Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Фрей Александр Ильич

Точные оценки обобщающей способности для симметричных множеств алгоритмов и рандомизированных методов обучения

010656 — Математические и информационные технологии

Магистерская диссертация

Научный руководитель:
с.н.с. ВЦ РАН, д.ф.-м.н.
Воронцов Константин Вячеславович

Москва
2010

Содержание

Аннотация

В комбинаторном подходе к проблеме переобучения основной задачей является получение вычислительно эффективных формул для вероятности переобучения. В работе предлагается теоретико-групповой подход, который позволяет проще выводить такие формулы в тех случаях, когда множество алгоритмов наделено некоторой группой симметрий. Приводятся примеры таких множеств. Для рандомизированного метода обучения доказывается общая оценка вероятности переобучения. Показывается её применение для модельных множеств алгоритмов: слоя булева куба, булева куба, связки монотонных цепочек, монотонных и унимодальных сеток произвольной размерности.

1 Введение

При решении задач распознавания образов, восстановления регрессии, прогнозирования всегда возникает проблема выбора по неполной информации. Имея лишь конечную обучающую выборку объектов, требуется из заданного множества алгоритмов выбрать алгоритм, который ошибался бы как можно реже не только на объектах наблюдаемой обучающей выборки, но и на объектах скрытой контрольной выборки, которая в момент выбора алгоритма ещё неизвестна. Если частота ошибок на контрольной выборке оказывается значительно выше, чем на обучающей, то говорят, что произошло «переобучение» (overtraining) или «переподгонка» (overfitting) алгоритма — он слишком хорошо описывает конкретные данные, но не обладает способностью к обобщению этих данных, не восстанавливает порождающую их зависимость и не пригоден для построения прогнозов.

Частоту ошибок на обучающей выборке называют также *эмпирическим риском*. *Минимизация эмпирического риска* — это метод обучения, который выбирает из заданного множества алгоритм, допускающий наименьшее число ошибок на обучающей выборке [1, 2]. В следующей таблице показан пример, когда минимизация эмпирического риска приводит к переобучению. Столбцы таблицы соответствуют алгоритмам, строки — объектам обучающей выборки $\{x_1, x_2, x_3\}$ и контрольной выборки $\{x_4, x_5, x_6\}$. Единица в $[i, d]$-й ячейке таблицы означает, что алгоритм a_d допускает ошибку на объекте x_i.

	a_1	a_2	…	a_d	…	a_D
x_1	0	1	…	0	…	1
x_2	1	1	…	0	…	0
x_3	0	0	…	0	…	0
x_4	1	1	…	1	…	1
x_5	1	0	…	1	…	0
x_6	0	0	…	1	…	0

В данном примере переобучение могло быть следствием «неудачного» разбиения генеральной выборки на обучение и контроль. Поэтому вводится функционал *вероятности переобучения*, равный доле разбиений выборки, при которых возникает переобучение [3, 4]. Этот функционал инвариантен относительно выбора разбиения и характеризует качество данного метода обучения на данной генеральной выборке.

Для некоторых семейств простой структуры (монотонных и унимодальных цепочек и h-мерных сеток) в [3, 5] найдены точные выражения вероятности переобучения. При этом использовалась техника производящих и запрещающих объектов [3]. Применение этой техники для получения формул в ряде других случаев (полный куб алгоритмов, шар алгоритмов) не представлялось возможным.

В данной работе развивается новый, теоретико-групповой подход [6], позволяющий выводить эффективные оценки вероятности переобучения для множеств алгоритмов, обладающих свойствами симметрии. Учет этих свойств позволил на порядки сократить число слагаемых в функционале вероятности переобучения, и тем самым упростить вывод формул в конкретных случаях.

Вместо метода пессимистической минимизации эмпирического риска в работе предлагается исследовать рандомизированный метод. Это позволяет окончательно отказаться от использования контрольной выборки на этапе обучения алгоритмов и дает возможность сравнить оценку худшего и среднего случаев.

В первом разделе работы вводятся формальные определения, включая определения рандомизированного метода минимизации эмпирического риска и определение функционала вероятности переобучения.

Во втором разделе определяется группа симметрии множества алгоритмов, и доказывается точная формула для вероятности переобучения, учитывающая свойства симметрии.

В третьем разделе полученные результаты применяются для вывода явных формул вероятности переобучения для ряда конкретных семейств алгоритмов: полного слоя, полного куба, связки из монотонных цепочек, монотонной и унимодальной сеток произвольной размерности. Приводятся результаты численных экспериментов.

1.1 Определения

Пусть задана генеральная выборка $\mathbb{X} = \left(x_1, \ldots, x_L\right)$, состоящая из L объектов. Произвольный алгоритм классификации, примененный к данной выборке, порождает бинарный вектор ошибок $a \equiv \left(a(x_i)\right)_{i=1}^{L}$, где $a(x_i) = 1$ означает, что алгоритм a допускает ошибку на объекте x_i. Генеральная выборка $\mathbb{X}$ предполагается фиксированной, поэтому алгоритмы отождествляются со своими векторами ошибок.

Обозначим через $\mathbb{A} = \{0,1\}^L$ множество всех возможных векторов ошибок

длины L, тогда $2^{\mathbb{A}}$ — это множество всех подмножеств $\mathbb{A}$. Заметим, что $|\mathbb{A}| = 2^L$, $|2^{\mathbb{A}}| = 2^{2^L}$.

Через $[\mathbb{X}]^\ell$ обозначим множество всех разбиений генеральной выборки $\mathbb{X}$ на обучающую выборку X длины ℓ и контрольную выборку $\bar{X}$ длины $k = L - \ell$.

Число ошибок алгоритма a на выборке $U \subseteq \mathbb{X}$ обозначим через $n(a, U) = \sum\limits_{x \in U} a(x)$.

Детерминированным методом обучения назовем произвольное отображение вида $\mu\colon 2^{\mathbb{A}} \times [\mathbb{X}]^\ell \to \mathbb{A}$. Метод обучения μ по обучающей выборке X выбирает некоторый алгоритм $a = \mu(A, X)$ из подмножества $A \subseteq \mathbb{A}$. Метод обучения называется *минимизацией эмпирического риска*, если возвращаемый им алгоритм допускает наименьшее число ошибок на обучении: для всех $X \in [\mathbb{X}]^\ell$ и $A \subseteq \mathbb{A}$ выполнено $\mu(A, X) \in A(X)$, где

$$A(X) = \operatorname*{Argmin}_{a \in A} n(a, X).$$

При минимизации эмпирического риска может возникать неоднозначность — несколько алгоритмов из $A(X)$ могут иметь одинаковое число ошибок на обучающей выборке. В [4] для устранения неоднозначности и получения точных верхних оценок вероятности переобучения использовалась *пессимистичная* минимизация эмпирического риска — предполагалось, что в случае неоднозначности выбирается алгоритм с наибольшим числом ошибок на генеральной выборке $\mathbb{X}$. Это не устраняет неоднозначность окончательно. Возможны ситуации, когда несколько алгоритмов имеют наименьшее число ошибок на обучающей выборке X и одинаковое число ошибок на генеральной выборке $\mathbb{X}$. В таких случаях на множестве алгоритмов вводился линейный порядок, и среди неразличимых алгоритмов выбирался алгоритм с бóльшим порядковым номером. Введение приоритетности алгоритмов является искусственным приемом, не имеющим адекватных аналогов среди известных методов обучения.

1.2 Рандомизированный метод обучения

Рандомизированный метод обучения произвольному множеству алгоритмов $A \subseteq \mathbb{A}$ и произвольной обучающей выборке $X \in [\mathbb{X}]^\ell$ ставит в соответствие функцию распределения весов на множестве алгоритмов:

$$\mu : 2^{\mathbb{A}} \times [\mathbb{X}]^\ell \to \{f : \mathbb{A} \to [0, 1]\}. \tag{1.1}$$

Естественно полагать, что эта функция нормирована и может быть интерпретирована как вероятность получить каждый алгоритм в результате обучения.

Детерминированный метод обучения является частным случаем рандомизированного, когда функция распределения весов $f(a)$ принимает единичное значение ровно на одном алгоритме и нулевое на всех остальных.

Заметим, что вместо определения (1.1) можно пользоваться эквивалентным способом задать то же самое отображение:

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

Рассмотрим группу S_L — симметрическую группу из L элементов, действующую на множестве объектов генеральной выборки перестановками $S_L = \{\pi \colon \mathbb{X} \to \mathbb{X}\}$.

Для каждого $\pi \in S_L$ определим действие π на произвольную выборку $X \in [\mathbb{X}]^{\ell}$ поэлементным действием отображения $\pi \colon \mathbb{X} \to \mathbb{X}$ на каждый объект выборки X: $\pi X = \{\pi x \colon x \in X\}$. Это отображение не меняет числа объектов: $|X| = |\pi X|$, поэтому можно говорить о действии π на множестве разбиений генеральной выборки на обучение и контроль фиксированной длины $\pi : [\mathbb{X}]^{\ell} \to [\mathbb{X}]^{\ell}$.

Определим действие S_L на множестве всех алгоритмов $\mathbb{A}$ перестановкой координат векторов ошибок алгоритмов: $(\pi a)(x_i) = a(\pi^{-1} x_i)$. Здесь на объекты действует обратная перестановка π^{-1}, поскольку именно в этом случае корректно говорить, что группа S_L *действует* на множестве $\mathbb{A}$.

Лемма 1.1. *Число ошибок алгоритма a на подвыборке $U \subseteq \mathbb{X}$ не меняется от одновременного применения перестановки $\pi \in S_L$ к алгоритму и к подвыборке:*

$$n(a, U) = n(\pi a, \pi U). \tag{1.2}$$

□ **Доказательство.** Запишем определение числа ошибок алгоритма и воспользуемся определенным выше действием перестановки π на алгоритм a:

$$n(\pi a, \pi U) = \sum_{x_i \in \pi U} (\pi a)(x_i) = \sum_{x'_i \in U} (\pi a)(\pi x'_i) =$$
$$\sum_{x'_i \in U} a(\pi^{-1}(\pi x'_i)) = \sum_{x'_i \in U} a(x'_i) = n(a, U). \quad \blacksquare$$

Расстоянием между алгоритмами $\rho(a, a')$ будем называть расстояние Хэмминга между их векторами ошибок:

$$\rho(a, a') = \sum_{x \in \mathbb{X}} |a(x) - a'(x)|.$$

Лемма 1.2. *Произвольная $\pi \in S_L$ является изометрией на множестве алгоритмов:*

$$\rho(a, a') = \rho(\pi a, \pi a').$$

□ **Доказательство.** Рассмотрим алгоритм $b \equiv |a - a'|$. Тогда $\rho(a, a') = n(b, \mathbb{X})$. Непосредственной проверкой убеждаемся, что $\pi b = |\pi a - \pi a'|$. Следовательно, $\rho(\pi a, \pi a') = n(\pi b, \mathbb{X})$. Из леммы 1.1 следует, что $n(b, \mathbb{X}) = n(\pi b, \mathbb{X})$, т.е. $\rho(a, a') = \rho(\pi a, \pi a')$. ■

Действие группы S_L на множестве всевозможных алгоритмов $\mathbb{A}$ естественным образом продолжается до действия на системе всех подмножеств — $S_L \colon 2^{\mathbb{A}} \to 2^{\mathbb{A}}$ по правилу $\pi A = \{\pi a \colon a \in A\}$. В дальнейшем будет использоваться единое обозначение π для описанных выше действий.

Теперь можно дать более строгое определение рандомизированного метода обучения.

Определение 1.1. *Рандомизированным методом обучения назовем отображение вида*

$$\mu : 2^{\mathbb{A}} \times [\mathbb{X}]^{\ell} \times \mathbb{A} \to [0, 1], \tag{1.3}$$

удовлетворяющее при любых $A \in 2^{\mathbb{A}}$, $X \in [\mathbb{X}]^{\ell}$, $a, b \in A$ и $\pi \in S_L$ условиям:

1) *нормировка:*

$$\sum_{a \in A} \mu(A, X, a) = 1; \tag{1.4}$$

2) *неразличимость алгоритмов с одинаковой частотой ошибок на обучении:*

$$n(a, X) = n(b, X) \;\to\; \mu(A, X, a) = \mu(A, X, b); \tag{1.5}$$

3) *инвариантность результата обучения относительно замены множества алгоритмов A на $\pi(A)$:*

$$\mu(A, X, a) = \mu\big(\pi A, \pi X, \pi a\big). \tag{1.6}$$

Первое условие означает «вероятностную» нормировку весов алгоритмов и обеспечивает нулевую «вероятность» алгоритмам, не принадлежащих множеству A. Второе условие означает, что при любом разбиении $\mathbb{X} = X \sqcup \bar{X}$, $X \in [\mathbb{X}]^{\ell}$ вероятность получить алгоритм в результате обучения зависит только от количества ошибок алгоритма на обучении. Третье условие означает, что результат обучения не изменится, если подействовать перестановкой π одновременно и на множество объектов $[\mathbb{X}]^{\ell}$, и на множество алгоритмов $\mathbb{A}$.

Конструктивным примером рандомизированного метода обучения является следующее отображение, которые мы назовем *рандомизированным методом минимизации эмпирического риска*:

$$\mu(A, X, a) = \frac{[a \in A(X)]}{|A(X)|}. \tag{1.7}$$

Тут и далее квадратные скобки — нотация Айверсона [7], переводящая логическое выражение в число 0 или 1 по правилам [истина] = 1, [ложь] = 1.

Лемма 1.3. *Для всех $\pi \in S_L$ алгоритм $a_0 \in A(X)$ тогда и только тогда, когда $\pi a_0 \in (\pi A)(\pi X)$.*

□ **Доказательство.**

Перепишем утверждение леммы в виде

$$a_0 \in \operatorname*{Argmin}_{a \in A} n(a, X) \quad \Leftrightarrow \quad \pi a_0 \in \operatorname*{Argmin}_{a \in \pi A} n(a, \pi X).$$

Используя лемму 1.1, проведем следующую цепочку равносильных утверждений:

$$\begin{aligned}
a_0 \in \operatorname*{Argmin}_{a \in A} n(a, X) \quad &\Leftrightarrow \\
&\Leftrightarrow \quad \forall a \in A \to n(a_0, X) \leqslant n(a, X) \quad \Leftrightarrow \\
&\Leftrightarrow \quad \forall a \in A \to n\big(\pi a_0, \pi X\big) \leqslant n\big(\pi a, \pi X\big) \quad \Leftrightarrow \\
&\Leftrightarrow \quad \forall a' \in \pi A \to n\big(\pi a_0, \pi X\big) \leqslant n\big(a', \pi X\big) \quad \Leftrightarrow \\
&\Leftrightarrow \quad \pi a_0 \in \operatorname*{Argmin}_{a \in \pi A} n(a, \pi X). \quad \blacksquare
\end{aligned}$$

Теорема 1.1. *Отображение* (1.7) *является рандомизированным методом обучения.*

□ **Доказательство.** Первое условие проверяется явно:

$$\sum_{a \in A} \mu(A, X, a) = \sum_{a \in A(X)} \frac{1}{|A(X)|} = 1.$$

Для доказательства второго утверждения достаточно заметить, что два алгоритма a_1 и a_2 с равным числом ошибок на обучении могут лежат в множестве $A(X)$ только одновременно. Следовательно, вероятность получить каждый из алгоритмов в результате обучения равна либо нулю, либо $\frac{1}{|A(X)|}$.

Третье условие непосредственно следует из доказанной выше леммы 1.3.

Теорема доказана. ■

1.3 Вероятность переобучения

Величину $\nu(a,U) = n(a,U)/|U|$ будем называть *частотой ошибок* алгоритма a на выборке U. *Уклонение частот* на разбиении $\mathbb{X} = X \sqcup \bar{X}$ определим как разность частот ошибок на контроле и на обучении: $\delta(a,X) = \nu(a,\bar{X}) - \nu(a,X)$.

Зафиксируем параметр $\varepsilon \in (0,1]$. Будем говорить, что алгоритм a *переобучен* при разбиении $X \sqcup \bar{X}$, если $\delta(a,X) \geqslant \varepsilon$.

Сделаем основное (и единственное) вероятностное предположение, что все разбиения генеральной выборки на наблюдаемую и скрытую подвыборки равновероятны [3, 4].

Если $\varphi\colon [\mathbb{X}]^\ell \to \{\text{истина}, \text{ложь}\}$ — некоторый предикат, то *вероятностью события* $\varphi(X)$ будем называть долю разбиений выборки, при которых предикат $\varphi(X)$ истинен:

$$\mathbf{P}\big[\varphi(X)\big] = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \big[\varphi(X)\big].$$

Соответственно, математическое ожидание произвольной функции $\xi\colon [\mathbb{X}]^\ell \to \mathbb{R}$ есть

$$\mathbf{E}\xi(X) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \xi(X).$$

Вероятностью получить алгоритм $a \in A$ в результате обучения назовем величину

$$P_\mu(a,A) = \mathbf{E}\mu(A,X,a). \tag{1.8}$$

Для произвольного $\varepsilon \in (0,1]$ определим *вклад* алгоритма $a \in A$ в вероятность переобучения:

$$Q_\mu(\varepsilon,a,A) = \mathbf{E}\mu(A,X,a)\big[\delta(a,X) \geqslant \varepsilon\big]. \tag{1.9}$$

Вероятность переобучения определим как сумму вкладов по всем алгоритмам:

$$Q_\mu(\varepsilon,A) = \sum_{a \in A} Q_\mu(\varepsilon,a,A) = \mathbf{E} \sum_{a \in A} \mu(A,X,a)\big[\delta(a,X) \geqslant \varepsilon\big]. \tag{1.10}$$

Для детерминированного метода обучения $\mu\colon 2^{\mathbb{A}} \times [\mathbb{X}]^\ell \to \mathbb{A}$ это определение можно упростить:

$$\begin{aligned} Q_\mu(\varepsilon,A) &= \mathbf{E} \sum_{a \in A} \big[\mu(A,X) = a\big]\big[\delta(a,X) \geqslant \varepsilon\big] = \\ &= \mathbf{E}\big[\delta(\mu(A,X),X) \geqslant \varepsilon\big]. \end{aligned}$$

Полученное выражение буквально означает «долю разбиений выборки на обучение и контроль, при которых выбранный алгоритм $a = \mu(A,X)$ оказался переобученным».

Определение 1.2. *Методы минимизации эмпирического риска*

$$\mu_o X = \arg\min_{a \in A(X)} n(a, \bar{X});$$
$$\mu_p X = \arg\max_{a \in A(X)} n(a, \bar{X});$$

называются, соответственно, оптимистичным и пессимистичным.

Теорема 1.2. *Путь μ — рандомизированный метод минимизации эмпирического риска. Тогда для произвольного множества алгоритмов $A \subseteq \mathbb{A}$ и каждого $\varepsilon \in (0, 1]$ справедлива цепочка неравенств:*

$$Q_{\mu_o}(\varepsilon, A) \leqslant Q_{\mu}(\varepsilon, A) \leqslant Q_{\mu_p}(\varepsilon, A). \tag{1.11}$$

Эта теорема позволяет называть методы μ, μ_p и μ_o соответственно выбором случайного, худшего и лучшего алгоритма из лучших на обучении.

□ **Доказательство.** Для краткости обозначений будем опускать аргумент A у отображений μ_o и μ_p. Покажем, что утверждение верно для каждого разбиения выборки:

$$\big[\delta(\mu_o(X), X) \geqslant \varepsilon\big] \leqslant \sum_{a \in A(X)} \frac{1}{|A(X)|} \big[\delta(a, X) \geqslant \varepsilon\big] \leqslant \big[\delta(\mu_p(X), X) \geqslant \varepsilon\big].$$

Введем обозначения:

$$F_o \equiv \big[\delta(\mu_o(X), X) \geqslant \varepsilon\big];$$
$$F_p \equiv \big[\delta(\mu_p(X), X) \geqslant \varepsilon\big];$$
$$F \equiv \frac{1}{|A(X)|} \sum_{a \in A(X)} \big[\delta(a, X) \geqslant \varepsilon\big].$$

Рассмотрим неравенство $F_o \leqslant F$. Заметим, что F_o может принимать только два значения — 0 и 1, а значение выражения F ограничено отрезком $[0, 1]$. Следовательно, если $F_o = 0$ неравенство выполнено автоматически.

Докажем, что из $F_o = 1$ следует $F = 1$. Обозначим $a_o \equiv \mu_o(X)$. По определению μ_o это значит, что $a_o \in A(X)$ и $\forall a \in A(X)$ выполнено $n(a_o, \bar{X}) \leqslant n(a, \bar{X})$. Следовательно, $\forall a \in A(X)$ выполнено $\delta(a, X) \geqslant \delta(a_o, X) \geqslant \varepsilon$. Значит $F = \sum\limits_{a \in A(X)} \frac{1}{|A(X)|} = 1$.

Для доказательства утверждения $F \leqslant F_p$ достаточно рассмотреть два случая: $F_p = 0$ и $F_p = 1$ и провести аналогичные рассуждения. ■

2 Симметрия множества алгоритмов

Введённые выше понятия позволяют определить группу симметрии множества алгоритмов и с её помощью получать вычислительно эффективные формулы вероятности переобучения.

2.1 Инвариантность вероятности переобучения к действию группы S_L

Определения рассмотренных выше функционалов $P_\mu(a, A)$, $Q_\mu(\varepsilon, a, A)$ и $Q_\mu(\varepsilon, A)$ опирались на упорядоченность объектов в генеральной выборке $\mathbb{X}$. Докажем инвариантность указанных функционалов к изменению нумерации объектов в $\mathbb{X}$.

Для краткости обозначений будем опускать аргумент ε у функции $Q_\mu(\varepsilon, a, A)$.

Лемма 2.1. *Вероятность $P_\mu(a, A)$ получить алгоритм a в результате обучения, а также вклад $Q_\mu(a, A)$ алгоритма a вероятность переобучения сохраняются при одновременном применении произвольной перестановки $\pi \in S_L$ к множеству A и алгоритму a:*

$$P_\mu(a, A) = P_\mu(\pi a, \pi A), \tag{2.1}$$

$$Q_\mu(a, A) = Q_\mu(\pi a, \pi A). \tag{2.2}$$

□ **Доказательство.** Заметим, что для произвольной функции $f(X)$ от разбиения выборки $X \sqcup \bar{X}$ на обучение и контроль выполнено $\mathbf{E} f(X) = \mathbf{E} f(\pi X)$. Воспользуемся также свойством $\delta(\pi a, \pi X) = \delta(a, X)$, которое следует из леммы 1.1 и определения уклонения частот ошибок алгоритма. Тогда

$$\begin{aligned} Q_\mu(\pi a, \pi A) &= \mathbf{E}\mu(\pi A, X, \pi a)\,[\delta(\pi a, X) \geqslant \varepsilon] = \\ &= \mathbf{E}\mu(\pi A, \pi X, \pi a)\,[\delta(\pi a, \pi X) \geqslant \varepsilon] = \\ &= \mathbf{E}\mu(A, X, a)\,[\delta(a, X) \geqslant \varepsilon] = Q_\mu(a, A). \end{aligned}$$

Равенство $P_\mu(\pi a, \pi A) = P_\mu(a, A)$ получается из выражения $Q_\mu(a, A) = Q_\mu(\pi a, \pi A)$ подстановкой $\varepsilon = -1$. ■

Следствие 2.0.1. *Вероятность переобучения сохраняется при применении произвольной перестановки $\pi \in S_L$ к множеству алгоритмов:*

$$Q_\mu(A) = Q_\mu(\pi A). \tag{2.3}$$

□ **Доказательство.**

$$Q_\mu(\pi A) = \sum_{a\in\pi A} Q_\mu(a,\pi A) = \sum_{a\in A} Q_\mu(\pi a,\pi A) = \sum_{a\in A} Q_\mu(a,A) = Q_\mu(A). \ \blacksquare$$

Последнее утверждение выглядит очень естественно, поскольку в большинстве задач обучения по прецедентам порядок объектов в выборке не имеет значения.

2.2 Группа симметрии множества алгоритмов

Напомним, что выше было определено действие группы S_L на множестве всех возможных наборов алгоритмов $2^{\mathbb{A}}$.

Определение 2.1. *Группой симметрий* $\mathrm{Sym}(A)$ *множества алгоритмов* $A \in 2^{\mathbb{A}}$ *будем называть его стационарную подгруппу:*

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

Пример 2.1. Рассмотрим множество алгоритмов, заданное следующей матрицей ошибок:

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

Строки матрицы соответствуют объектам генеральной выборки $\mathbb{X}$, столбцы — алгоритмам $a \in A$. Группа симметрии данного множества алгоритмов является диэдральной группой: $\mathrm{Sym}(A) \cong S_2 \ltimes \mathbb{Z}/5\,\mathbb{Z}$. Образующими элементами группы являются циклическая перестановка $\pi_{\circlearrowleft} = (x_1, x_2, x_3, x_4, x_5) \in S_5$ и пара транспозиций $\pi_{\leftrightarrow} = (x_2, x_5)(x_3, x_4)$.

Важно отметить, что группа симметрии $\mathrm{Sym}(A)$ *действует* на множестве алгоритмов A. Действительно, каждый элемент группы симметрий $\pi \in \mathrm{Sym}(A)$ переставляет алгоритмы a только *внутри* множества A. Значит, для любого $a \in A$ и любого $\pi \in \mathrm{Sym}(A)$ выполнено $\pi a \in A$. Поэтому для группы $\mathrm{Sym}(A)$, в отличии от всей группы S_L, естественным образом определено действие на множестве A.

Орбитой элемента m множества M, на котором действует группа G, называется подмножество $Gm = \{gm\colon g \in G\} \subseteq M$. Орбиты двух элементов m_1 и m_2 либо не пересекаются, либо совпадают. Это позволяет говорить о разбиении множества M на непересекающиеся орбиты: $M = Gm_1 \sqcup \ldots \sqcup Gm_k$.

В дальнейшем будут рассматриваться орбиты действия группы симметрии $\mathrm{Sym}(A)$ на множестве алгоритмов. Совокупность всех орбит множества алгоритмов A обозначим через $\Omega(A)$. Представителя орбиты $\omega \in \Omega(A)$ обозначим через $a_\omega \in A$.

В теории групп точки одной орбиты принято называть эквивалентными. Однако в [1] *эквивалентными алгоритмами* называют алгоритмы с равными векторами ошибок на генеральной выборке $\mathbb{X}$. Поэтому различных представителей одной и той же орбиты будем называть *идентичными алгоритмами*.

Лемма 2.2. *Идентичные алгоритмы имеют равное число ошибок на полной выборке.*

□ Доказательство утверждения автоматически следует из леммы 1.1:

$$n(a, \mathbb{X}) = n(\pi a, \pi\mathbb{X}) = n(\pi a, \mathbb{X}). \quad \blacksquare$$

Согласно данному выше определению *алгоритм* $a \equiv \big(a(x_i)\big)_{i=1}^{L}$ является вектором, следовательно, зависит от нумерации объектов выборки. Однако ни группа симметрий $\mathrm{Sym}(A)$, ни разбиение на классы идентичных алгоритмов $\Omega(A)$, уже не зависят от этой нумерации.

Лемма 2.3. *Для любого множества алгоритмов $A \in 2^{\mathbb{A}}$ и любой перестановки $\pi \in S_L$ группы* $\mathrm{Sym}(A)$ *и* $\mathrm{Sym}(\pi A)$ *сопряжены:* $\mathrm{Sym}(\pi A) = \pi \circ \mathrm{Sym}(A) \circ \pi^{-1}$.

Эта лемма эквивалентна известному утверждению из теории групп: стационарные подгруппы точек, лежащих на одной орбите действия, получаются друг из друга сопряжением [8].

Лемма 2.4. *Пусть алгоритмы a_1 и a_2 идентичны в множестве алгоритмов A. Тогда $\forall \pi \in S_L$ алгоритмы πa_1 и πa_2 идентичны в множестве алгоритмов πA.*

□ Пусть $\gamma \in \mathrm{Sym}(A)$ — перестановка, такая что $a_2 = \gamma a_1$. Тогда $\pi a_2 = \pi\gamma a_1 = (\pi\gamma\pi^{-1})\pi a_1 = \tilde{\gamma}\pi a_1$. Из леммы 2.3 получаем, что $\tilde{\gamma} = \pi\gamma\pi^{-1}$ — элемент $\mathrm{Sym}(\pi A)$. ■

2.3 Теоремы о равном вкладе идентичных алгоритмов в вероятность переобучения

Теоремы, приведенные в данном параграфе, позволяют в ряде случаев существенно упростить получение явных формул для вероятности переобучения.

Теорема 2.1. *Идентичные алгоритмы имеют равную вероятность реализоваться в результате обучения, а также дают равный вклад в вероятность переобучения:*

$$P_\mu(a, A) = P_\mu(\pi a, A), \tag{2.4}$$

$$Q_\mu(a, A) = Q_\mu(\pi a, A), \tag{2.5}$$

где $\pi \in \mathrm{Sym}(A)$.

□ Доказательство автоматически следует из леммы 2.1 и определения группы симметрии: $P_\mu(\pi a, A) = P_\mu(\pi a, \pi A) = P_\mu(a, A)$, и аналогично для $Q_\mu(a, A)$. ■

Следствие 2.1.1. *Пусть группа симметрии действует на множестве алгоритмов транзитивно:* $A = \{\pi a_0,\ \pi \in \mathrm{Sym}(A)\}$, *где* $a_0 \in A$ *— произвольный алгоритм множества* A. *Тогда все алгоритмы множества имеют равную вероятность реализоваться в результате обучения.*

Теорема 2.1 позволяет перейти от суммирования по всем алгоритмам множества к суммированию по орбитам действия группы $\mathrm{Sym}(A)$.

Теорема 2.2. *Вероятность переобучения* $Q_\mu(A)$ *для рандомизированного метода минимизации эмпирического риска можно записать в следующем виде:*

$$Q_\mu(A) = \sum_{\omega \in \Omega(A)} |\omega|\, \mathbf{E} \frac{\big[a_\omega \in A(X)\big]}{|A(X)|} \left[\delta(a_\omega, X) \geqslant \varepsilon\right]. \tag{2.6}$$

□ Воспользуемся теоремой о равном вкладе идентичных алгоритмов в вероятность переобучения, затем определениями (1.9) и (1.7):

$$Q_\mu(A) = \sum_{a \in A} Q_\mu(a, A) = \sum_{\omega \in \Omega(A)} |\omega|\, Q_\mu(a_\omega, A) =$$

$$= \sum_{\omega \in \Omega(A)} |\omega|\, \mathbf{E} \frac{\big[a_\omega \in A(X)\big]}{|A(X)|} \left[\delta(a_\omega, X) \geqslant \varepsilon\right]. \quad \blacksquare$$

Формула (2.6) является основным инструментом вывода точных оценок вероятности переобучения для рандомизированного метода минимизации эмпирического риска.

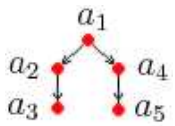
2.4 Группа автоморфизмов графа смежности множества алгоритмов

Графом смежности множества алгоритмов A назовем направленный ациклический граф $T(A) = (A, E)$, вершины которого соответствуют алгоритмам из A, а ребро $(a_1, a_2) \in E$ соединяет пары алгоритмов, чьи вектора ошибок отличаются только на одном объекте: $\rho(a_1, a_2) = 1$, причем число ошибок алгоритма a_2 на единицу больше, чем у a_1.

Пример 2.2. Рассмотрим множество алгоритмов, заданное следующей матрицей ошибок:

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

Данное множество алгоритмов мы будем называть унимодальной цепочкой. Нетрудно убедиться, что граф смежности данного множества будет состоять из ребер $E = \{(a_1, a_2), (a_2, a_3), (a_1, a_4), (a_4, a_5)\}$:



Заметим, что разным множествам алгоритмов могут соответствовать изоморфные графы смежности.

Пример 2.3. Рассмотрим множество алгоритмов, заданное следующей матрицей ошибок:

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

Его граф смежности изоморфен графу унимодальной цепочки, рассмотренной в предыдущем примере.

Таким образом, невозможно восстановить множество алгоритмов по его графу смежности. Граф смежности сохраняет информацию только о близком со-

седстве алгоритмов. Тем не менее граф смежности остается самым естественным способом визуализировать множество алгоритмов.

Определение 2.2. *Группой автоморфизмов графа смежности $T(A) = (A, E)$ множества алгоритмов A называют максимальную подгруппу $Aut(T(A))$ группы перестановок вершин графа, такую что каждый ее элемент $\pi \in Aut(T(A))$ удовлетворяет двум условиям:*

- *Сохранение ребер графа и их ориентации:*

$$(a_1, a_2) \in E \to (\pi a_1, \pi a_2) \in E; \tag{2.7}$$

- *Сохранение числа ошибок алгоритмов:*

$$n(a, \mathbb{X}) = n(\pi a, \mathbb{X}). \tag{2.8}$$

Данное определение использует, помимо структуры графа, дополнительную информацию о числе ошибок алгоритмов. Ниже мы покажем, что для широкого класса *связных графов* группа автоморфизмов зависит только от структуры самого графа.

Определение 2.3. *Граф смежности $T(A) = (A, E)$ назовем связным, если в соответствующем неориентированном графе существует путь между каждой парой вершин: для всех $a, a' \in A$ существует конечная последовательность вершин $\{a_i\}_{i=1}^n$, такая что $a = a_1$, $a' = a_n$, и для всех $i = 2, \ldots, n$ одно из ребер (a_{i-1}, a_i) или (a_i, a_{i-1}) лежит в E.*

Теорема 2.3. *Пусть граф смежности $T(A) = (A, E)$ связен. Тогда в определении 2.2 условие $n(a, \mathbb{X}) = n(\pi a, \mathbb{X})$ выполняется автоматически.*

□ **Доказательство.**

Для краткости обозначений будем опускать аргумент $\mathbb{X}$ у функционала числа ошибок: $n(a, \mathbb{X}) \equiv n(a)$. Рассмотрим произвольную перестановку вершин графа π, такую что для всех ребер $(a, a') \in E$ выполнено $(\pi a, \pi a') \in E$. Покажем, что для всех $a \in A$ выполнено $n(a) = n(\pi a)$.

Шаг 1. Покажем, что перестановка π сохраняет разности между числом ошибок любой пары алгоритмов: $n(a') - n(a) = n(\pi a') - n(\pi a)$. Для этого рассмотрим путь $\{a_i\}_{i=1}^n$, который соединяет вершины $a = a_1$ и $a' = a_n$.

Пусть $\sigma \in \{0, 1\}$. Введем обозначение

$$(a_{i-1}, a_i)^\sigma = \begin{cases} (a_{i-1}, a_i), \text{ при } \sigma = 0, \\ (a_i, a_{i-1}), \text{ при } \sigma = 1. \end{cases}$$

Тогда, согласно определению связного графа, существует конечная последовательность $\{\sigma_i\}_{i=2}^n$, такая что для всех $i = 2, \ldots, n$ ребро $(a_{i-1}, a_i)^{\sigma_i} \in E$. Непосредственной проверкой убеждаемся, что разность $n(a') - n(a)$ записывается в виде

$$n(a') - n(a) = \sum_{i=2}^{n} [\sigma_i = 0] - \sum_{i=2}^{n} [\sigma_i = 1].$$

Заметим, что конечная последовательность $\{\pi a_i\}_{i=1}^n$ задает путь между вершинами $\pi a = \pi a_1$ и $\pi a' = \pi a_n$, причем ребро $(\pi a_{i-1}, \pi a_i)^{\sigma_i} \in E$. Значит, $n(\pi a') - n(\pi a)$ дается тем же выражением:

$$n(\pi a') - n(\pi a) = \sum_{i=2}^{n} [\sigma_i = 0] - \sum_{i=2}^{n} [\sigma_i = 1].$$

Шаг 2. Покажем, что существует алгоритм $a \in A$, такой что $n(\pi a) = n(a)$. Рассмотрим алгоритм с минимальным и максимальным числом ошибок:

$$a_{min} \in \operatorname*{Argmin}_{a \in A} n(a),$$

$$a_{max} \in \operatorname*{Argmax}_{a \in A} n(a).$$

Для них выполнено $n(a_{max}) - n(a_{min}) = n(\pi a_{max}) - n(\pi a_{min})$. Выразим отсюда $n(\pi a_{max})$:

$$n(\pi a_{max}) = n(a_{max}) + n(\pi a_{min}) - n(a_{min}).$$

Разность $n(\pi a_{min}) - n(a_{min})$ неотрицательна, следовательно $n(\pi a_{max}) \geqslant n(a_{max})$. Но с другой стороны $a_{max} \in \operatorname*{Argmax}_{a \in A} n(a)$. Поэтому $n(\pi a_{max}) = n(a_{max})$.

Соединяя вместе результаты, полученные на первом и втором шаге, приходим к утверждению теоремы. ■

Следующая теорема устанавливает связь между группой симметрии множества алгоритмов и группой автоморфизмов соответствующего графа смежности. В ряде случаев это утверждение помогает угадать орбиты действия группы симметрии $\mathrm{Sym}(A)$.

Теорема 2.4. *Орбиты действия группы симметрии* $\mathrm{Sym}(A)$ *на множестве алгоритмов* A *вложены в орбиты действия группы автоморфизмов* $Aut(T(A))$ *на* A.

□ **Доказательство.**

Рассмотрим пару алгоритмов a, a' из одной орбиты действия группы $\mathrm{Sym}(A)$. Это значит, что существует $\pi \in \mathrm{Sym}(A)$, такая что $\pi a = a'$. Каждый элемент группы $\mathrm{Sym}(A)$ действует на A, следовательно можно рассматривать π и как

элемент группы перестановок вершин A. Нам необходимо показать, что $\pi \in Aut(T(A))$.

Рассмотрим произвольное ребро $(a, a') \in E$. Рассматривая перестановку π как элемент симметрической группы S_L, применим леммы 1.1 и 1.2 к алгоритмам a, a'. Получим, что $n(a, \mathbb{X}) = n(\pi a, \mathbb{X})$, $n(a', \mathbb{X}) = n(\pi a', \mathbb{X})$, $\rho(\pi a, \pi a') = \rho(a, a') = 1$. Следовательно, $(\pi a, \pi a') \in E$, а значит $\pi \in Aut(T(A))$. ■

2.5 Орбиты разбиений выборки

Напомним, что вероятность переобучения для рандомизированного метода минимизации эмпирического риска может быть записана следующим образом:

$$Q_\mu(\varepsilon, A) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \sum_{a \in A(X)} \frac{\big[\delta(a, X) \geqslant \varepsilon\big]}{|A(X)|},$$

где $A(X) = \underset{a \in A}{\operatorname{Argmin}}\, n(a, X)$.

Заметим, что коэффициент $\frac{1}{|A(X)|}$ не зависит алгоритма a, и потому может быть вынесен за знак суммирования по $a \in A$:

$$Q_\mu(\varepsilon, A) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \frac{1}{|A(X)|} \sum_{a \in A(X)} \big[\delta(a, X) \geqslant \varepsilon\big]$$

Обозначим через $\Delta_A^\varepsilon(X) \subset A$ множество алгоритмов из A, переобученных на разбиении X: $\Delta_A^\varepsilon(X) = \{a \in A \colon \delta(a, X) \geqslant \varepsilon\}$.

Тогда

$$Q_\mu(\varepsilon, A) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \frac{|\Delta_{A(X)}^\varepsilon(X)|}{|A(X)|}. \tag{2.9}$$

В этой интерпретации вероятность переобучения — это усредненная по всем разбиениям доля переобученных алгоритмов в $A(X)$.

Лемма 2.5. *В множестве $A(X)$ переобученными на разбиении X являются те и только те алгоритмы, у которых число ошибок на полной выборке не меньше порога $m_\delta(A, X) = \varepsilon\ell + \frac{L}{k} \min\limits_{a \in A} n(a, X)$.*

□ **Доказательство.**

По определению $A(X)$ все алгоритмы $a \in A(X)$ имеют равное число ошибок на обучении. Обозначим это число через $n_0 = \min\limits_{a \in A} n(a, X)$. Тогда $\delta(a, X) = \frac{n(a, \mathbb{X}) - n_0}{\ell} - \frac{n_0}{k}$, и неравенство $\delta(a, X) \geqslant \varepsilon$ можно записать как $n(a, \mathbb{X}) \geqslant \varepsilon\ell + \frac{L}{k} \min\limits_{a \in A} n(a, X) \equiv m_\delta(A, X)$. Это число означает минимальное количество ошибок на полной выборке, начиная с которого алгоритмы из $A(X)$ будут переобученными. ■

Напомним, что определенная выше группа симметрий $\mathrm{Sym}(A)$ являлась подгруппой в S_L, и потому действовала на $[\mathbb{X}]^\ell$. Обозначим орбиты этого действия через $\Omega([\mathbb{X}]^\ell)$. Произвольного представителя орбиты $\tau \in \Omega([\mathbb{X}]^\ell)$ обозначим через X_τ.

Теорема 2.5. *Вероятность переобучения для рандомизированного метода минимизации эмпирического риска можно записать в виде*

$$Q_\mu(\varepsilon, A) = \frac{1}{C_L^\ell} \sum_{\tau \in \Omega([\mathbb{X}]^\ell)} |\tau| \frac{|\Delta^\varepsilon_{A(X)}(X)|}{|A(X)|}. \tag{2.10}$$

□ **Доказательство.**

Достаточно показать, что для каждой $\pi \in \mathrm{Sym}(A)$ и разбиения $X \in [\mathbb{X}]^\ell$ выполнено $|A(X)| = |A(\pi X)|$ и $|\Delta^\varepsilon_{A(X)}(X)| = |\Delta^\varepsilon_{A(\pi X)}(\pi X)|$.

Для доказательства обоих тождеств нам понадобится лемма 1.3 о инвариантности множества $A(X)$ к действию $\pi \in S_L$. Данная лемма утверждает, что $a_0 \in A(X) \Leftrightarrow \pi a_0 \in (\pi A)(\pi X)$. Для $\pi \in \mathrm{Sym}(A)$ выполнено $\pi A = A$, значит $a_0 \in A(X) \Leftrightarrow \pi a_0 \in A(\pi X)$. Это устанавливает взаимно-однозначное соответствие между множествами $A(X)$ и $A(\pi X)$. Следовательно $|A(X)| = |A(\pi X)|$.

Пусть $a_0 \in \Delta^\varepsilon_{A(X)}(X)$. Тогда $\pi a_0 \in A(\pi X)$. Согласно лемме 1.1 имеем $n(a_0, \mathbb{X}) = n(\pi a_0, \mathbb{X})$. Воспользовавшись леммой 2.5 получаем, что $\pi a_0 \in \Delta^\varepsilon_{A(\pi X)}(\pi X)$. Тем самым установлено соответствие между множествами $\Delta^\varepsilon_{A(X)}(X)$ и $\Delta^\varepsilon_{A(\pi X)}(\pi X)$. Данное соответствие является взаимно-однозначным, поскольку в группе $\mathrm{Sym}(A)$ существует обратный элемент π^{-1}. Следовательно, мощности рассматриваемых множеств совпадают. ■

3 Точные оценки вероятности переобучения

В данном параграфе будут получены явные комбинаторные формулы для функционала $Q_\mu(\varepsilon, A)$ для некоторых множеств алгоритмов A, обладающих свойством симметрии.

3.1 Полный слой алгоритмов

Полным m-слоем алгоритмов будем называть множество, состоящее из всех алгоритмов $a \in \mathbb{A}$ с фиксированным числом ошибок: $n(a, \mathbb{X}) = m$.

Теорема 3.1. *При обучении рандомизированным методом минимизации эмпирического риска вероятность переобучения для полного m-слоя алгоритмов*

есть

$$Q_\mu(\varepsilon, A) = \big[\varepsilon k \leqslant m \leqslant L - \varepsilon\ell\big]. \tag{3.1}$$

□ **Доказательство.**

В рассматриваемом случае группой симметрии $\mathrm{Sym}(A)$ будет вся симметрическая группа S_L. Следовательно, действие группы симметрии на множестве алгоритмов транзитивно, и в рассматриваемом множестве есть только один класс из C_L^m идентичных алгоритмов. Согласно теореме 2.2 запишем:

$$Q_\mu(\varepsilon, A) = C_L^m \mathbf{E} \frac{\big[a_0 \in A(X)\big]}{|A(X)|} \left[\delta(a_0, X) \geqslant \varepsilon\right].$$

где a_0 — произвольный алгоритм рассматриваемого семейства.

Алгоритм a_0 будет выбран только если он имеет минимальное число ошибок на обучении. Рассмотрим два случая.

Случай 1, $m \leqslant k$. Все ошибки a_0 помещаются в контроль, и переобучение наступает при условии $m \geqslant \varepsilon k$. Этим фиксируются m объектов контроля, следовательно число слагаемых в сумме по разбиениям X определяется числом способов выбрать $k - m$ объектов, на которых алгоритм a_0 не ошибается. Это число равно C_{L-m}^{k-m}.

Мощность множества лучших на обучении алгоритмов $A(X)$ не зависит от X и равна C_k^m — числу способов расставить m ошибок алгоритма на k позициях контрольной выборки. Таким образом,

$$Q_\mu(\varepsilon, A) = \frac{C_L^m}{C_L^\ell} \frac{C_{L-m}^{k-m}}{C_k^m} \big[m \geqslant \varepsilon k\big], \text{ при } m \leqslant k.$$

Случай 2, $m > k$. Контрольная выборка должна содержать только объекты, на которых a_0 ошибается. Тогда в обучении останется $m - k$ ошибок, а условие переобучения примет вид $1 - \frac{m-k}{\ell} \geqslant \varepsilon$, откуда $m \leqslant L - \varepsilon\ell$.

Число разбиений выборки, при которых $a_0 \in A(X)$, равно C_m^k — числу способов выбрать k ошибок алгоритма a_0 в контрольную выборку. Мощность множества $A(X)$ вновь не зависит от X, и равна C_ℓ^{m-k} — числу способов отобрать $m - k$ ошибок в обучающую выборку.

$$Q_\mu(\varepsilon, A) = \frac{C_L^m}{C_L^\ell} \frac{C_\ell^{m-k}}{C_m^k} \big[m \leqslant L - \varepsilon\ell\big], \text{ при } m > k.$$

Записав для каждого комбинаторного коэффициента тождество $C_L^k = \frac{L!}{k!(L-k)!}$, убеждаемся, что в обеих формулах комбинаторные множители равны единице. Соединяя вместе условия $\varepsilon k \leqslant m \leqslant k$ и $k < m \leqslant L = \varepsilon\ell$, получаем утверждение теоремы. ■

3.2 Куб алгоритмов

Кубом алгоритмов $\mathbb{A}$ называется множество, содержащее все возможные $a \in \{0,1\}^L$.

Теорема 3.2. *Вероятность переобучения для куба алгоритмов дается формулой:*

$$Q_\mu(\varepsilon, \mathbb{A}) = \frac{1}{2^k} \sum_{m=\lceil \varepsilon k \rceil}^{k} C_k^m.$$

□ **Доказательство.**

Очевидно, что в данном случае группа симметрии — это вся S_L. Тогда орбитами ее действия будут слои алгоритмов с одинаковым числом ошибок. Поэтому, согласно теореме 2.2,

$$Q_\mu(\varepsilon, \mathbb{A}) = \sum_{m=0}^{L} C_L^m \mathbf{E} \frac{[a_m \in A(X)]}{|A(X)|} [\delta(a, X) \geqslant \varepsilon].$$

Алгоритм может быть выбран в результате обучения только в том случае, когда он не допускает ошибок на обучении. Поэтому все его ошибки должны помещаться в контрольную выборку, значит можно ограничить индекс суммирования $m \leqslant k$.

Раз все ошибки выбранного алгоритма расположены в контрольной выборке, то, вне зависимости от разбиения, уклонение частот равно $\delta(a, X) = \frac{m}{k}$. Следовательно, переобучение наступает при $m \geqslant \lceil \varepsilon k \rceil$.

В множестве $A(X)$ всегда 2^k алгоритмов. Это алгоритмы с нулевым числом ошибок на обучении и всеми возможными векторами ошибок на контрольной выборке.

Собирая вместе установленные выше факты, получаем формулу

$$Q_\mu(\varepsilon, \mathbb{A}) = \sum_{m=\lceil \varepsilon k \rceil}^{k} C_L^m \frac{\mathbf{E}[a_m \in A(X)]}{2^k}.$$

Осталось вычислить число разбиений, на которых алгоритм a_m будет выбран методом обучения. Этих разбиений столько, сколько способов выбрать ℓ объектов обучающей выборки из $L-m$ правильных ответов алгоритма a_m. Итого получаем

$$Q_\mu(\varepsilon, \mathbb{A}) = \sum_{m=\lceil \varepsilon k \rceil}^{k} C_L^m \frac{C_{L-m}^{\ell}}{C_L^{\ell} 2^k} = \frac{1}{2^k} \sum_{m=\lceil \varepsilon k \rceil}^{k} C_k^m. \quad \blacksquare$$

3.3 Унимодальная цепочка

Напомним, что расстояние между алгоритмами $\rho(a, a')$ определялось как расстояние Хэмминга между их векторами ошибок:

$$\rho(a, a') = \sum_{x \in \mathbb{X}} |a(x) - a'(x)|.$$

Определение 3.1. *Множество алгоритмов $\{a_0, \ldots, a_D\}$ называется унимодальной цепочкой, если выполнены два условия:*

1) *монотонность числа ошибок: $n(a_i, \mathbb{X}) = m + i, \;\; i = 0, \ldots, D$ при некотором фиксированном m;*
2) *поглощение ошибок предыдущего алгоритма: $\rho(a_i, a_{i-1}) = 1, \;\; i = 1, \ldots, D$.*

Таким образом, в монотонной цепочке каждый следующий алгоритм ошибается на тех же объектах, что и предыдущий, и допускает еще одну дополнительную ошибку.

Монотонная цепочка алгоритмов — это простейшая модель однопараметрического *связного семейства алгоритмов*, предполагающая, что при непрерывном удалении некоторого параметра от оптимального значения число ошибок на полной выборке только увеличивается.

Определение 3.2. *Множество алгоритмов $\{a_0, a_1, \ldots, a_D, a'_1, \ldots, a'_D\}$ называется унимодальной цепочкой, если выполнены два условия:*

1) *левая ветвь $\{a_0, a_1, \ldots, a_D\}$ и правая ветвь $\{a_0, a'_1, \ldots, a'_D\}$ являются монотонными цепочками.*
2) *пересечение множества ошибок алгоритмов a_D и a'_D равно множеству ошибок алгоритма a_0.*

Унимодальная цепочка является более реалистичной моделью однопараметрического *связного семейства*, по сравнению с монотонной цепочкой. Если мы имеем лучший алгоритм a_0 с оптимальным значением некоторого вещественного параметра, то отклонение значения этого параметра как в бóльшую, так и в меньшую, сторону приводит к увеличению числа ошибок.

Теорема 3.3. *Для унимодальной цепочки с ветвями длины D вероятность переобучения рандомизированного метода минимизации эмпирического риска равна*

$$Q_\mu(\varepsilon, A) = \sum_{h=0}^{D} \sum_{t_1=h}^{D} \sum_{t_2=0}^{D} \frac{|\omega_h|}{1 + t_1 + t_2} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell', m}(s(\varepsilon)), \tag{3.2}$$

где $L' = L - t_1 - t_2 - F$, $F = [t_1 \neq D] + [t_2 \neq D]$, $\ell' = \ell - F$, $s(\varepsilon) = \left\lfloor \frac{\ell}{L}(m + h - \varepsilon k) \right\rfloor$; $|\omega_h| = 1$ *при* $h = 0$ *и* $|\omega_h| = 2$ *при* $h \geqslant 1$; $H_{L'}^{\ell',m}(z) = \frac{1}{C_{L'}^{\ell'}} \sum\limits_{s=0}^{\lfloor z \rfloor} C_m^s C_{L'-m}^{\ell'-s}$ *— функция гипергеометрического распределения [4].*

□ **Доказательство.**

Пронумеруем объекты генеральной выборки $\mathbb{X}$ таким образом, как показано в следующей таблице:

	a_0	a_1	a_2	$\cdots$	a_D	a'_1	a'_2	$\cdots$	a'_D
x_1	0	1	1	$\cdots$	1	0	0	$\cdots$	0
x_2	0	0	1	$\cdots$	1	0	0	$\cdots$	0
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_D	0	0	0	$\cdots$	1	0	0	$\cdots$	0
x'_1	0	0	0	$\cdots$	0	1	1	$\cdots$	1
x'_2	0	0	0	$\cdots$	0	0	1	$\cdots$	1
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x'_D	0	0	0	$\cdots$	0	0	0	$\cdots$	1

Перестановками объектов выборки $(x_1 \leftrightarrow x'_1, \ldots, x_D \leftrightarrow x'_D)$ можно поменять левую и правую ветви местами. Поэтому идентичные алгоритмы в унимодальной цепочке — это пары алгоритмов с равным числом ошибок на полной выборке.

Согласно теореме 2.2 вероятность переобучения записывается в виде:

$$Q_\mu(\varepsilon, A) = \sum_{h=0}^{D} |\omega_h| \sum_{t_1=h}^{D} \sum_{t_2=0}^{D} \frac{1}{C_L^\ell} \sum_{X \in N(t_1,t_2)} \frac{1}{|A(X)|} \left[\delta(a_h, X) \geqslant \varepsilon\right].$$

Здесь индекс h обозначает номер класса идентичных алгоритмов (таким образом, что все алгоритмы класса ω_h имеют $m + h$ ошибок); $|\omega_0| = 1$, и $|\omega_h| = 2$ при $h \geqslant 1$. Для определенности будем брать представителя a_h класса ω_h из левой ветви цепочки.

Индексы t_1 и t_2 параметризуют состав множества $A(X)$. Для произвольного разбиения $X \in [\mathbb{X}]^\ell$ определим t_1 как максимальное число, для которого все объекты $x_1, x_2, \ldots, x_t$ находятся в контроле, а x_{t_1+1} (при его наличии) — в обучении. Индекс t_2 определяется аналогично для объектов правой ветви. Множество $N(t_1, t_2) \subset [\mathbb{X}]^\ell$ есть множество всех разбиений выборки с параметрами t_1 и t_2.

Из определения t_1 и t_2 следует, что $|A(X)| = \frac{1}{1+t_1+t_2}$. Индексы t_1 и t_2 при суммировании пробегают разные множества значений, поскольку рассматрива-

ются только разбиения, при которых выбранный из левой ветви представитель a_h лежит в $A(X)$.

Обозначим $F = [t_1 \neq D] + [t_2 \neq D]$, $L' = L - t_1 - t_2 - F$, $\ell' = \ell - F$. Параметр F позволяет учитывать вклад последних алгоритмов a_D и a'_D цепочки.

Вычислим мощность подмножества тех разбиений из $N(t_1, t_2)$, на которых алгоритм a_h оказывается переобученным. Пусть $s_0(\varepsilon)$ — максимальное число ошибок на обучении, при котором наблюдается переобучение. По определению уклонения частот находим $s_0(\varepsilon) = \lfloor \frac{\ell}{L}(m + h - \varepsilon k) \rfloor$. Нам необходимо из L' объектов выбрать ℓ' для обучения таким образом, что бы из m свободных ошибок алгоритма a_h в обучении оказалось не более $s_0(\varepsilon)$ ошибок. Это число способов дается выражением $\sum\limits_{s=0}^{s_0(\varepsilon)} C_m^s C_{L'-m}^{\ell'-s}$.

Собирая все результаты, приходим к окончательной формуле:

$$Q_\mu(\varepsilon, A) = \sum_{h=0}^{D} |\omega_h| \sum_{t_1=h}^{D} \sum_{t_2=0}^{D} \frac{1}{1 + t_1 + t_2} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell', m}(s_0(\varepsilon)). \quad \blacksquare$$

3.4 Связка из монотонных цепочек

Связкой из p монотонных цепочек называется множество алгоритмов, полученное объединением p монотонных цепочек равной длины, с общим первым алгоритмом. Как и в случае унимодальной цепочки, предполагается, что множества объектов, на которых ошибаются алгоритмы ветвей, не пересекаются.

Группа симметрии связки из p монотонных цепочек является симметрической группой S_p, действующей на ветви связки всевозможными перестановками. Таким образом, классы идентичных алгоритмов — это подмножества алгоритмов с одинаковым числом ошибок на полной выборке, называемые *слоями* [4].

В следующей теореме будет дана явная формула вероятности переобучения для связки из p монотонных цепочек. Введём *комбинаторный коэффициент* $R_{D,p}^h(S, F)$, который зависит от параметров S и F, от числа монотонных цепочек p и от их длины D, а также от h — минимального значения параметра S. Коэффициент $R_{D,p}^h(S, F)$ равен числу способов представить число S в виде суммы p неотрицательных слагаемых, $S = t_1 + \ldots + t_p$, каждое из которых не превосходит D. При этом ровно F слагаемых не должно равняться D, а на первое слагаемое накладывается дополнительное ограничение $t_1 \geqslant h$.

Теорема 3.4. *Пусть в связке из p монотонных цепочек лучший алгоритм допускает m ошибок на полной выборке, длина каждой ветви без учета лучшего*

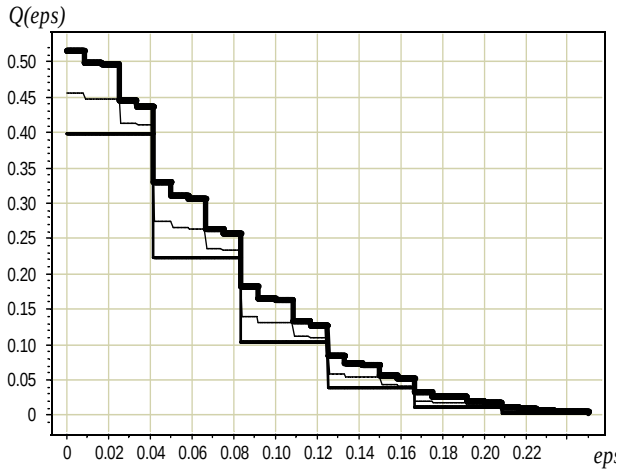


Рис. 1: Зависимость $Q_\mu(\varepsilon, A)$ от ε для монотонной цепочки при $L = 100$, $\ell = 60$, $D = 40$, $m = 20$.

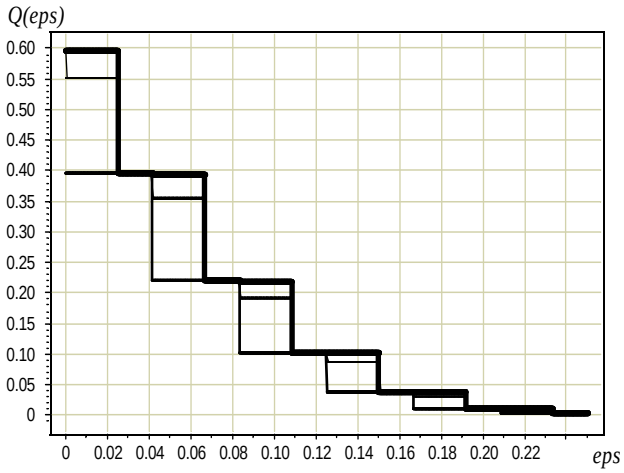


Рис. 2: Зависимость $Q_\mu(\varepsilon, A)$ от ε для единичной окрестности при $L = 100$, $\ell = 60$, $p = 10$, $m = 20$.

алгоритма равна D. Тогда при обучении рандомизированным методом вероятность переобучения может быть записана в виде:

$$Q_\mu(\varepsilon, A) = \sum_{h=0}^{D} \sum_{S=h}^{pD} \sum_{F=0}^{p} \frac{|\omega_h| R^h_{D,p}(S, F)}{1 + S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell', m}_{L'}(s(\varepsilon)), \tag{3.3}$$

где $L' = L - S - F$, $\ell' = \ell - F$, $s(\varepsilon) = \left\lfloor \frac{\ell}{L}(m + h - \varepsilon k) \right\rfloor$; $|\omega_h| = 1$ при $h = 0$ и $|\omega_h| = p$ при $h \geqslant 1$; $H^{\ell', m}_{L'}(s)$ — функция гипергеометрического распределения [4].

□ **Доказательство.** Естественным образом обобщая рассуждения, приведенные для унимодальной цепочки, получаем формулу

$$Q_\mu(\varepsilon, A) = \sum_{h=0}^{D} |\omega_h| \sum_{t_1=h}^{D} \sum_{t_2=0}^{D} \dots \sum_{t_p=0}^{D} \frac{1}{1 + t_1 + t_2 + \dots + t_p} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell', m}_{L'}(s(\varepsilon)),$$

где $L' = L - \sum\limits_{i=1}^{p} t_i - \sum\limits_{i=1}^{p} [t_i \neq D]$, $\ell' = \ell - \sum\limits_{i=1}^{p} [t_i \neq D]$, $s_0(\varepsilon) = \lfloor \frac{\ell}{L}(m + h - \varepsilon k) \rfloor$.

Упростим запись, введя дополнительные обозначения $S = \sum\limits_{i=1}^{p} t_i$, $F = \sum\limits_{i=1}^{p} [t_i \neq D]$. Параметр S определяет мощность множества $A(X)$.

$$Q_\mu(\varepsilon, A) = \sum_{h=0}^{D} |\omega_h| \sum_{t_1=h}^{D} \sum_{t_2=0}^{D} \dots \sum_{t_p=0}^{D} \frac{1}{1 + S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell', m}_{L'}(s_0(\varepsilon)),$$

где $L' = L - S - F$, $\ell' = \ell - F$, $s_0(\varepsilon) = \lfloor \frac{\ell}{L}(m + h - \varepsilon k) \rfloor$.

Теперь от суммирования по параметрам t_i можно перейти к суммированию

по множеству возможных значений S и F:

$$Q_\mu(\varepsilon, A) = \sum_{h=0}^{D} |\omega_h| \sum_{S=h}^{pD} \sum_{F=0}^{p} \frac{R^h_{D,p}(S,F)}{1+S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s_0(\varepsilon)),$$

где $R^h_{D,p}(S,F)$ — определенный выше комбинаторный коэффициент. ■

Связка из $2p$ монотонных цепочек является моделью p-параметрического семейства алгоритмов, в котором разрешено изменять любой из p параметров при фиксированных остальных, а одновременное изменение нескольких параметров не допускается. Данное семейство можно также рассматривать как обобщение трёх частных случаев, рассмотренных в [3]: монотонной цепочки ($p = 1$), унимодальной цепочки ($p = 2$) и единичной окрестности лучшего алгоритма ($D = 1$).

Формула для вероятности переобучения унимодальной цепочки уже была получена в теореме 3.3. Для получения явных формул для двух оставшихся семейств достаточно найти явное выражение для комбинаторного коэффициента $R^h_{D,p}(S,F)$.

Следствие 3.4.1. *Для монотонной цепочки длины $D+1$ вероятность переобучения равна*

$$Q_\mu(\varepsilon, A) = \frac{1}{C^{\ell}_{L}} \sum_{h=0}^{D} \sum_{S=h}^{D} \frac{1}{1+S} H^{\ell',m}_{L'}(s(\varepsilon)), \tag{3.4}$$

где $L' = L - S - [S \neq D]$, $\ell' = \ell - [S \neq D]$.

Следствие 3.4.2. *Для единичной окрестности из $p+1$ алгоритма вероятность переобучения равна*

$$Q_\mu(\varepsilon, A) = \frac{1}{C^{\ell}_{L}} \sum_{h=0}^{1} \sum_{S=h}^{p} \frac{|\omega_h| C^{S-h}_{p-h}}{1+S} H^{\ell',m}_{L'}(s(\varepsilon)), \tag{3.5}$$

где $L' = L - p$, $\ell' = \ell + S - p$.

На рис. 1 и рис. 2 представлены результаты численных экспериментов, в которых сравнивались вероятности переобучения для различных вариантов минимизации эмпирического риска. Из четырех кривых на каждом графике верхняя (жирная) соответствует пессимистической минимизации эмпирического риска [3, 4], нижняя — оптимистической. Две почти сливающиеся кривые между ними соответствуют рандомизированной минимизации эмпирического риска. Одна из них вычислена по доказанным формулам, вторая построена методом Монте-Карло по 10^5 случайных разбиений, при равновероятном выборе лучшего алгоритма в случаях неопределенности. Различия этих двух кривых находятся в пределах погрешности метода Монте-Карло.

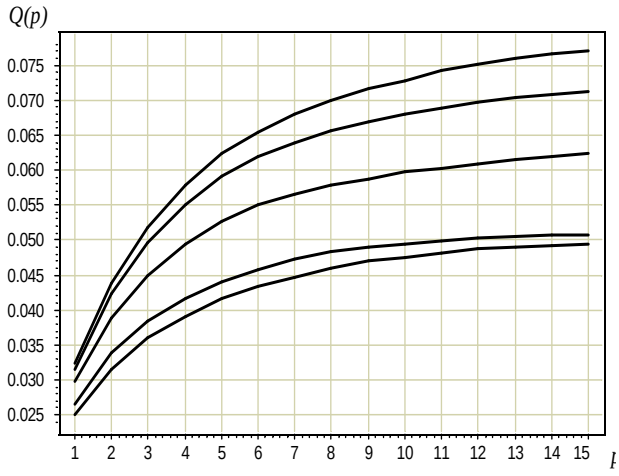


Рис. 3: Зависимость $Q_\mu(\varepsilon, A)$ от p для связки из монотонных цепочек при $L = 300$, $\ell = 150$, $m = 15$, $D = 1, 2, 3, 5, 10$, $\varepsilon = 0.05$.

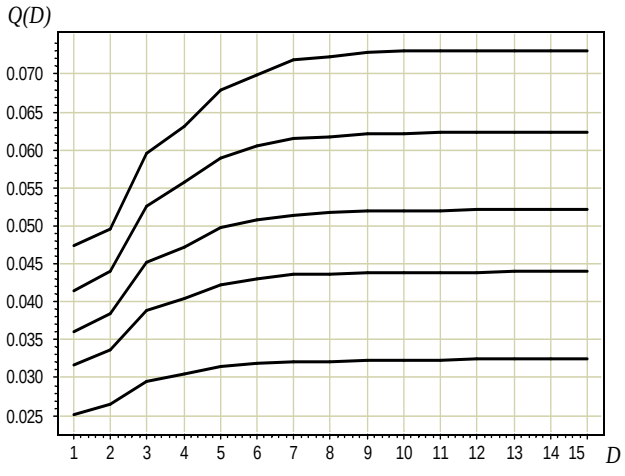


Рис. 4: Зависимость $Q_\mu(\varepsilon, A)$ от D для связки из $p = 1, 2, 3, 5, 10$ монотонных цепочек при $L = 300$, $\ell = 150$, $m = 15$, $\varepsilon = 0.05$.

На рис. 3 и рис. 4 представлены зависимости вероятности переобучения от числа p ветвей в связке и от их длины D. Графики построены для рандомизированного метода минимизации эмпирического риска. Рис. 4 показывает, что при увеличении длин цепочек D вероятность переобучения практически перестаёт расти уже при $D = 7$. Это связано с *эффектом расслоения* — лишь алгоритмы из нижних слоёв имеют существенно отличную от нуля вероятность быть выбранными методом минимизации эмпирического риска. Добавление «слишком плохих» алгоритмов не увеличивает вероятность переобучения. Рис. 3 показывает, что при увеличении числа p цепочек в связке вероятность переобучения продолжает расти. Однако скорость роста сублинейна по p, благодаря *эффекту связности* — все алгоритмы находятся на хэмминговом расстоянии не более D от лучшего алгоритма.

3.5 Монотонная сетка

Введём целочисленный вектор индексов $\boldsymbol{d} = (d_1, \ldots, d_h) \in \mathbb{Z}^h$. Обозначим $\|\boldsymbol{d}\| = \max\limits_{j=1,\ldots,h} |d_j|$, $|\boldsymbol{d}| = |d_1| + \cdots + |d_h|$. На множестве векторов индексов введём покомпонентное отношение сравнения: $\boldsymbol{d} < \boldsymbol{d}'$, если $d_j \leqslant d'_j$, $j = 1, \ldots, h$, и хотя бы одно из неравенств строгое.

Определение 3.3. *Множество алгоритмов* $A_M = \{a_{\boldsymbol{d}}\}$, *где* $\boldsymbol{d} \geqslant 0$ *и* $\|\boldsymbol{d}\| \leqslant D$

называется монотонной h-мерной сеткой алгоритмов, если существует $h \in \mathbb{N}$ и упорядоченные наборы объектов $X_j = \{x_j^1, x_j^2, \ldots, x_j^D\} \subset \mathbb{X}$, для всех $j = 1, \ldots, h$, а так же множества $U_1 \subset \mathbb{X}$ и $U_0 \subset \mathbb{X}$, такие что выполнены условия:

1) *Набор* $\left\{U_0, U_1, \{X_j\}_{j=1}^h\right\}$ *является разбиением множества* $\mathbb{X}$ *на непересекающиеся множества;*
2) $a_d(x_j^i) = [i \leqslant d_j]$, *где* $x_j^i \in X_j$;
3) $a_d(x_0) = 0$ *при всех* $x_0 \in U_0$;
4) $a_d(x_1) = 1$ *при всех* $x_1 \in U_1$.

Обозначим $|U_1| = m$. Из определения следует, что $n(a_{\boldsymbol{d}}, \mathbb{X}) = m + |d|$. Алгоритм $a_{\mathbf{0}}$ является *лучшим в сетке*. Множество алгоритмов с равным числом ошибок $t + m = n(a_{\boldsymbol{d}}, \mathbb{X})$ называются t-*слоем* сетки.

Пример 3.1. Монотонная двумерная сетка при $m = 0$ и $L = 4$:

	$a_{0,0}$	$a_{1,0}$	$a_{2,0}$	$a_{0,1}$	$a_{1,1}$	$a_{2,1}$	$a_{0,2}$	$a_{1,2}$	$a_{2,2}$
x_1	0	**1**	**1**	0	**1**	**1**	0	**1**	**1**
x_2	0	0	**1**	0	0	**1**	0	0	**1**
x_3	0	0	0	**1**	**1**	**1**	**1**	**1**	**1**
x_4	0	0	0	0	0	0	**1**	**1**	**1**

Число алгоритмов в h-мерной монотонной сетке с ветвями длины D равно $(D+1)^h$. *Укороченной* h-мерной монотонной сеткой $\tilde{A}_M \subset A_M$ назовем первые D слоев из A_M. Таким образом $\tilde{A}_M = \{a_{\boldsymbol{d}} \in A_M, |\boldsymbol{d}| \leqslant D\}$. Число алгоритмов в $\tilde{A}_M$ равно C_{D+h}^h.

Впервые укороченные монотонны сетки произвольной размерности были изучены П. Ботовым в [5]. Там же были получены формулы для вероятности переобучения *пессимистического* метода минимизации эмпирического риска.

Численные эксперименты показывают, что при разумных сочетаниях параметров вероятности переобучения для укороченной $\tilde{A}_M$ и простой A_M монотонных сеток различаются крайне мало. Поэтому в дальнейшем мы ограничимся исследованием не-укороченных монотонных сеток. Для этого класса семейств алгоритмов будут получены явные формулы вероятности переобучения рандомизированного метода минимизации эмпирического риска.

Лемма 3.1. *Группа симметрии монотонной сетки размерности h содержит в качестве подгруппы группу S_h всевозможных перестановок множеств $X_1, \ldots, X_h$.*

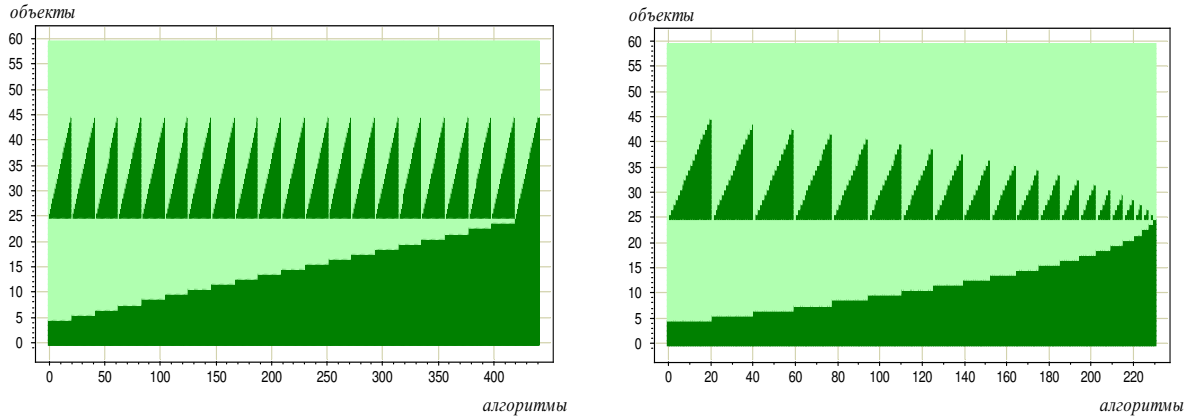


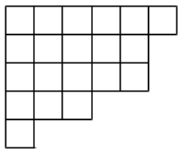
Рис. 5: Матрица ошибок монотонной сетки (слева) и укороченной монотонной сетки (справа) при $D = 20$, $h = 2$, $m = 5$, $L = 60$.

□ **Доказательство.**

Рассмотрим алгоритм $a_{\boldsymbol{d}} \in A_M$ и произвольную $\pi \in S_h$. По данному выше определению действия π на $\mathbb{X}$ получаем, что $\pi a_{\boldsymbol{d}} = a_{\pi \boldsymbol{d}}$, где действие π на вектор $\boldsymbol{d}$ определяется соответствующей перестановкой его координат. Множество $\{0, \ldots, D\}^h$ сохраняется при применении к нему произвольной перестановки координат $\pi \in S_h$. Поэтому $\forall \boldsymbol{d} \in \{0, \ldots, D\}^h$ выполнено $\pi \boldsymbol{d} \in \{0, \ldots, D\}^h$. А следовательно $a_{\pi \boldsymbol{d}} \in A_M$. ■

Определение 3.4. *Диаграммой Юнга порядка p будем называть не-возрастающую последовательность неотрицательных чисел $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n \geqslant 0$, такую что $\sum_{j=1}^{n} \lambda_j = p$.*

Диаграммы Юнга находятся во взаимно-однозначном соответствии с гистограммами следующего вида:



Множество диаграмм Юнга порядка p будем обозначать через Y_p. Множество диаграмм Юнга из h столбцов, в которых $\lambda_1 \leqslant D$, обозначим через $Y_p^{h,D}$. Обозначим $Y_*^{h,D} = \bigcup\limits_{p=0}^{D*h} Y_p^{h,D}$

Лемма 3.2. *Множество орбит монотонной сетки $A_M = \{a_{\boldsymbol{d}}\}$ размерности h, $\|d\| \leqslant D$ под действием S_h индексировано всевозможными диаграммами Юнга из $Y_*^{h,D}$. Число алгоритмов в орбите ω_λ, где $\lambda = (\lambda_1, \ldots, \lambda_h)$ равно числу различных слов длины h, состоящих из символов $\lambda_1, \ldots, \lambda_h$: $|\omega_\lambda| = |S_h \lambda|$.*

□ **Доказательство.** Напомним, что вместо действия S_h на $A_M = \{a_{\boldsymbol{d}}\}$ можно рассматривать действие S_h на вектор индексов $\boldsymbol{d}$, заданное перестановками координат.

Рассмотрим орбиту произвольного алгоритма $a_{\boldsymbol{d}}$. Возьмем перестановку $\pi \in S_h$, упорядочивающую координаты $\boldsymbol{d}$ в порядке не-возрастания, и положим $\lambda = \pi\boldsymbol{d}$. Получаем, что $\lambda \in Y_*^{h,D}$ — диаграмма Юнга. При этом различным диаграммам Юнга λ_1 и λ_2 будут соответствовать различные орбиты действия группы S_h на $\{a_{\boldsymbol{d}}\}$.

Взаимно-однозначное соответствие между словами длины h из символов $\lambda_1, \ldots, \lambda_h$ и количеством элементами орбиты $|\omega_\lambda|$ очевидно. ■

Теорема 3.5. *Вероятность переобучения рандомизированного метода минимизации эмпирического риска, примененного к монотонной сетке $A_M = \{a_{\boldsymbol{d}}\}$ размерности h, $\|d\| \leqslant D$, дается выражением:*

$$Q_\mu(\varepsilon, A_M) = \sum_{\lambda \in Y_*^{h,D}} \sum_{\substack{\boldsymbol{t} \geqslant \lambda, \\ \|\boldsymbol{t}\| \leqslant D}} \frac{|S_h\lambda|}{T(\boldsymbol{t})} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell',m}(s_0),$$

где $T(\boldsymbol{t}) = \prod_j (t_j + 1)$, $\ell' = \ell - \sum_{j=1}^h [t_j \neq D]$, $k' = k - |\boldsymbol{t}|$, $L' = \ell' + k'$, $s_0 = \frac{\ell}{L}[m + |\lambda| - \varepsilon k]$, $H_{L'}^{\ell',m}(s)$ — функция гипергеометрического распределения [4].

□ **Доказательство.**

Согласно теореме 2.2 вероятность переобучения записывается в виде:

$$Q_\mu(\varepsilon, A_M) = \frac{1}{C_L^\ell} \sum_{\lambda \in Y_*^{h,D}} |S_h\lambda| \sum_{X \in [\mathbb{X}]^\ell} \frac{[a_\lambda \in A_M(X)]}{|A_M(X)|} [\delta(a_\lambda, X) \geqslant \varepsilon].$$

Шаг 1. Зафиксируем $X \in [\mathbb{X}]^\ell$. Обозначим через t_j максимальный индекс из $\{1, \ldots, h\}$, при котором все объекты $\{x_j^1, \ldots, x_j^{t_j}\}$ содержатся в $\bar{X}$, а $x_j^{t_j+1}$, при его наличии, лежит в X. Положим $\boldsymbol{t} = \{t_j\}_{j=1}^h$. Тогда условие $a_\lambda \in A_M(X)$ перепишется как $\boldsymbol{t} \geqslant \lambda$.

Действительно, заметим что для всех $a \in A_M$ и $X \in [\mathbb{X}]^\ell$ выполнено $n(a, X) \geqslant n(a_0, X)$. Следовательно, алгоритм a_λ может быть выбран, только если объекты x_j^i при всех $j = 1, \ldots, h$ и $i \leqslant \lambda_j$ лежат в контроле. В терминах $\boldsymbol{t}$ это записывается как $\boldsymbol{t} \geqslant \lambda$.

Обозначим множество разбиений на обучение и контроль с фиксированным значением параметра $\boldsymbol{t}$ через $[\mathbb{X}]_{\boldsymbol{t}}^\ell$. Тогда

$$Q_\mu(\varepsilon, A_M) = \frac{1}{C_L^\ell} \sum_{\lambda \in Y_*^{h,D}} |S_h\lambda| \sum_{\substack{\boldsymbol{t} \geqslant \lambda, \\ \|\boldsymbol{t}\| \leqslant D}} \sum_{X \in [\mathbb{X}]_{\boldsymbol{t}}^\ell} \frac{1}{|A_M(X)|} [\delta(a_\lambda, X) \geqslant \varepsilon].$$

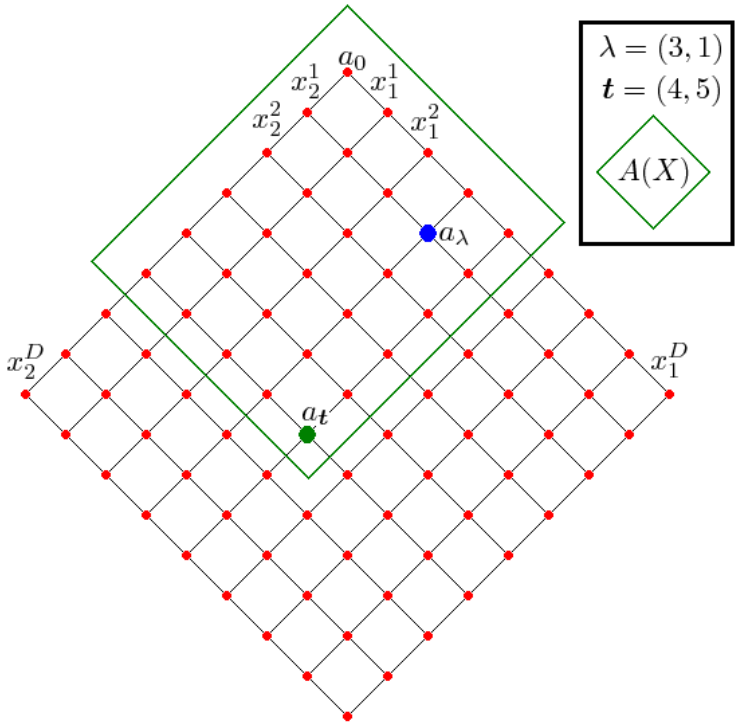


Рис. 6: Строение множества $A_M(X)$ для двумерной монотонной сетки.

Шаг 2. Пусть $X \in [\mathbb{X}]^\ell_{\boldsymbol{t}}$. Заметим, что алгоритм $a_{\boldsymbol{d}} \in A_M(X)$ тогда и только тогда, когда $\boldsymbol{d} \leqslant \boldsymbol{t}$. Следовательно $|A_M(X)| = (t_1+1)(t_2+1)\dots(t_h+1)$. Обозначим $T(\boldsymbol{t}) = \prod_j (t_j+1)$.

Шаг 3. Обозначим через $s = |U_1 \cap X|$ число объектов из U_1, лежащих в обучении. Тогда $\delta(a_\lambda, X) = \frac{m-s+|\lambda|}{k} - \frac{s}{\ell}$, и условие $\delta(a_\lambda, X) \geqslant \varepsilon$ запишется в виде $s \leqslant \frac{\ell}{L}[m + |\lambda| - \varepsilon k] \equiv s_0$. Множество всех разбиений из $[\mathbb{X}]^\ell_{\boldsymbol{t}}$ с фиксированным параметром s обозначим через $[\mathbb{X}]^\ell_{\boldsymbol{t},s}$. Тогда

$$Q_\mu(\varepsilon, A_M) = \frac{1}{C_L^\ell} \sum_{\lambda \in Y_*^{h,D}} |S_h \lambda| \sum_{\substack{\boldsymbol{t} \geqslant \lambda, \\ \|\boldsymbol{t}\| \leqslant D}} \frac{1}{T(\boldsymbol{t})} \sum_{s=0}^{s_0} |[\mathbb{X}]^\ell_{\boldsymbol{t},s}|.$$

Шаг 4. Вычислим мощность множества $[\mathbb{X}]^\ell_{\boldsymbol{t},s}$.

Введем обозначения $\ell' = \ell - \sum_{j=1}^h [t_j \neq D]$, $k' = k - |\boldsymbol{t}|$, $L' = \ell' + k'$. Тогда простое комбинаторное вычисление показывает, что $|[\mathbb{X}]^\ell_{\boldsymbol{t},s}| = C_m^s C_{L'-m}^{k'-s}$. Следовательно,

$$Q_\mu(\varepsilon, A_M) = \frac{1}{C_L^\ell} \sum_{\lambda \in Y_*^{h,D}} |S_h \lambda| \sum_{\substack{\boldsymbol{t} \geqslant \lambda, \\ \|\boldsymbol{t}\| \leqslant D}} \frac{1}{T(\boldsymbol{t})} \sum_{s=0}^{s_0} C_m^s C_{L'-m}^{k'-s}.$$

Напомним, что $H_{L'}^{\ell',m}(z) = \frac{1}{C_{L'}^{\ell'}} \sum\limits_{s=0}^{\lfloor z \rfloor} C_m^s C_{L'-m}^{\ell'-s}$ — функция гипергеометрического распределения [4]. Тогда

$$Q_\mu(\varepsilon, A_M) = \sum_{\lambda \in Y_*^{h,D}} \sum_{\substack{\boldsymbol{t} \geqslant \lambda, \\ \|\boldsymbol{t}\| \leqslant D}} \frac{|S_h \lambda|}{T(\boldsymbol{t})} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell',m}(s_0). \blacksquare$$

3.6 Унимодальная сетка

Определение 3.5. *Множество алгоритмов $A_U = \{a_{\boldsymbol{d}}\}$, где $\|\boldsymbol{d}\| \leqslant D$ называется унимодальной h-мерной сеткой алгоритмов, если существует $h \in \mathbb{N}$ и упорядоченные наборы объектов $X_j = \{x_j^1, x_j^2, \ldots, x_j^D\} \subset \mathbb{X}$, $Y_j = \{y_j^1, y_j^2, \ldots, y_j^D\} \subset \mathbb{X}$, для всех $j = 1, \ldots, h$, а так же множества $U_1 \subset \mathbb{X}$ и $U_0 \subset \mathbb{X}$, такие что выполнены условия:*

1) *Набор $\left\{U_0, U_1, \{X_j\}_{j=1}^h, \{Y_j\}_{j=1}^h\right\}$ является разбиением множества $\mathbb{X}$ на непересекающиеся множества;*
2) *$a_d(x_j^i) = [d_j > 0]\,[i \leqslant |d_j|]$, где $x_j^i \in X_j$;*
3) *$a_d(y_j^i) = [d_j < 0]\,[i \leqslant |d_j|]$, где $y_j^i \in Y_j$;*
4) *$a_d(x_0) = 0$ при всех $x_0 \in U_0$;*
5) *$a_d(x_1) = 1$ при всех $x_1 \in U_1$.*

Заметим, что данное определение отличается от определения монотонной сетки отсутствием ограничения $\boldsymbol{d} \geqslant 0$. Число алгоритмов в h-мерной унимодальной сетке с ветвями длины D составляет $(2D+1)^h$. *Укороченной* h-мерной унимодальной сеткой $\tilde{A}_U$ назовем множество первых D слоев из A_U: $\tilde{A}_U = \{a_{\boldsymbol{d}} \in A_U \colon n(a_{\boldsymbol{d}}, \mathbb{X}) \leqslant m + D\}$.

Формула для вероятности переобучения *пессимистического* метода минимизации эмпирического риска на укороченных унимодальных сетках так же была получена в [5]. Ниже рассматриваются не-укороченные унимодальные сетки и случай *рандомизированного* метода минимизации эмпирического риска.

Лемма 3.3. *Группа симметрии унимодальной сетки размерности h содержит в качестве подгруппы группу $\mathrm{Sym}(A_U) = (S_2)^h \times S_h$. Группа S_h действует на множестве пар $(X_j, Y_j)_{j=1}^h$ всеми возможными перестановками; j-тая группа S_2 переставляет объекты множества X_j и Y_j местами, сохраняя относительный порядок объектов.*

□ **Доказательство.**

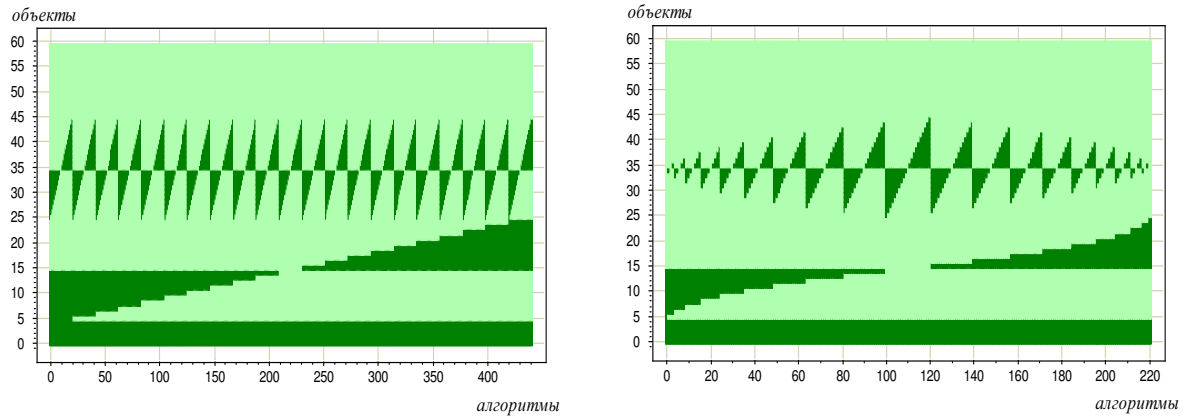


Рис. 7: Матрица ошибок унимодальной сетки (слева) и укороченной унимодальной сетки (справа) при $D = 10$, $h = 2$, $m = 5$, $L = 60$.

Рассмотрим алгоритм $a_{\boldsymbol{d}} \in A_U$ и произвольную $\pi = (z_1, \ldots, z_h) \times \pi_0 \in \mathrm{Sym}(A_U)$, где $z_j \in S_2$, $\pi_0 \in S_h$. По данному выше определению действия π на $\mathbb{X}$ получаем, что $\pi a_{\boldsymbol{d}} = a_{\pi \boldsymbol{d}}$, где действие π на вектор $\boldsymbol{d}$ определяется перестановкой его координат с помощью π_0 и инверсией знаков для всех j, таких что $z_j \neq id$ — транспозиция. Множество $\{-D, \ldots, D\}^h$ сохраняется при применении к нему произвольной перестановки координат $\pi \in \mathrm{Sym}(A_U)$. Поэтому $\forall \boldsymbol{d} \in \{-D, \ldots, D\}^h$ выполнено $\pi \boldsymbol{d} \in \{-D, \ldots, D\}^h$. А следовательно $a_{\pi \boldsymbol{d}} \in A_U$. ■

Лемма 3.4. *Множество орбит унимодальной сетки $A_U = \{a_{\boldsymbol{d}}\}$ размерности h, $\|d\| \leqslant D$ под действием $\mathrm{Sym}(A_U)$ индексировано всевозможными диаграммами Юнга из $Y_*^{h,D}$. Пусть $\lambda = (\lambda_1, \ldots, \lambda_h) \in Y_*^{h,D}$. Обозначим через $|S_h \lambda|$ число различных слов длины h, состоящих из символов $\lambda_1, \ldots, \lambda_h$ Пусть $|\lambda > 0|$ — число строго положительных компонент вектора λ.*

Тогда число алгоритмов в орбите ω_λ равно $|S_h \lambda| * 2^{|\lambda > 0|}$.

□ **Доказательство** полностью повторяет рассуждения леммы 3.2. Множитель $2^{|\lambda > 0|}$ соответствует возможности сменить знак у всех не-нулевых компонент вектора $\boldsymbol{d}$. ■

Теорема 3.6. *Вероятность переобучения рандомизированного метода минимизации эмпирического риска, примененного к унимодальной сетке $A_U = \{a_{\boldsymbol{d}}\}$ размерности h, $\|d\| \leqslant D$, дается выражением:*

$$Q_\mu(\varepsilon, A_U) = \sum_{\lambda \in Y_*^{h,D}} \sum_{\substack{\boldsymbol{t} \geqslant \lambda, \\ \|\boldsymbol{t}\| \leqslant D}} \sum_{\substack{\boldsymbol{t'} \geqslant 0, \\ \|\boldsymbol{t'}\| \leqslant D}} \frac{|S_h \lambda| * 2^{|\lambda > 0|}}{T(\boldsymbol{t}, \boldsymbol{t'})} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell', m}(s_0),$$

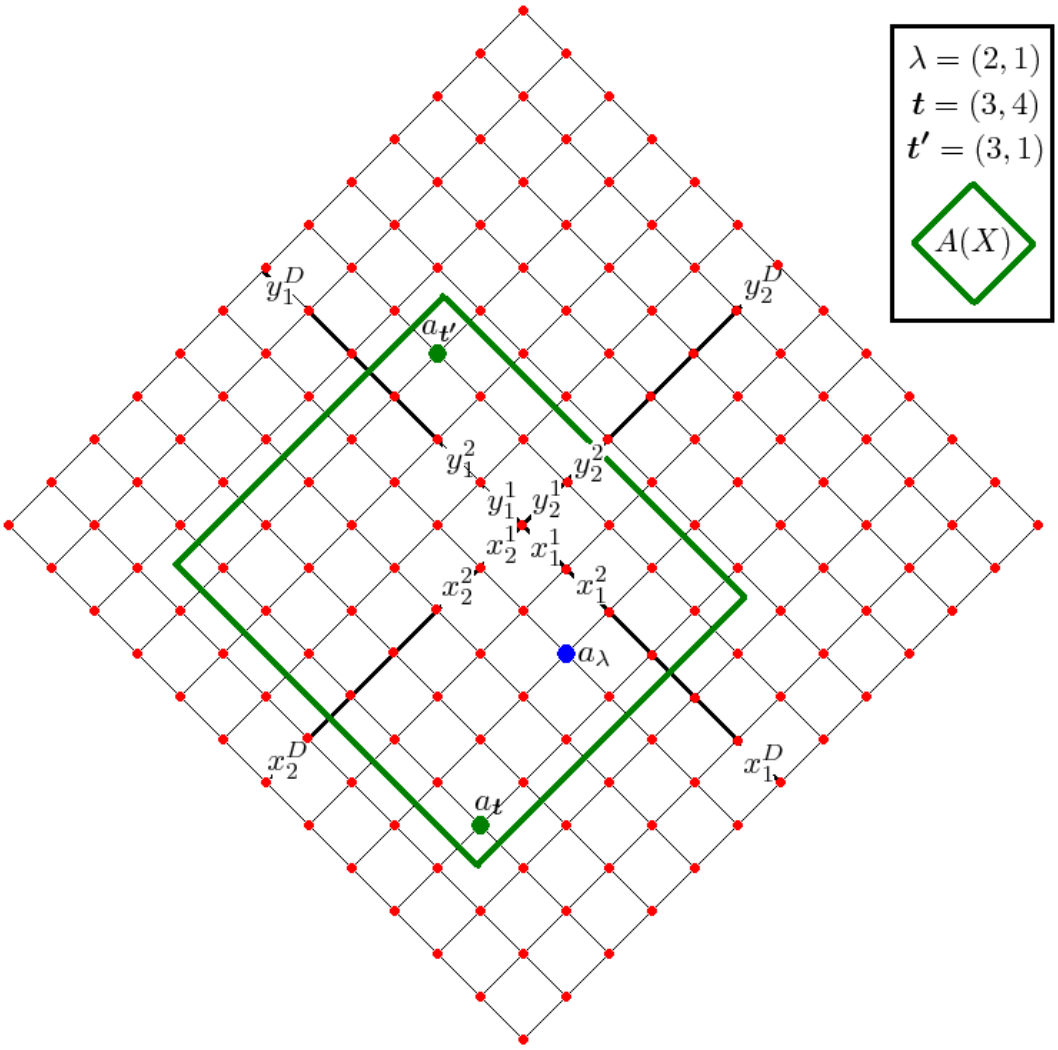


Рис. 8: Строение множества $A_U(X)$ для двумерной унимодальной сетки.

где $T(\boldsymbol{t},\boldsymbol{t}') = \prod_j (t_j + t'_j + 1)$, $\ell' = \ell - \sum\limits_{j=1}^{h} \Big([t_j \neq D] + [t'_j \neq D] \Big)$, $k' = k - |\boldsymbol{t}| - |\boldsymbol{t}'|$, $L' = \ell' + k'$, $s_0 = \frac{\ell}{L}[m + |\lambda| - \varepsilon k]$, $H_{L'}^{\ell',m}(s)$ — *функция гипергеометрического распределения [4].*

□ **Доказательство.**

Шаг 1. Выберем в качестве представителя a_λ орбиты ω_λ алгоритм, не допускающий ошибок на множестве $Y = \bigcup_{j=1}^{h} Y_j$. Этого можно добиться, взяв произвольный $a_{\boldsymbol{d}} \in \omega_\lambda$ и поменяв знаки у всех $d_j < 0$ с помощью транспозиции z_j.

Введя обозначения $\boldsymbol{t}$ и $[\mathbb{X}]^\ell_{\boldsymbol{t}}$ так же, как и на первом шаге вывода формулы для монотонной сетки, получим

$$Q_\mu(\varepsilon, A_U) = \frac{1}{C^\ell_L} \sum_{\lambda \in Y^{h,D}_*} |S_h\lambda| * 2^{|\lambda>0|} \sum_{\substack{\boldsymbol{t} \geqslant \lambda, \\ \|\boldsymbol{t}\| \leqslant D}} \sum_{X \in [\mathbb{X}]^\ell_{\boldsymbol{t}}} \frac{1}{|A_U(X)|} \left[\delta(a_\lambda, X) \geqslant \varepsilon\right].$$

Шаг 2. Обозначим через t'_j максимальный индекс из $\{1, \ldots, h\}$, при котором все объекты $\{y^1_j, \ldots, y^{t'_j}_j\}$ содержатся в $\bar{X}$, а $y^{t'_j+1}_j$, при его наличии, лежит в X. Положим $\boldsymbol{t'} = \{t'_j\}_{j=1}^h$. Заметим, что вектор $\boldsymbol{t'}$ играет для набора $\{Y_j\}$ ту же роль, что $\boldsymbol{t}$ для $\{X_j\}$. Обозначим через $[\mathbb{X}]^\ell_{\boldsymbol{t},\boldsymbol{t'}}$ множество разбиений с фиксированными параметрами $\boldsymbol{t}$ и $\boldsymbol{t'}$.

Пусть $X \in [\mathbb{X}]^\ell_{\boldsymbol{t},\boldsymbol{t'}}$. Заметим, что алгоритм $[a_{\boldsymbol{d}} \in A_U(X)] = [-\boldsymbol{t'} \leqslant \boldsymbol{d} \leqslant \boldsymbol{t}]$. Следовательно $|A_U(X)| = (t_1 + t'_1 + 1)(t_2 + t'_2 + 1)\ldots(t_h + t'_h + 1)$. Обозначим $T(\boldsymbol{t}, \boldsymbol{t'}) = \prod_j (t_j + t'_j + 1)$.

Шаг 3. Обозначим через $s = |U_1 \cap X|$ число объектов из U_1, лежащих в обучении. Пусть $s_0 \equiv \frac{\ell}{L}[m + |\lambda| - \varepsilon k]$. Повторяя рассуждения аналогичного шага доказательства для монотонной сетки получим

$$Q_\mu(\varepsilon, A_U) = \frac{1}{C^\ell_L} \sum_{\lambda \in Y^{h,D}_*} |S_h\lambda| * 2^{|\lambda>0|} \sum_{\substack{\boldsymbol{t} \geqslant \lambda, \\ \|\boldsymbol{t}\| \leqslant D}} \sum_{\substack{\boldsymbol{t'} \geqslant 0, \\ \|\boldsymbol{t'}\| \leqslant D}} \frac{1}{T(\boldsymbol{t}, \boldsymbol{t'})} \sum_{s=0}^{s_0} |[\mathbb{X}]^\ell_{\boldsymbol{t},\boldsymbol{t'},s}|.$$

Шаг 4. Посчитаем мощность множества $[\mathbb{X}]^\ell_{\boldsymbol{t},\boldsymbol{t'},s}$.

Обозначим $\ell' = \ell - \sum_{j=1}^h \Big([t_j \neq D] + [t'_j \neq D]\Big)$, $k' = k - |\boldsymbol{t}| - |\boldsymbol{t'}|$, $L' = \ell' + k'$. Тогда $|[\mathbb{X}]^\ell_{\boldsymbol{t},\boldsymbol{t'},s}| = C^s_m C^{k'-s}_{L'-m}$. Воспользовавшись определением функции гипергеометрического распределения получим:

$$Q_\mu(\varepsilon, A_U) = \sum_{\lambda \in Y^{h,D}_*} \sum_{\substack{\boldsymbol{t} \geqslant \lambda, \\ \|\boldsymbol{t}\| \leqslant D}} \sum_{\substack{\boldsymbol{t'} \geqslant 0, \\ \|\boldsymbol{t'}\| \leqslant D}} \frac{|S_h\lambda| * 2^{|\lambda>0|}}{T(\boldsymbol{t}, \boldsymbol{t'})} \frac{C^{\ell'}_{L'}}{C^\ell_L} H^{\ell',m}_{L'}(s_0). \blacksquare$$

3.7 Опорное подмножество алгоритмов

Множество ошибок алгоритма $a \in A$ обозначим через $E(a) \subset \mathbb{X}$.

Определение 3.6. *Подмножество $B \subset A$ множества алгоритмов A будем называть опорным, если для всех $a \in A$ найдется натуральное число $k \in \mathbb{N}$ и набор $\{b_i\}_{i=1}^k$, такой что для всех $i = 1, \ldots, k$ множество ошибок $E(b_i) \subset E(a)$, и кроме того $\bigcup_{i=1}^k E(b_i) = E(a)$. Опорное множество $B \subset A$ назовем минимальным, если любое его собственное подмножество уже не является опорным.*

Перечислим без доказательства очевидные свойства произвольного опорного множества $B \subset A$.

Утверждение 3.1. *Пусть $X \in [\mathbb{X}]^\ell$, $a \in A$, $\{b_i\}_{i=1}^k \subset B$ — некоторое разложение a по опорному множеству B. Тогда*

- $\min\limits_{a\in A} n(a,X) = \min\limits_{b\in B} n(b,X)$;
- $B(X) = A(X) \cap B$;
- *Если $a \in A(X)$ то все b_i тоже лежат в $A(X)$;*
- *Пусть $a \in A(X)$. Тогда для всех b_i выполнено $\delta(b_i, X) \leqslant \delta(a, X)$.*

Легко установить, что связка из h монотонных цепочек является минимальным опорным подмножеством для монотонной сетки размерности h. Связка из $2h$ монотонных цепочек является минимальным опорном множеством унимодальной сетки размерности h. Заметим также, что монотонная сетка размерности $2h$ содержит в качестве опорного множества унимодальную сетку размерности h.

Гипотеза 3.1. *Пусть $B \subset A$ — опорное подмножество. Тогда $Q_\mu(\varepsilon, B) \leqslant Q_\mu(\varepsilon, A)$.*

На рис. 9 и 10 приведены результаты численных экспериментов, подтверждающих данную гипотезу для случая связок из монотонных цепочек, монотонных и унимодальных сеток. Верхняя кривая на всех рисунках соответствует монотонной сетке, средняя кривая — унимодальной сетке, нижняя — связке монотонных цепочек. В эксперименте использовались точные формулы, полученные в предыдущих параграфах.

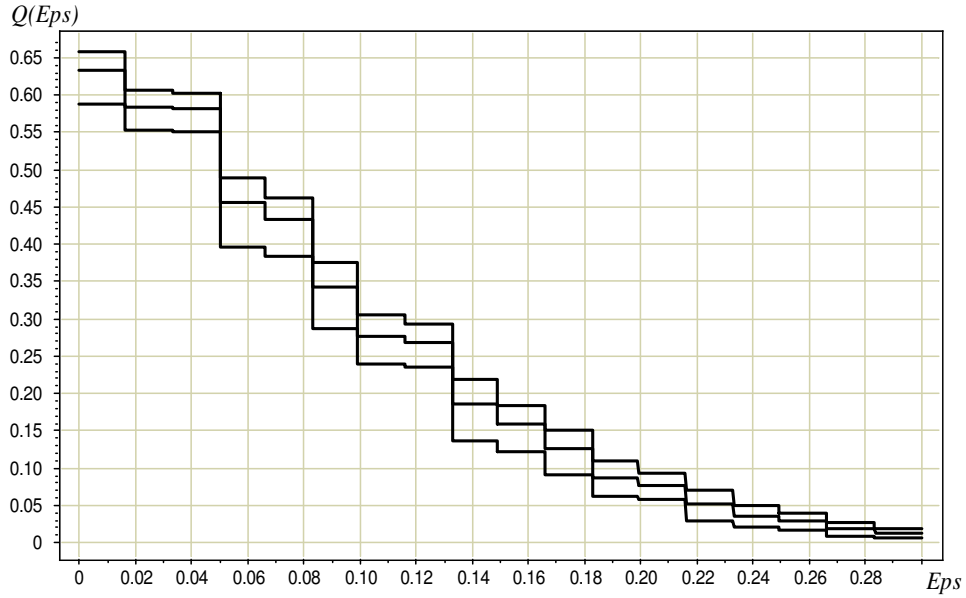


Рис. 9: Сравнение $Q(\varepsilon)$ для связки из $p = 4$ монотонных цепочек, монотонной сетки размерности $h = 4$ и унимодальной сетки размерности $h = 2$. Значения параметров $D = 5$, $m = 5$, $L = 50$, $\ell = 30$.

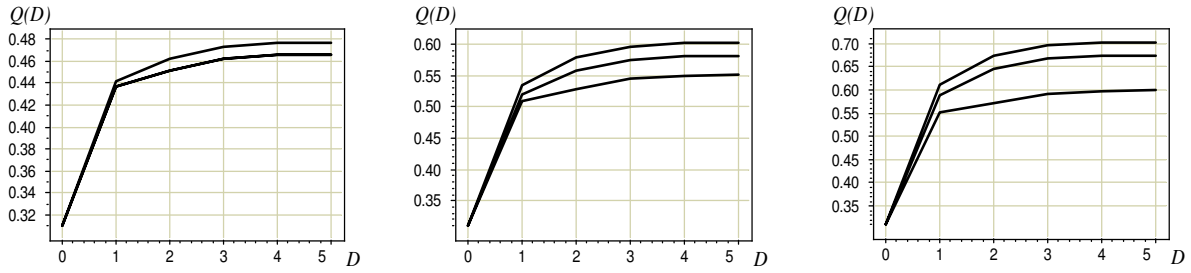


Рис. 10: Сравнение $Q(D)$ для связки из $2h$ монотонных цепочек, $2h$-мерной монотонной сетки и h-мерной унимодальной сетки. Значения параметров $\varepsilon = 0.04$, $m = 5$, $L = 50$, $\ell = 30$, $D = 1, \ldots, 5$; размерности унимодальных сеток $h = 1, 2, 3$ (слева направо).

4 Заключение

Свойство симметрии семейств алгоритмов позволяет получать вычислительно эффективные формулы вероятности переобучения. Для монотонной цепочки, унимодальной цепочки и единичной окрестности такие формулы получены как следствие одной теоремы, в то время как ранее аналогичные оценки доказывались независимо и при неестественном предположении об априорной упорядоченности алгоритмов в семействе [3]. Примененный подход позволил получать оценки для семейств с экспоненциально растущим числом алгоритмов (полный слой алгоритмов, куб алгоритмов).

Получены формулы для вероятности переобучения рандомизированного метода минимизации эмпирического риска на монотонных и унимодальных сетках произвольной размерности. Экспериментально показано, что в широком диапазоне параметров вероятность переобучения и монотонных, и унимодальных сеток оценивается снизу вероятностью переобучения связки из соответствующего количества монотонных цепочек.

Работа поддержана РФФИ (проект № 08-07-00422) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Список литературы

[1] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1974.

[2] *Vapnik V.* Statistical Learning Theory. — New York: Wiley, 1998.

[3] *Воронцов К. В.* Точные оценки вероятности переобучения // Доклады РАН, 2009. — Т. 429, № 1. — С. 15–18.

[4] *Воронцов К. В.* Комбинаторный подход к проблеме переобучения // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009. — С. 18–21.

[5] *Ботов П. В.* Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009. — С. 7–10.

[6] *Фрей А. И.* Точные оценки вероятности переобучения для симметричных семейств алгоритмов // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009. — С. 66–69.

[7] *Грэхем Р., Кнут Д., Паташник О.* Конкретная математика. — М.: Мир, 1998.

[8] *Винберг Э. Б.* Курс алгебры // М.: Факториал Пресс, 2001. — 544 с.