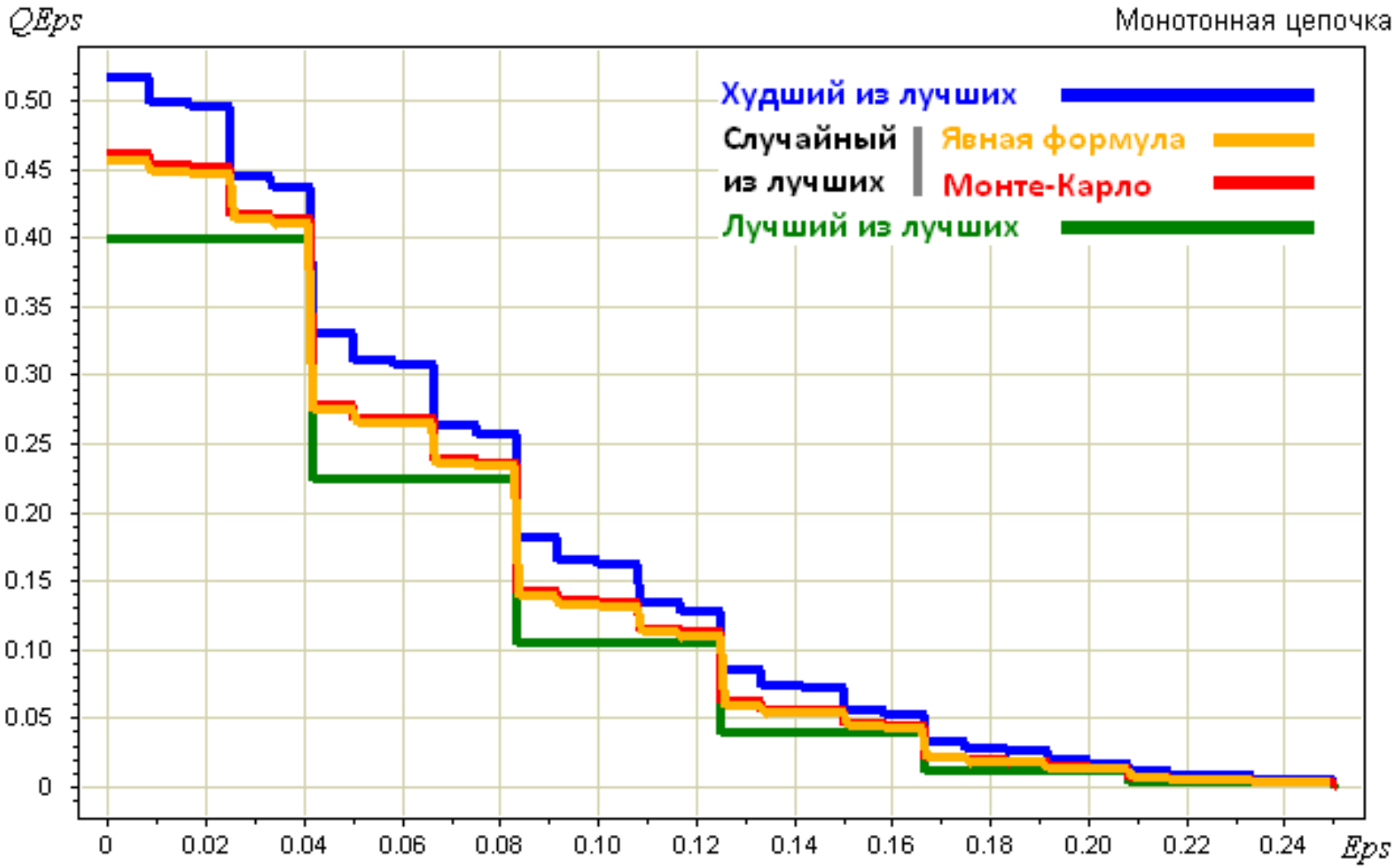



Рис. 1. $L = 100$, $\ell = 60$, $D = 40$, $m = 20$

# Перестановки объектов

$S_L$ — группа перестановок объектов выборки $\mathbb{X} = (x_1, \ldots, x_L)$.

- действие перестановки $\pi \in S_L$ на подмножество объектов:

$$\pi X \equiv \{\pi x \colon x \in X\};$$

- действие перестановки $\pi \in S_L$ на алгоритм:

$$\pi a \equiv \big(I(a, \pi^{-1} x_i)\big)_{i=1}^{L};$$

- действие перестановки $\pi \in S_L$ на множество алгоритмов:

$$\pi A \equiv \{\pi a \colon a \in A\}.$$

# Группа симметрий множества алгоритмов

- Генеральная выборка $\mathbb{X} = \left(x_i\right)_{i=1}^{L}$
- Алгоритм — бинарный вектор $a \equiv \left(a(x_i)\right)_{i=1}^{L}$ длины $L$
- Множество $\mathbb{A} = \{0,1\}^L$ — все алгоритмы длины L
- Аналогия:

| | |
|---|---|
| Точка на плоскости | Алгоритм |
| Плоскость $\mathbb{R}^2$ | Множество всех алгоритмов $\mathbb{A}$ |
| Плоская фигура $F \subset \mathbb{R}^2$ | Множество алгоритмов $A \subset \mathbb{A}$ |
| Группа движений плоскости | Группа перестановок $S_L$ |

**Определение (Группа симметрий)**

*Группой симметрий* $\mathrm{Sym}(A)$ *множества алгоритмов* $A \subset \mathbb{A}$ *назовем его стационарную подгруппу (стабилизатор):*

$$\mathrm{Sym}(A) = \{\pi \in S_L : \pi(A) = A\}.$$

# Упражнение: найти $\mathrm{Sym}(A)$...

... для булева куба:

$$\mathbb{A} = \{0,1\}^L;$$

... для шара алгоритмов:

$$B_r(a_0) = \{a \in \mathbb{A} \colon \rho(a, a_0) \leq r\},$$

где $\rho(a, a_0) = \sum_{i=1}^{L} [I(a, x_i) \neq I(a_0, x_i)]$ — расстояние Хэмминга;
... для цикла алгоритмов:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccccc} a_1 & a_2 & a_3 & a_4 & a_5 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

# Орбиты алгоритмов

- Орбитой элемента $m$ множества $M$, на котором действует группа $G$, называется подмножество $Gm = \{gm\colon g \in G\}$.
- Две орбиты либо не пересекаются, либо совпадают.
- Разбиение на орбиты: $M = Gm_1 \sqcup Gm_2 \sqcup \cdots \sqcup Gm_k$.

- Группа $\mathrm{Sym}(A)$ *действует* на множестве алгоритмов $A$.
- Обозначим $\Omega(A)$ — множества всех орбит, $\omega \in \Omega(A)$ — орбиты.

**Лемма**

*Алгоритмы одной орбиты имеют равное число ошибок на полной выборке.*

# Равный вклад алгоритмов одной орбиты

- Вероятность переобучения — сумма вкладов алгоритмов:

$$Q_\mu(\varepsilon, A) = \sum_{a \in A} Q_\mu(\varepsilon, a, A), \text{ где}$$

$$Q_\mu(\varepsilon, a, A) = \mathbf{E}\mu(A, X, a)[\delta(a, X) \geq \varepsilon];$$

- Алгоритмы одной орбиты дают равный вклад:

$$Q_\mu(\varepsilon, a, A) = Q_\mu(\varepsilon, \pi a, A), \text{ где } \pi \in \mathrm{Sym}(A)$$

- Обозначим $\Omega(A)$ — множество орбит $\mathrm{Sym}(A)$ на $A$;
- Вероятность переобучения с учетом симметрий:

$$Q_\mu(\varepsilon, A) = \sum_{\omega \in \Omega(A)} |\omega| \, \mathbf{E}\mu(A, X, a) \, [\delta(a_\omega, X) \geq \varepsilon] \, . \qquad (1)$$

# Равный вклад разбиений одной орбиты [И. Толстихин]

- Вклад разбиения $X \in [\mathbb{X}]^\ell$ в вероятность переобучения РМЭР:

$$\phi(A, X, \varepsilon) = \frac{1}{|A(X)|} \sum_{a \in A(X)} [\delta(a, X) \geq \varepsilon];$$

$$Q_\mu(\varepsilon, A) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \phi(A, X, \varepsilon);$$

- Разбиения одной орбиты дают равный вклад:

$$\phi(A, X, \varepsilon) = \phi(A, \pi X, \varepsilon), \text{ где } \pi \in \mathrm{Sym}(A);$$

- Обозначим $\Omega(\mathbb{X})$ — множество орбит $\mathrm{Sym}(A)$ на $[\mathbb{X}]^\ell$;
- Вероятность переобучения с учетом симметрий:

$$Q_\mu(\varepsilon, A) = \frac{1}{C_L^\ell} \sum_{\tau \in \Omega(\mathbb{X})} |\tau| \phi(A, X_\tau, \varepsilon).$$

# Центральный слоя хэммингова шара

*Центральным слоем шара* радиуса $r$ называют множество алгоритмов, заданное следующим условием:

$$B_r^m(a_0) = \{a \in \mathbb{A} \colon n(a, \mathbb{X}) = n(a_0, \mathbb{X}) \text{ и } \rho(a, a_0) \leq r\},$$

где $a_0$ — фиксированный алгоритм.

## Теорема (И. Толстихин)

*Вероятность переобучения ПМЭР для центрального слоя шара алгоритмов дается формулой*

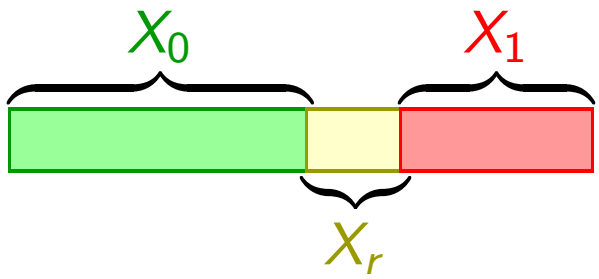
$$Q_\varepsilon(B_r^m(a_0)) = H_L^{\ell,m}\left(s_d(\varepsilon) + \left\lfloor r/2 \right\rfloor\right) \cdot [m \geq \varepsilon k], \qquad (2)$$

*где* $s_d(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$, $H_L^{\ell,m}(s)$ — *функция гипергеометрического распределения.*

# Слой интервала булева куба

Пусть объекты из $\mathbb{X}$ разделены на три множества:

- надежно классифицируемые объекты $X_0$,
- ошибочно классифицируемые объекты $X_1$,
- пограничные объекты $X_r$.




*Слоем интервала булева куба* будем называть множество алгоритмов $B$, такое что

- ни один алгоритм из $B$ не ошибается на $X_0$,
- все алгоритмы из $B$ ошибаются на всех объектах из $X_1$,
- все алгоритмы из $B$ допускают ровно $\rho$ ошибок на $X_r$.

# Слой интервала булева куба

Пусть $\mathbb{X} = X_0 \sqcup X_1 \sqcup X_r$. Обозначим $|X_r| = r$ и $|X_1| = m$, $\rho$ — целочисленный параметр, $\rho \leq r$.

## Теорема

*Вероятность переобучения ПМЭР для слоя интервала булева куба дается формулой*

$$Q_\varepsilon(\hat{B}^m_{r,\rho}) = \frac{1}{C^\ell_L} \sum_{i=0}^{\min(m,\ell)} \sum_{j=0}^{\min(r,\ell-i)} C^i_m C^j_r C^{\ell-i-j}_{L-m-r} \Big[\delta(i,j) \geq \varepsilon\Big], \quad (3)$$

*где* $t(i,j) = i + \max(0, \rho - r - j)$, *и* $\delta(i,j) = \frac{m+\rho-t(i,j)}{k} - \frac{t(i,j)}{\ell}$.

# Разложение и покрытие - Шаг 1.

1. Шаг 1. Разбить множество алгоритмов на кластеры.
2. Шаг 2. Пополнить кластер до множества с известной оценкой.

**Теорема (Фрей, Толстихин)**

*Пусть $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ — разложение $A$ на кластеры. Пусть $A_i \subset B_i$, где в $B_i$ все алгоритмы имеют равное число ошибок. Пусть метод обучения $\mu$ — ПМЭР. Тогда*

$$Q_\varepsilon(A) \leq \sum_{i=1}^{t} Q_\varepsilon(A_i) \leq \sum_{i=1}^{t} Q_\varepsilon(B_i).$$

# Разложение и покрытие - Шаг 2.

1. Шаг 1. Разбить множество алгоритмов на кластеры.
2. Шаг 2. Пополнить кластер до множества с известной оценкой.

**Теорема (Фрей, Толстихин)**

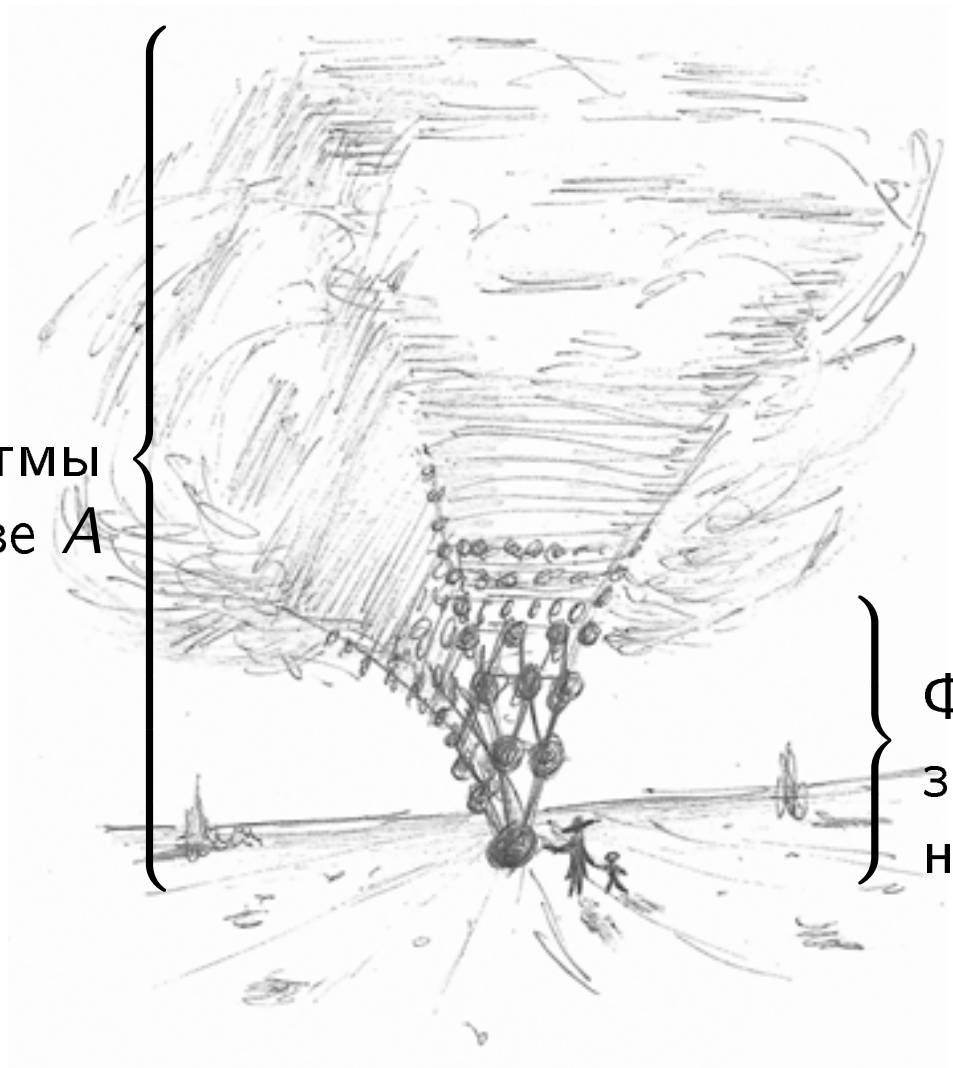
*Пусть $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ — разложение $A$ на кластеры. Пусть $A_i \subset B_i$, где в $B_i$ все алгоритмы имеют равное число ошибок. Пусть метод обучения $\mu$ — ПМЭР. Тогда*

$$Q_\varepsilon(A) \leq \sum_{i=1}^{t} Q_\varepsilon(A_i) \leq \sum_{i=1}^{t} Q_\varepsilon(B_i).$$

# Промежуточные выводы

- Теоретико-групповой метод орбит позволяет выводить оценки вероятности переобучения в тех случаях, когда множество алгоритмов обладает симметрией;
- с помощью доказанных теорем удалось вывести точные формулы для вероятности переобучения РМЭР, примененного к модельным семействам алгоритмов: слой хэммингова шара, слой интервала булева куба;
- полученные формулы вероятности переобучения можно использовать в оценке, основанной на разложении и покрытии множества алгоритмов;
- оценку разложений и покрытий требуется уточнить так, чтобы она учитывала расслоение алгоритмов по числу ошибок (по аналогии с другими комбинаторными оценками, такими как оценка расслоения-связности).

# Расслоение алгоритмов по числу ошибок



Все алгоритмы в семействе $A$

Фактически задействованы только нижние слои $A$

# Пример: монотонная цепь алгоритмов

$$
\begin{array}{lccccccc}
 & x_1 & x_2 & x_3 & & x_D & & \overbrace{\phantom{1,\ldots,1}}^{m} \\
\vec{a}_0 = ( & 0, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 \ ); \\
\vec{a}_1 = ( & 1, & 0, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 \ ); \\
\vec{a}_2 = ( & 1, & 1, & 0, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 \ ); \\
\vec{a}_3 = ( & 1, & 1, & 1, & \ldots & 0, & 0,\ldots,0, & 1,\ldots,1 \ ); \\
\ldots & & & & \ldots & & \ldots & \ldots \\
\vec{a}_D = ( & 1, & 1, & 1, & \ldots & 1, & 0,\ldots,0, & 1,\ldots,1 \ );
\end{array}
$$

## Монотонная цепь

Множество алгоритмов $A = \{a_0, \ldots, a_D\}$ — *монотонная цепь*, если $\rho(a_i, a_{i-1}) = 1$ и $n(a_i) = n(a_0) + i$.

# Порождающие и запрещающие множества

Пусть $\mu$ — пессимистическая минимизация эмпирического риска (выбор алгоритма по принципу «худший из лучших»)

$$A(X) = \operatorname{Arg}\min_{a \in A} n(a, X); \quad \mu X = \arg \max_{a \in A(X)} n(a, \bar{X}).$$

## Факты о монотонной цепи алгоритмов

$$[\mu X{=}a_t] = [x_1, \ldots, x_D \in \bar{X}], \text{ при } t = D,$$

$$[\mu X{=}a_t] = [x_{t+1} \in X][x_1, \ldots, x_t \in \bar{X}], \text{ при } t \leq D,$$

$$P_d = P[\mu X = a_d] = \frac{C_{L-d-1}^{\ell-1}}{C_L^{\ell}},$$

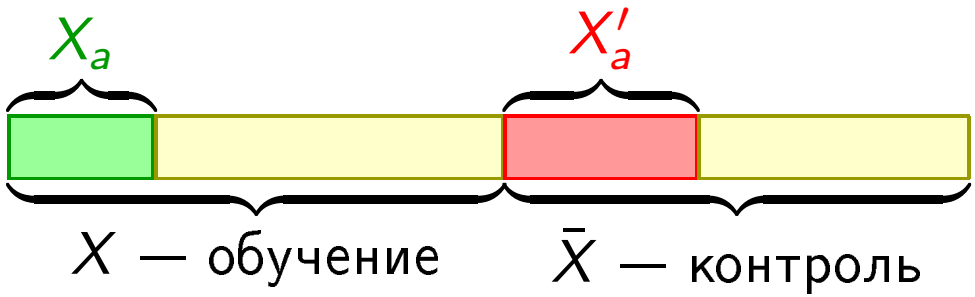
$$Q_\varepsilon(\mu, \mathbb{X}) = \sum_{d=0}^{k} P_d H_{L-d-1}^{\ell-1,m}\left(\tfrac{\ell}{L}(m + d - \varepsilon k)\right).$$

# Гипотеза о порождающих и запрещающих объектах

## Гипотеза

*Для каждого $a \in A$ можно указать пару непересекающихся подмножеств объектов $X_a \subset \mathbb{X}$, $X'_a \subset \mathbb{X}$ такую, что:*

$$\mu X{=}a \Leftrightarrow X_a \subseteq X \text{ и } X'_a \subseteq \bar{X}, \quad \forall X \in [\mathbb{X}]^\ell. \tag{4}$$




**Опр.** $X_a$ — множество объектов, порождающих алгоритм $a$.

**Опр.** $X'_a$ — множество объектов, запрещающих алгоритм $a$.

**Опр.** $\mathbb{X} \setminus X_a \setminus X'_a$ — множество объектов, нейтральных для $a$.

# Оценки на основе порождающих и запрещающих множеств

## Лемма (вероятность получить конкретный алгоритм)

*Если* ***Гипотеза*** *верна, то для любых* $\mu$, $X$, $a \in A$

$$\mathsf{P}\left[\mu X{=}a\right] = P_a = C_{L_a}^{\ell_a} / C_L^{\ell}.$$

*где* $L_a = L - |X_a| - |X'_a|, \quad \ell_a = \ell - |X_a|.$

## Теорема (вероятность переобучения)

*Если* ***Гипотеза*** *верна, то для любых* $\mathbb{X}$, $\mu$, $A$ *и* $\varepsilon \in (0,1)$

$$Q_\varepsilon = \sum_{a \in A} P_a H_{L_a}^{\ell_a, m_a}\left(\tfrac{\ell}{L}\left(n(a,\mathbb{X}) - \varepsilon k\right) - n(a, X_a)\right),$$

*где* $m_a = n(a,\mathbb{X}) - n(a, X_a) - n(a, X'_a).$

# Граф расслоения–связности

**Определим** бинарные отношения на множестве алгоритмов:

- *частичный порядок* $a \leq b$: $I(a,x) \leq I(b,x), \ \forall x \in \mathbb{X}$;
- *предшествование* $a \prec b$: $a \leq b$ и $n(a)+1 = n(b)$.

**Определение (Граф расслоения–связности)**

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

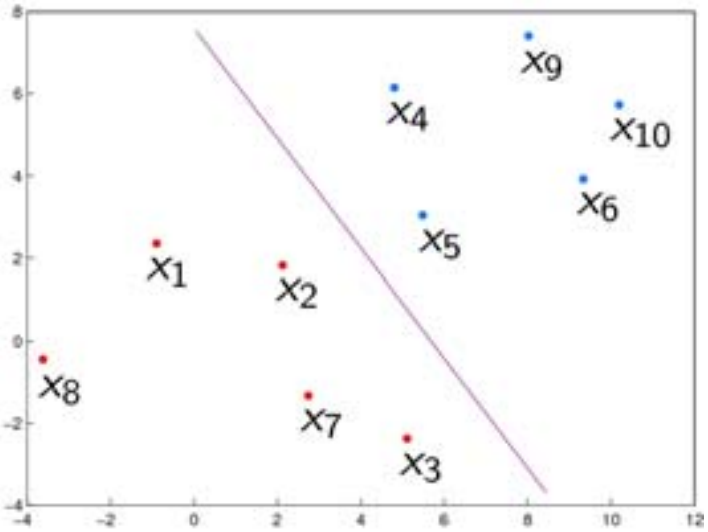
$A$ — *множество попарно различных векторов ошибок;*

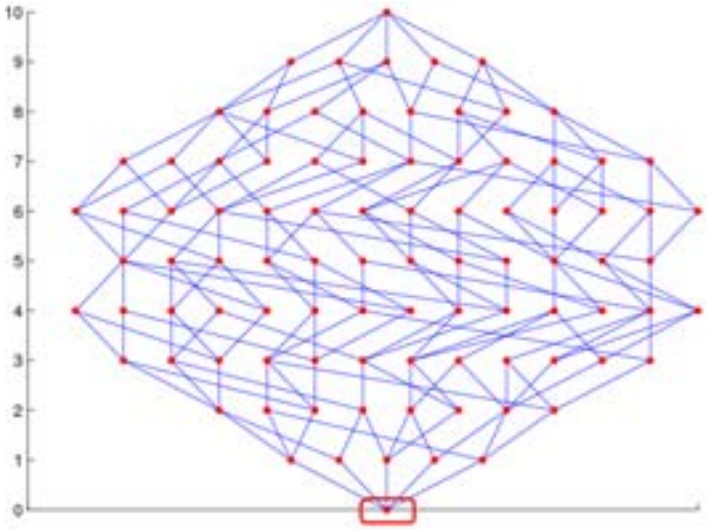
$E = \{(a,b) \colon a \prec b\}$.

**Свойства графа расслоения–связности:**

- каждое ребро $(a,b)$ помечено объектом $x_{ab} \in \mathbb{X}$, для которого $0 = I(a,x_{ab}) < I(b,x_{ab}) = 1$;
- граф является многодольным со слоями $A_m = \{a \in A \colon n(a) = m\}, \ m = 0, \ldots, L+1$;

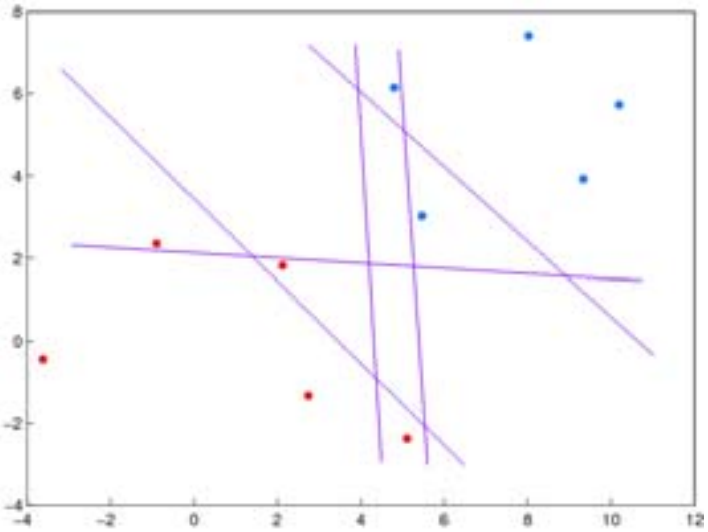
# Пример: граф расслоения-связности

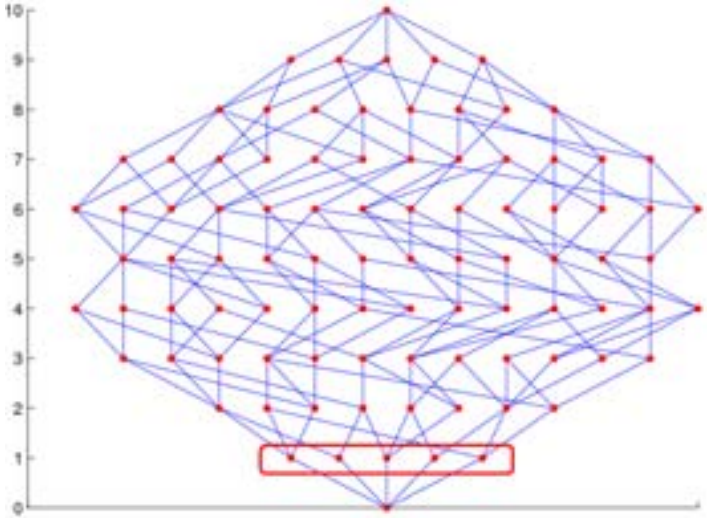


| | слой 0 |
|---|---|
| $x_1$ | 0 |
| $x_2$ | 0 |
| $x_3$ | 0 |
| $x_4$ | 0 |
| $x_5$ | 0 |
| $x_6$ | 0 |
| $x_7$ | 0 |
| $x_8$ | 0 |
| $x_9$ | 0 |
| $x_{10}$ | 0 |

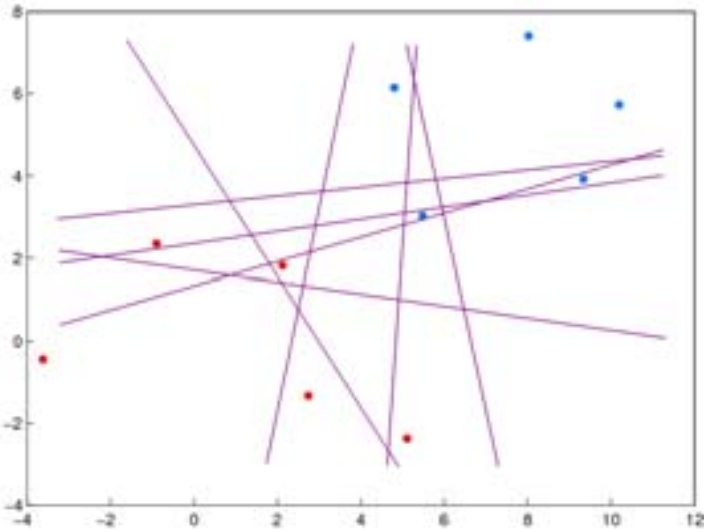
# Пример: граф расслоения-связности

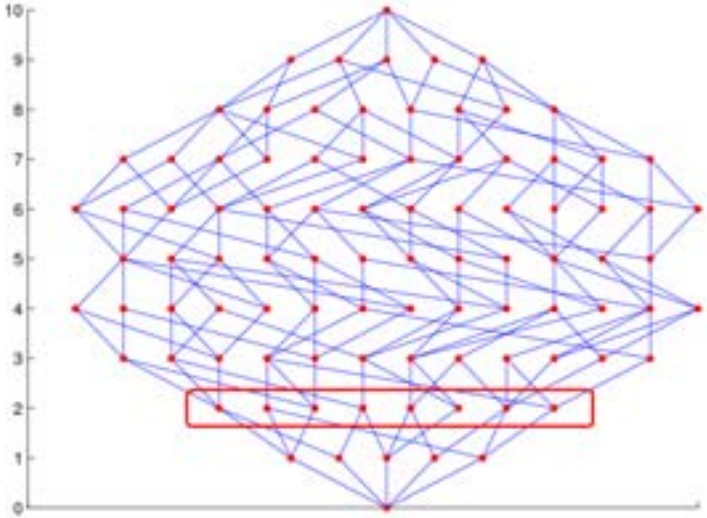




| | слой 0 | слой 1 | | | | |
|---|---|---|---|---|---|---|
| $x_1$ | 0 | 1 | 0 | 0 | 0 | 0 |
| $x_2$ | 0 | 0 | 1 | 0 | 0 | 0 |
| $x_3$ | 0 | 0 | 0 | 1 | 0 | 0 |
| $x_4$ | 0 | 0 | 0 | 0 | 1 | 0 |
| $x_5$ | 0 | 0 | 0 | 0 | 0 | 1 |
| $x_6$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_7$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_8$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

# Пример: граф расслоения-связности





| | 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 | ... |

# Связность и неполноценность

**Опр.** *Связность* алгоритма $a \in A$

$u(a) = \#\{x_{ab} \in \mathbb{X}: a \prec b\}$ — верхняя связность;

$p(a) = \#\{x_{ba} \in \mathbb{X}: b \prec a\}$ — нижняя связность.

**Опр.** *Неполноценность* алгоритма $a \in A$

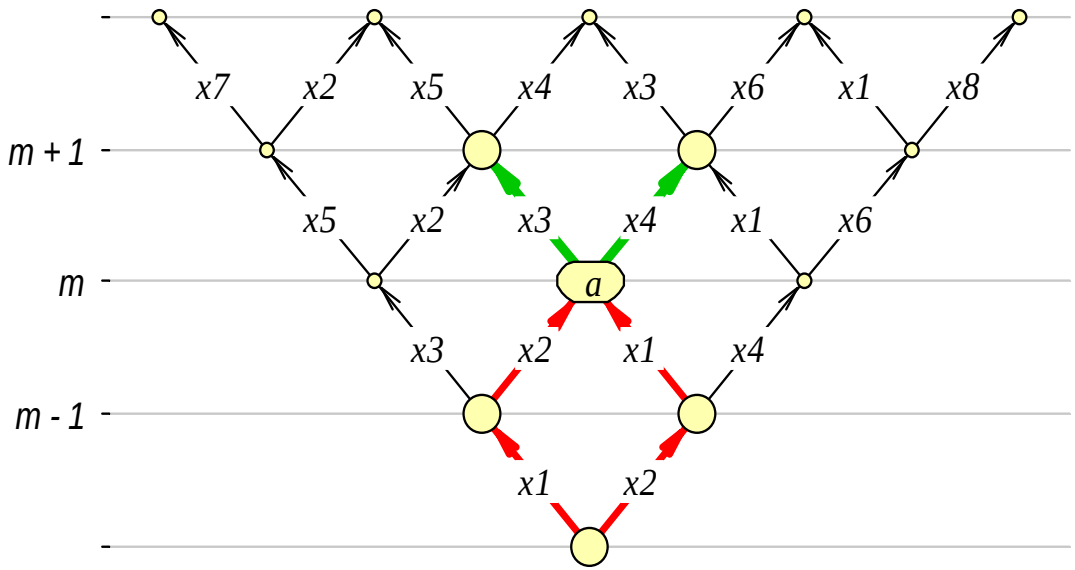
$q(a) = \#\{x_{cb} \in \mathbb{X}: c \prec b \leq a\}, \quad q(a) \leq n(a).$

**Пример:**

$u(a) = \#\{x3, x4\} = 2,$

$p(a) = \#\{x1, x2\} = 2,$

$q(a) = \#\{x1, x2\} = 2.$



# Оценка расслоения-связности

## Теорема (Оценка расслоения-связности)

*Для любых $\mathbb{X}$, $A$, пессимистического МЭР $\mu$ и $\varepsilon \in (0,1)$*

$$Q_\varepsilon \leq \sum_{a \in A} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} H_{L-u-q}^{\ell-u,m-q}(\varepsilon),$$

*где* $m = n(a,\mathbb{X})$, $u = u(a)$, $q = q(a)$,

$$H_L^{\ell,m}(\varepsilon) = \sum_{s=0}^{\lfloor (m-\varepsilon k)\ell/L \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$$ *— правый хвост гипергеометрического распределения.*

# Свойства оценки расслоения-связности

$$Q_\varepsilon \leq \sum_{a \in A} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} H_{L-u-q}^{\ell-u,m-q}(\varepsilon)$$

1. При $|A| = 1$ SC-оценка является точной:

$$Q_\varepsilon = \mathsf{P}\left[\nu(a,\bar{X}) - \nu(a,X) > \varepsilon\right] = H_L^{\ell,m}(\varepsilon) \overset{\ell=k}{\leq} \frac{3}{2} e^{-\varepsilon^2 \ell}$$

2. Замена $u(a) \equiv q(a) \equiv 0$ превращает SC-оценку в VC-оценку:

$$Q_\varepsilon \leq \sum_{a \in A} H_L^{\ell,m}(\varepsilon) \overset{\ell=k}{\leq} |A| \cdot \frac{3}{2} e^{-\varepsilon^2 \ell}$$

# Свойства оценки расслоения-связности

$$Q_\varepsilon \leq \sum_{a \in A} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} H_{L-u-q}^{\ell-u,m-q}(\varepsilon)$$

4. Вероятность реализации алгоритма $a$ в результате обучения:
$$\mathsf{P}[\mu X = a] \leq \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell}$$
5. Вклад $a \in A$ убывает экспоненциально по $u(a) \Rightarrow$ **связные множества меньше переобучаются**; $q(a) \Rightarrow$ **только нижние слои вносят существенный вклад to $Q_\varepsilon$.**
6. Оценка является ***точной*** для некоторых нетривиальных семейств алгоритмов.

# Эксперимент: центральный слой хэммингова шара

$B_r^m$ — центральный слой шара (попарно близкие алгоритмы),
$R_n^m$ — алгоритмы со случайными векторами ошибок.

| $r$ | $\lvert B_r^m \rvert$ | $\lvert R_n^m \rvert$ | $EOF(\mu, \mathbb{X})$ | $\varepsilon$: $Q_\varepsilon(B_r^m) = 0.5$ |
|---|---|---|---|---|
| 2 | 401 | 2 | 0.079 | 0.320 |
| 4 | 35.501 | 7 | 0.160 | 0.400 |
| 6 | 1.221.101 | 39 | 0.240 | 0.400 |
| 8 | 20.413.001 | 378 | 0.319 | 0.400 |

Таблица : Сравнение $\lvert R_n^m \rvert$ и $\lvert B_r^m \rvert$ при $L = 50$, $\ell = 25$, $m = 10$

**Вывод:** Множество из семи «случайных» алгоритмов может переобучиться также сильно, как и множество из 35 тыс. «похожих» алгоритмов.

**Проблема:** Оценка расслоения-связности не ловит этот важный эффект.

# ПЗМ для разложений и покрытий множества алгоритмов

Напомним простую оценку на основе разложения и покрытий:

Теорема. Пусть $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ — разложение $A$ на кластеры. Пусть $A_i \subset B_i$, а метод обучения $\mu$ — ПМЭР. Тогда

$$Q_\varepsilon(A, \mathbb{X}) \leq \sum_{i=1}^{t} Q_\varepsilon(A_i, \mathbb{X}) \leq \sum_{i=1}^{t} Q_\varepsilon(B_i, \mathbb{X}).$$

Аналогичная оценка с учетом эффекта расслоения:

$$Q_\varepsilon(A, \mathbb{X}) \leq \sum_{i=1}^{t} P_i \, Q_{\varepsilon_i}(B_i, \mathbb{Y}_i),$$

где $P_i = C_{L-u_i-q_i}^{\ell-u_i} / C_L^\ell$ — оценка на $\mathsf{P}[\mu X \in A_i]$, $u_i$ — верхняя связность, $q_i$ — неполноценность кластера $A_i$.

# Верхняя связность и неполноценность кластера

Обозначения, не определенные на прошлом слайде:

$u_i \equiv |X_i|$, где $X_i \equiv \bigcap\limits_{a \in A_i} X_a$ – верхняя связность $A_i$,

$q_i \equiv |X'_i|$, где $X'_i \equiv \bigcap\limits_{a \in A_i} X'_a$ – неполноценность $A_i$,

$\mathbb{Y}_i = \mathbb{X} \backslash X_i \backslash X'_i$ – множество нейтральных объектов $A_i$,

$$\varepsilon_i = \frac{L_i}{\ell_i k_i} \frac{\ell\, k}{L} \varepsilon + \Big(1 - \frac{\ell\, L_i}{L \ell_i}\Big) \frac{m_i}{k_i} - \frac{|X'_i|}{k_i},$$

$$Q_\varepsilon(B_i, \mathbb{Y}_i) = \frac{1}{C_{L_i}^{\ell_i}} \sum_{Y \in [\mathbb{Y}_i]^{\ell_i}} [\max_{a \in B_i} \delta(a, Y) \geq \varepsilon],$$

$L_i = L - u_i - q_i$, $\ell_i = \ell - u_i$, $k_i = k - q_i$,

$m_i$ – число ошибок алгоритмов из $A_i$.

# Экспериментальная проверка полученной оценки

| | | Комбинаторные оценки | | | PAC-Bayes | |
|---|---|---|---|---|---|---|
| Task | $EOF(\mu)$ | VC | SC | SS | DI | DD |
| glass | 0.067 | 0.191 | 0.127 | 0.106 | 1.268 | 0.740 |
| Liver dis. | 0.046 | 0.249 | 0.192 | 0.161 | 1.207 | 1.067 |
| Ionosphere | 0.042 | 0.138 | 0.099 | 0.084 | 1.219 | 1.149 |
| Australian | 0.023 | 0.130 | 0.101 | 0.086 | 1.145 | 0.678 |
| pima | 0.021 | 0.151 | 0.117 | 0.098 | 0.971 | 0.749 |
| faults | 0.008 | 0.091 | 0.070 | 0.060 | 1.110 | 1.054 |
| statlog | 0.008 | 0.072 | 0.060 | 0.051 | 1.102 | 0.746 |
| wine | 0.003 | 0.061 | 0.047 | 0.040 | 0.776 | 0.637 |
| waveform | 0.003 | 0.043 | 0.033 | 0.023 | 0.561 | 0.354 |
| pageblocks | 0.003 | 0.030 | 0.022 | 0.018 | 0.739 | 0.186 |
| Optdigits | 0.003 | 0.043 | 0.034 | 0.026 | 1.068 | 0.604 |

# ПЗМ для РМЭР

## Теорема

*Введем множество* $\mathfrak{A}(A) = \{A(X) : X \in [\mathbb{X}]^\ell\}$, *где* $A(X) \equiv \underset{a \in \mathbb{A}}{\operatorname{Argmin}}\, n(a, X)$.

*Пусть для каждого* $\alpha \in \mathfrak{A}(A)$ *существуют порождающее и запрещающее множества* $X_\alpha$ *и* $X'_\alpha$, *такие что:*

$$\big[A(X){=}\alpha\big] = \big[X_\alpha \subseteq X\big]\big[X'_\alpha \subseteq \bar{X}\big], \quad \forall X \in [\mathbb{X}]^\ell.$$

*Тогда вероятность переобучения записывается в виде:*

$$Q_\varepsilon(A) = \sum_{a \in A} \sum_{\alpha \in \mathfrak{A}(A)} \frac{[a \in \alpha]}{|\alpha|} \frac{C^{\ell_\alpha}_{L_\alpha}}{C^\ell_L} H^{\ell_\alpha, m^a_\alpha}_{L_\alpha}(s^a_\alpha(\varepsilon)).$$

- $A_B$ — Связка $h$ монотонных цепей длины $D$,
- $A_M$ — Монотонная сеть размерности $h$
- $A_U$ — Унимодальная сеть размерности $h^{\text{II}}$.

**Теорема (Вероятность переобучения для $A_B$, $A_M$, и $A_U$.)**

$$Q_\varepsilon(A_B, \mathbb{X}) = \sum_{p=0}^{D} \sum_{S=p}^{hD} \sum_{F=0}^{h} \frac{|\omega_p| R^p_{D,h}(S,F)}{1+S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s_0),$$

$$Q_\varepsilon(A_M, \mathbb{X}) = \sum_{\vec{\lambda} \in Y^{h,D}_*} \sum_{\substack{\vec{t} \geq \vec{\lambda}, \\ \|\vec{t}\| \leq D}} \frac{|S_h \vec{\lambda}|}{T(\vec{t})} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s_0),$$

$$Q_\varepsilon(A_U, \mathbb{X}) = \sum_{\vec{\lambda} \in Y^{h,D}_*} \sum_{\substack{\vec{t} \geq \vec{\lambda}, \\ \|\vec{t}\| \leq D}} \sum_{\substack{\vec{t'} \geq \vec{0}, \\ \|\vec{t'}\| \leq D}} \frac{|S_h \vec{\lambda}| \cdot 2^{n(\vec{\lambda})}}{T(\vec{t} + \vec{t'})} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s_0),$$

*где $H^{\ell',m}_{L'}(s_0)$ — гипергеометрическое распределение.*

Новые результаты:

1. Предложен теоретико-групповой метод орбит для вывода оценок вероятности переобучения рандомизированного метода минимизации эмпирического риска;
2. получена общая оценка вероятности переобучения, основанная на разложении и покрытии множества алгоритмов;
3. экспериментальные результаты подтверждают низкую завышенность новой оценки;
4. получены неулучшаемые оценки вероятности переобучения для модельных семейств алгоритмов.