Московский физико-технический институт

Факультет Управления и Прикладной Математики

Кафедра «Интеллектуальные Системы»

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ СТУДЕНТА 774 ГРУППЫ

«Концентрация меры в комбинаторных оценках обобщающей способности»

Выполнил:

студент 6 курса 774 группы

Животовский Никита Кириллович

Научный руководитель:

д.ф-м.н., профессор

Воронцов Константин Вячеславович

Москва, 2013

Содержание

Аннотация

В работе в рамках комбинаторного подхода [21] исследуются возможность получения оценок обобщающей способности, не зависящих от ненаблюдаемой контрольной выборки. Для решения этой проблемы в первой части работы уточняются оценки ожидаемой переобученности, а во второй части выводятся специальные неравенства концентрации меры, с помощью которых затем оценивается частота ошибок на ненаблюдаемой контрольной выборке.

1 Введение

1.1 Статистическая теория обучения

Одной из важнейших задач теории статистического обучения [3, 10] является получение оценок обобщающей способности. В простейшем случае данная задача может быть сформулирована следующим образом:

Рассмотрим X — *обучающую* выборку, состоящую из n объектов, каждый из которых является реализаций некоторой случайной величины, определённой на фиксированном вероятностном пространстве, одинаковом для всех объектов. Общую для всех объектов вероятностную меру будем обозначать $\mathbf{P}$.

Пусть также нам дан класс измеримых функций $\mathcal{F}$, которые мы будем называть *алгоритмами*. Наша цель найти среди них тот $\mathfrak{f}$, который минимизирует *ожидаемый риск*

$$\mathfrak{f} = \arg\min_{f \in \mathbf{F}} \int f d\mathbf{P}$$

Заметим, что в нашем случае алгоритмы фактически отождествлены с их функциями потерь. Минимизация ожидаемого риска — это поиск такого алгоритма, ожидаемая ошибка которого минимальна. Но в реальных задачах мера $\mathbf{P}$ неизвестна.

Таким образом, решение о выборе алгоритма $\mathfrak{f}$ приходится производить лишь по реализации обучающей выборки. Разумным в этой ситуации может представиться следующий подход. Имея обучающую выборку, можно построить эмпирическую меру $\mathbf{P}_n$, являющуюся в широком смысле приближением реальной меры $\mathbf{P}$. Затем в качестве оценки для $\mathfrak{f}$ выступает алгоритм

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \int f d\mathbf{P}_n$$

Связь между алгоритмами $\hat{f}$ и $\mathfrak{f}$ впервые была подробно исследована в рамках теории Вапника-Червоненкиса [18]. В ней исследовалась равномерная по классу $\mathcal{F}$ оценка разности

$$\left| \int f d\mathbf{P}_n - \int f d\mathbf{P} \right|$$

и её уклонений от нуля, не зависящих от $\mathbf{P}$.

Практически сразу стало ясно, что оцениваемая величина в первую очередь зависит от «геометрии» класса $\mathcal{F}$. В результате возникла серия оценок, зависящих

от глобальных характеристик класса алгоритмов. Однако полученные оценки были очень общими, существенно завышенными и минимально опирающимися на наблюдаемую обучающую выборку. Более того, фактически они давали лишь достаточные условия, которые нужно наложить на $\mathcal{F}$, при которых происходит равномерная по классу сходимость по вероятности $\left|\int f d\mathbf{P}_n - \int f d\mathbf{P}\right| \to 0$ при $n \to \infty$.

Следующим важным шагом в исследовании проблемы было развитие теории эмпирических процессов [13, 8] и неравенств концентрации меры [15]. Современный подход к оценке $\sup\limits_{f\in\mathcal{F}}\left|\int f d\mathbf{P}_n - \int f d\mathbf{P}\right|$ заключается в следующем. На первом этапе производится мажорирование данной невычислимой по наблюдаемой выборке величины ожидаемым значением *Радемахеровского процесса* [10]

$$\mathbf{E}\sup_{f\in\mathcal{F}}\left|\int f d\mathbf{P}_n - \int f d\mathbf{P}\right| \leq 2\mathbf{E}\left(\mathcal{R}_n\right) = 2\mathbf{E}\left(\mathbf{E}_\varepsilon \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\varepsilon_i f(x_i)\right|\right),$$

где $\mathbf{E}$ – математическое ожидание соответствующее мере $\mathbf{P}$ и влияющее на эмпирическую меру $\mathbf{P}_n$, ε_i - независимые случайные величины, принимающие равновероятно значения ± 1, а $\mathbf{E}_\varepsilon$ - соответствующее им математическое ожидание. Отметим, что введённый процесс зависит лишь от выборок конечной длины и фактически является мерой сложности класса $\mathcal{F}$. Как и ожидается, величина Радемахеровского процесса контролируется «геометрией» класса $\mathcal{F}$ [10, 8, 13]. В недавней работе [5] в более общих терминах описано, какие глобальные характеритиски $\mathcal{F}$ контролируют реализации $\mathcal{R}_n$ с точностью до универсальных констант как сверху так и снизу.

Затем с помощью неравенств концентрации меры выводятся вероятностные оценки, контролирующие отклонение величины $\sup\limits_{f\in\mathcal{F}}\left|\int f d\mathbf{P}_n - \int f d\mathbf{P}\right|$ от её математического ожидания, которое мы можем оценить с помощью введённого Радемахеровского процесса. Основным инструментом здесь является неравенство Талаграна [15] и различные его вариации [10]. Примечательно, что и эти неравенства существенно используют глобальные характеристики семейства $\mathcal{F}$. В итоге получаются оценки вычислимые по наблюдаемой выборке и не зависящие от неизвестной меры $\mathbf{P}$.

1.2 Комбинаторная теория переобучения

Альтернативным подходом является комбинаторная теория переобучения, развитая в [19, 21]. Введем основные понятия и обозначения, которые будут использоваться на протяжении всей работы.

Пусть задана конечная генеральная совокупность *объектов* $\mathbb{X} = \{x_1, \ldots, x_L\}$, конечное множество *алгоритмов* $A = \{a_1, \ldots, a_D\}$ и бинарная *функция потерь* $I\colon A \times \mathbb{X} \to \{0, 1\}$, где $I(a, x) = 1$ означает, что алгоритм a ошибается на объекте x. *Вектором ошибок* алгоритма a называется бинарный вектор $\big(I(a, x_1), \ldots, I(a, x_L)\big)$ размерности L. Предполагается, что векторы ошибок всех алгоритмов из A попарно различны.

Для произвольного $a \in A$ и выборки $X \subseteq \mathbb{X}$ вводятся обозначения для числа и частоты ошибок:

$$n(a, X) = \sum_{x \in X} I(a, x), \qquad \nu(a, X) = \frac{n(a, X)}{|X|}.$$

Методом обучения называется отображение $\mu\colon 2^{\mathbb{X}} \to A$, которое произвольной выборке $X \subset \mathbb{X}$ ставит в соответствие некоторый алгоритм $\mu X \in A$.

Метод обучения μ называется методом *минимизации эмпирического риска* (МЭР), если $\mu X \in A(X)$ для всех выборок $X \subset \mathbb{X}$, где

$$A(X) = \operatorname{Arg}\min_{a \in A} n(a, X), \quad X \subset \mathbb{X}.$$

В новых обозначениях в статистической теории обучения перед нами ставится задача получения оценок вероятности ошибки $P(a) = \mathbf{E} I(a, x)$ для алгоритма $a = \mu X$. Для этого оценивается правый хвост распределения *переобученности* алгоритма a — разности вероятности ошибки и частоты ошибок на обучающей выборке:

$$P_\varepsilon(\mu) = \mathsf{P}\big(X\colon P(\mu X) - \nu(\mu X, X) \geq \varepsilon\big), \quad \varepsilon \in (0, 1).$$

Чтобы избавиться от сложностей, связанных с анализом метода обучения μ, и получить универсальную оценку, справедливую для любого метода, оценивается вероятность большого равномерного отклонения частоты ошибок от вероятности:

$$P_\varepsilon(\mu) \leq \mathsf{P}\big(X\colon \sup_a \big(P(a) - \nu(a, X)\big) \geq \varepsilon\big).$$

Комбинаторная теория переобучения основана на более слабых вероятностных допущениях. Предполагается, что все $L!$ перестановок объектов конечной генеральной совокупности $\mathbb{X}$ равновероятны. Сами объекты предполагаются неслучайными, никакой меры на множестве объектов не вводится, и даже не предполагается существование каких-то других объектов кроме $\mathbb{X}$. Случайным считается только порядок появления объектов. Предполагается, что выборка X из первых ℓ объектов наблюдалась в прошлом, выборка $\bar{X}$ из оставшихся $k = L - \ell$ объектов пока неизвестна и будет наблюдаться в будущем. В этот момент методом μ выбирается алгоритм $a = \mu X$, и интересует оценка частоты его ошибок $\nu(a, \bar{X})$ на будущих данных. Эта оценка характеризует обобщающую способность метода μ и на практике может использоваться в качестве критерия выбора семейства алгоритмов A или метода обучения μ.

В комбинаторной теории понятие «вероятности ошибки» не определяется, оцениваются только частоты ошибок на конечных выборках. В дальнейшем все используемые функции выборок X и $\bar{X}$ будут инвариантны относительно перестановок объектов внутри этих выборок. Поэтому основное вероятностное предположение можно ещё немного ослабить, считая равновероятными все C_L^ℓ разбиений генеральной совокупности $\mathbb{X} = X \sqcup \bar{X}$ на две выборки — *наблюдаемую обучающую* X длины ℓ и *скрытую контрольную* $\bar{X}$ длины k.

Вероятность переобучения метода μ на выборке $\mathbb{X}$ определяется как доля разбиений, при которых частота ошибок на контроле превосходит частоту ошибок на обучении на ε или более:

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

В отличие от статистической теории обучения в комбинаторном подходе рассматриваются лишь бинарные функции потерь. Однако, это ограничение делает рассмотрение метрических свойств A очень удобным. Поэтому введём на множестве алгоритмов A, как на бинарных векторах ошибок, естественное отношение порядка и метрику

Хэмминга: для любых $a, b \in A$

$$(a \le b) \;\leftrightarrow\; (I(a,x) \le I(b,x),\ \forall x \in \mathbb{X});$$

$$(a < b) \;\leftrightarrow\; (a \le b \;\text{ и }\; a \ne b);$$

$$\rho(a,b) = \sum_{i=1}^{L} \big[I(a,x_i) \ne I(b,x_i)\big].$$

Если $a \le b$ и при этом $\rho(a,b) = 1$, то будем говорить, что a *предшествует* b и записывать $a \prec b$.

Графом расслоения–связности множества алгоритмов A будем называть направленный граф $\langle A, E\rangle$ с множеством рёбер $E = \big\{(a,b)\colon a \prec b\big\}$.

Граф расслоения–связности является многодольным, доли соответствуют *слоям алгоритмов* $A_m = \big\{a \in A\colon n(a,\mathbb{X}) = m\big\}$, рёбрами могут соединяться только алгоритмы соседних слоёв. Каждому ребру (a,b) соответствует единственный объект $x_{ab} \in \mathbb{X}$, такой, что $I(a,x_{ab}) = 0$ и $I(b,x_{ab}) = 1$.

Заметим, что если для любых $a, b \in A,\ a < b$ существует путь $a \prec a_1 \cdots \prec a_s = b$, то граф расслоения–связности совпадает с диаграммой Хассе отношения порядка, введённого на множестве алгоритмов A. Однако в общем случае он является лишь подграфом диаграммы Хассе.

Порождающим множеством X_a алгоритма a называется множество объектов, соответствующих исходящим из вершины a рёбрам:

$$X_a = \big\{x \in \mathbb{X} \;\big|\; \exists b \in A\colon a \prec b,\ I(a,x) < I(b,x)\big\}.$$

Запрещающим множеством X'_a алгоритма a называется множество объектов x, на которых алгоритм a ошибается, при том, что существует алгоритм $b \in A,\ b < a$, не ошибающийся на x:

$$X'_a = \big\{x \in \mathbb{X} \;\big|\; \exists b \in A\colon b < a,\ I(b,x) < I(a,x)\big\}.$$

Верхней связностью алгоритма a называется число рёбер графа, исходящих из вершины a:

$$q(a) = |X_a|.$$

Нижней связностью алгоритма a называется число рёбер графа, входящих в вершину a:

$$d(a) = |\{x_{ba} \in \mathbb{X} : b \prec a\}|.$$

Неполноценностью алгоритма a называется размер запрещающего множества алгоритма a:

$$r(a) = |X'_a|.$$

Также на понадобится понятие *ёмкости*:

Пусть $\mathbb{X}$ получена из некоторого (возможно бесконечного) множества $\mathcal{X}$ объектов. Будем считать, что имеется некоторое семейство $A(\mathcal{X})$ алгоритмов, из которых $A(\mathbb{X})$ получается сужением на конкретную генеральную выборку $\mathbb{X}$.

Тогда функцию $\Delta(L) = \max\limits_{\mathbb{X}\subset\mathcal{X}} |A(\mathbb{X})|$ называется функцией роста, где L соответственно длина генеральной выборки $\mathbb{X}$. Очевидно, что $\Delta(L) \leq 2^L$. Число V называется *ёмкостью* семейства $A(\mathcal{X})$, если V — наибольшее из натуральных чисел, для которого $\Delta(V) = 2^V$. Если такого числа не существует, то полагают, что $V = \infty$.

Данная характеристика семейства подробно описана в [8, 10, 18].

Определим функцию гипергеометрического распределения

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

Если множество $A(X)$ содержит более одного элемента, то выбор алгоритма методом МЭР не однозначен. Будем рассматривать худший случай. Метод МЭР называется *пессимистичным*, если

$$\mu X \in \operatorname{Arg}\max_{a\in A(X)} n(a, \bar{X}), \quad X \subset \mathbb{X}.$$

Точные оценки вероятности переобучения для пессимистичного МЭР являются верхними оценками вероятности переобучения для произвольного МЭР.

Следующая лемма, доказанная в [21], описывает важное свойство порождающего и запрещающего множеств.

Лемма 1.1. *Пусть μ — пессимистичная минимизация эмпирического риска, тогда $\forall a \in A$ выполнено следующее равенство:*

$$\big[\mu X = a\big] \leq \big[X_a \subset X\big]\big[X'_a \subset \bar{X}\big]$$

Таким образом, для того чтобы пессимистичная минимизация эмпирического риска выбрала некоторый алгоритм необходимо, чтобы его порождающее и запрещающее множества были соответственно подмножествами обучающей и контрольной подвыборок.

Комбинаторная оценка вероятности переобучения существенно зависит от характеристик $q(a)$, $r(a)$ каждого алгоритма $a \in A$.

Теорема 1.1 (Оценка расслоения–связности [21]). *Пусть μ — метод пессимистичной минимизации эмпирического риска. Тогда для любого $\varepsilon \in (0,1)$*

$$Q_\varepsilon(\mu, \mathbb{X}) \leq \sum_{a \in A} \frac{C_{L-q-r}^{\ell-q}}{C_L^\ell} H_{L-q-r}^{\ell-q,\, m-r}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right),$$

где q — верхняя связность, r — неполноценность алгоритма a, $m = m(a) = n(a, \mathbb{X})$.

Оценка может показаться недостаточно наглядной, тем не менее она имеет простой смысл. Вероятность переобучения оценивается в виде суммы по всем алгоритмам произведений величин $P(a) = \frac{C_{L-q-r}^{\ell-q}}{C_L^\ell} = \mathsf{P}\big[X_a \subset X\big]\big[X'_a \subset \bar{X}\big]$ и условной вероятности переобучения одного алгоритма, задаваемого хвостом гипергеометрического распределения.

Величина $P(a)$ определяет при этом «вес» каждого алгоритма в оценке, то есть долю разбиений, на которых он может быть выбран методом обучения. Фактически оценка представляет собой комбинаторный аналог оценок Вапника [19], в котором учтён метод обучения за счёт множителей $P(a)$.

1.3 Структура работы

Работа выполнена в рамках комбинаторного подхода. Все результаты связаны непосредственно с бинарной функцией потерь и вероятностным предположением, используемым в этом подходе.

Основным недостатком существующих комбинаторных оценок является то, что они зависят от матрицы ошибок на всей генеральной выборке. Поэтому целью работы является получение максимально точных оценок обобщающей способности, которые вычислимы лишь по наблюдаемой выборке. Заметим, что отказ от наблюдения всей матрицы ошибок при применении техник схожих с теми, что используются при доказательстве теоремы 1.1, не позволит включить в оценку данные, вычисленные по

наблюдениям. Это аналогично уже описанному случаю, когда оценки Вапника [18] практически не опирались на информацию, полученную из наблюдаемой обучающей выборки.

Для решения задачи предлагается следующий подход. Зафиксируем $\ell = k = \frac{L}{2}$. Введём функционал равномерной *ожидаемой переобученности*

$$\mathcal{EOF}_{\max}(\mathbb{X}) = \mathbf{E} \max_{a \in A} \left(\nu(a, \bar{X}) - \nu(a, X) \right)$$

и функционал ожидаемой переобученности, учитывающий метод обучения μ (которым в данной работе будет являться ПМЭР)

$$\mathcal{EOF}_{\mu}(\mathbb{X}) = \mathbf{E} \left(\nu(\mu X, \bar{X}) - \nu(\mu X, X) \right).$$

В разделах 2 и 3 будут даны верхние оценки для обоих функционалов, зависящие от характеристик матрицы ошибок на всей $\mathbb{X}$, а также показана их аналогия с Радемахеровским процессом. Заметим, что в рамках комбинаторного подхода и раньше существовали оценки ожидаемой переобученности, но они росли линейно по вкладам алгоритмов, когда как полученные оценки растут лишь логарифмически.

В разделе 3 также будет доказана общая теорема, относящаяся к факторам завышенности полученных оценок и обобщающая исследовавшиеся ранее модельные семейства алгоритмов.

В разделе 4 будут применяться специальные неравенства концентрации меры. Напомним, что в комбинаторном подходе мы имеем дело с выборками без возвращений, поэтому стандартные результаты применены быть не могут. Поэтому сначала доказывается вариант изопериметрического неравенства на слоях дискретного куба. Затем этот результат применяется для получения оценки частоты ошибок на ненаблюдаемой контрольной выборке в равномерном случае.

В этом же разделе используются известные логарифмические неравенства Соболева для слоев дискретного куба, которые влекут вариант неравенств концентрации типа МакДиармида. Далее вводятся естественные ограничения устойчивости на метод обучения и семейство алгоритмов и выводится оценка частоты ошибок на ненаблюдаемой контрольной выборке, вычислимая по наблюдаемой обучающей выборке.

2 Равномерные оценки обобщающей способности

В данном разделе будут доказана серия оценок ожидаемой переобученности и complete cross-validation

$$\mathcal{CCV}_{max} = \mathbf{E} \max_{a \in A} \nu(a, \bar{X})$$

Напомним, что в рамках комбинаторного подхода оценки ожидаемой переобученности и $\mathcal{CCV}_{max}$ были линейны по вкладам алгроитмов. Предлагаемые оценки имеют гарантированную асимптотику $O\left(\sqrt{\ln\left(|A|\right)}\right)$ по числу алгоритмов семейства.

Для получения оценок сначала будет оценена производящая функция моментов функционала переобученности, доказана субгауссовость с оптимальными значениями параметров. Затем с помощью неравенства Буля будет оценена производящая функция моментов максимума из конечного числа субгауссовских случайных величин.

Для получения нетривиальных результатов нужно учесть специфику вероятностного предположения, используемого в комбинаторном подходе. Сначала нам понадобится некоторая техническая лемма.

Функция

$${}_2F_1\big(a, b, c, z\big) = 1 + \sum_{k=1}^{\infty} \left[\prod_{l=0}^{k-1} \frac{(a+l)(b+l)}{(1+l)(c+l)} \right] z^k$$

с параметрами a, b, c, определённая в круге $|z| < 1$, называется *гипергеометрической*.

Лемма 2.1. *Для целых чисел m, l, таких что $0 \leq m \leq l$ и действительного $z \in [0, 1]$*

$${}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2} - \ell, z\right) \leq 1$$

Доказательство данной леммы чисто техническое и перенесено в конец раздела.

2.1 Оценка ожидаемой переобученности

Следующим шагом будет оценка производящей функции моментов

Лемма 2.2. *Пусть $a \in A$, $\ell = k = \frac{L}{2}$ и $m(a) \leq \ell$, тогда для всех $\lambda > 0$*

$$\mathbf{E} \exp\big(\lambda(n(a, \bar{X}) - n(a, X))\big) \leq (\cosh(\lambda))^{\ell} {}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2} - \ell, (\tanh(\lambda))^2\right)$$

Доказательство. Произвольному разбиению генеральной выборки сопоставим вектор $\sigma = (\sigma_1, \ldots, \sigma_L)$, где на половине позиций стоят -1, которые соответствуют позициям обучающей выборки в $\mathbb{X}$, а на остальных позициях 1. Введение такого определения объясняется тем, что в комбинаторном подходе мы имеем дело с равномерным распределением на слое $\{1, -1\}^L$ куба. Без ограничения общности перенумеруем генеральную выборку так чтобы алгоритм ошибался на первых m объектах, а на оставшихся не допускал ошибок. Производящая функция моментов может быть записана в виде

$$\mathbf{E}_\sigma \exp\left(\lambda \sum_{i=1}^{L} \sigma_i I(a, x_i)\right).$$

Обозначим $\hat{x}_i = \sigma_i I(a, x_i)$, тогда с учётом $m(a) \leq \ell$

$$\mathbf{E}_\sigma \exp\big(\lambda \sum_{i=1}^{L} \sigma_i I(a, x_i)\big) = \mathbf{E}_\sigma \prod_{i=1}^{L} \exp(\lambda \hat{x}_i) = \mathbf{E}_\sigma \prod_{i=1}^{\ell} \exp(\lambda \hat{x}_i)$$

Очевидно, что для всех i: $\lambda \hat{x}_i \in [-\lambda, \lambda]$, поэтому используем выпуклость экспоненты:

$$\exp(\lambda \hat{x}_i) \leq \frac{\lambda \hat{x}_i + \lambda}{\lambda + \lambda} \exp(\lambda) + \frac{-\lambda \hat{x}_i + \lambda}{\lambda + \lambda} \exp(-\lambda) = \hat{x}_i \sinh(\lambda) + \cosh(\lambda).$$

Подставляем в раннее выражение:

$$\mathbf{E}_\sigma \left(\prod_{i=1}^{\ell} \exp(\lambda \hat{x}_i)\right) \leq \mathbf{E} \prod_{i=1}^{\ell} (\hat{x}_i \sinh(\lambda) + \cosh(\lambda))$$

Теперь будем раскрывать скобки в полученном выражении и учтём, что последние $\ell - m$ значения $\hat{x}_i$ тождественно равны нулю, так как алгоритм не ошибается на эти объектах. Учтём, что

$$\mathbf{E} \prod_{i=1}^{\ell} (\hat{x}_i \sinh(\lambda) + \cosh(\lambda)) = (\cosh(\lambda))^{\ell - m} \mathbf{E} \prod_{i=1}^{m} (\hat{x}_i \sinh(\lambda) + \cosh(\lambda)).$$

Теперь раскроем скобки в произведении. Очевидно, что для всех $i \leq m$ математическое ожидание произведения одинакового числа $\hat{x}_i$, соответствующих различным

индексам, одинаково. Таким образом

$$(\cosh(\lambda))^{\ell-m}\mathbf{E}\prod_{i=1}^{m}(\hat{x}_i\sinh(\lambda)+\cosh(\lambda)) =$$

$$(\cosh(\lambda))^{\ell-m}\mathbf{E}(\cosh^m(\lambda)+C_m^1\hat{x}_1\sinh^1(\lambda)\cosh^{m-1}(\lambda)+$$

$$C_m^2\hat{x}_1\hat{x}_2\sinh^2(\lambda)\cosh^{m-2}(\lambda)+\ldots+\hat{x}_1\ldots\hat{x}_m\sinh^m(\lambda)) =$$

$$(\cosh(\lambda))^{\ell}\mathbf{E}(1+C_m^1\hat{x}_1\tanh^1(\lambda)+\ldots+\hat{x}_1\ldots\hat{x}_m\tanh^m(\lambda)).$$

Рассмотрим для неотрицательного целого r выражение $\mathbf{E}(\hat{x}_1\ldots\hat{x}_{2r+1})$.

По условию $2r+1\leq\ell$, тогда для каждого разбиения генеральной выборки на обучение и контроль знак $\hat{x}_1\ldots\hat{x}_{2r+1}$ зависит лишь от четности числа элементов $x_1,\ldots,x_{2r+1}$ попавших в контроль. Очевидно, что число разбиений на которой всё нечётное число этих объектов в обучении равно числу разбиений, когда все они в контроле. Вклад таких разбиений в математическое ожидание просто противоположен по знаку. Также одинаков по модулю и противоположен по знаку вклад разбиений, где лишь один из перечисленных объектов в обучении и где все кроме одного в обучении. И так далее компенсируем вклады всех 2^{2r+1} вариантов помещения части объектов $\hat{x}_1\ldots\hat{x}_{2r+1}$ в обучение.

Пусть $j=2r$. Теперь будем выводить точную формулу для $E(\hat{x}_1\ldots\hat{x}_j)$.

Как и ранее сосчитаем вклады разбиений в математическое ожидание. Введём суммирование по i — числу объектов $x_1,\ldots,x_j$, попавших в обучение. Очевидно, что вклад разбиения равен $(-1)^i$. Число разбиений сосчитать не сложно: сначала выберем i позиций среди j для для помещения в обучение, оставшиеся объекты помещаем в контроль. При этом мы еще не учли все возможные разбиения, а именно, нужно среди оставшихся $2l-j$ объектов выбрать $\ell-j+i$ в контроль. Объединяя результат получаем формулу

$$\mathbf{E}(\hat{x}_1\ldots\hat{x}_j)=\frac{\sum\limits_{i=0}^{j}(-1)^iC_{2\ell-j}^{\ell-j+i}C_j^i}{C_{2\ell}^{\ell}}.$$

Это же выражение верно и для нечётных j, при этом оно тождественно равно нулю. В итоге получаем

$$\mathbf{E}(1+C_m^1\hat{x}_1\tanh^1(\lambda)+\ldots+\hat{x}_1\ldots\hat{x}_m\tanh^m(\lambda))=\frac{\sum\limits_{j=0}^{m}C_m^j(\tanh(\lambda))^j\sum\limits_{i=0}^{j}(-1)^iC_{2\ell-j}^{\ell-j+i}C_j^i}{C_{2\ell}^{\ell}}.$$

Далее остается доказать, что выполнено равенство

$$\frac{\sum\limits_{j=0}^{m} C_m^j (\tanh(\lambda))^j \sum\limits_{i=0}^{j} (-1)^i C_{2\ell-j}^{\ell-j+i} C_j^i}{C_{2\ell}^{\ell}} = {}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2} - \ell, (\tanh(\lambda))^2\right).$$

Это утверждение чисто техническое, доказывается индукцией по параметрам. □

Если оценивать сумму несколько более аккуратно, то в условиях предыдущей леммы можно получить более точное выражение

$$\mathbf{E}\exp\left(\lambda(n(a,\bar{X}) - n(a,X))\right) \leq (\cosh(\lambda))^{m(a)} {}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2} - \ell, (\tanh(\lambda))^2\right).$$

Тем не менее, во избежание черезмерной громоздкости мы будем использовать более грубый результат леммы 2.2.

После этого можно сформулировать основную теорему данного раздела. Это некоторое обобщение леммы о ожидаемом максимуме субгауссовских случайных величин [8, 10].

Теорема 2.1. *Пусть* $\ell = k = \frac{L}{2}$ *и* $\max\limits_{a\in A} m(a) \leq \frac{L}{2}$, *тогда*

$$\mathcal{EOF}_{\max} \leq$$
$$\min_{\lambda>0} \frac{1}{\lambda}\left(\ln(\cosh(\lambda)) + \frac{1}{\ell}\ln\left(\sum_{s=0}^{l} \Delta_s \; {}_2F_1\left(\frac{1-s}{2}, -\frac{s}{2}, \frac{1}{2} - \ell, (\tanh(\lambda))^2\right)\right)\right) \leq$$
$$\sqrt{\frac{2\ln(|A|)}{l}},$$

где Δ_s *— число алгоритмов в* s*-ом слое семейства алгоритмов.*

Доказательство. Распишем выкладки максимально подробно. В обозначения предыдущей леммы оценивается величина

$$\lambda\mathbf{E}\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right) = \ln\left(\exp\left(\lambda\mathbf{E}\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\right).$$

По неравенству Йенсена:

$$\ln\left(\exp\left(\lambda\mathbf{E}\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\right) \leq \ln\left(\mathbf{E}\exp\left(\lambda\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\right).$$

Максимум может быть вынесен

$$\ln\left(\mathbf{E}\exp\left(\lambda\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\right) = \ln\left(\mathbf{E}\max_{a\in A}\exp\left(\lambda\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\right).$$

По неравенству Буля

$$\ln\left(\mathbf{E}\max_{a\in A}\exp\left(\lambda\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\right)\leq\ln\left(\mathbf{E}\sum_{a\in A}\exp\left(\lambda\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right).$$

По лемме 2.2

$$\ln\left(\mathbf{E}\sum_{a\in A}\exp\left(\lambda\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\leq$$
$$\ln\left(\sum_{a\in A}(\cosh(\lambda))^{\ell}{}_2F_1\left(\frac{1-m}{2},-\frac{m}{2},\frac{1}{2}-\ell,(\tanh(\lambda))^2\right)\right).$$

Из цепочки неравенств в предыдущем шаге имеем

$$\mathbf{E}\max_{a\in A}(\sum_{i=1}^{L}\sigma_i I(a,x_i))=\frac{1}{\lambda}\ln\left((\cosh(\lambda))^{\ell}\sum_{a\in A}{}_2F_1\left(\frac{1-m}{2},-\frac{m}{2},\frac{1}{2}-\ell,(\tanh(\lambda))^2\right)\right).$$

Для доказательства первого из неравенств теперь достаточно прологарифмировать данное выражение и минимизировать его по λ. Затем ввести суммирование по слоям, так как в каждом слагаемом суммы алгоритм характеризуется лишь числом ошибок.

Для того чтобы доказать верхнее неравенство нужно воспользоваться элементарными неравенствами

$$\cosh(\lambda)\leq\exp\left(\frac{\lambda^2}{2}\right)\quad\text{и}\quad(\cosh(\lambda))^{\ell}\leq\exp\left(\frac{\ell\lambda^2}{2}\right).$$

С учётом этих неравенств и леммы 2.1

$$\mathbf{E}\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\leq\min_{\lambda>0}\frac{1}{\lambda}\left(\ln(|A|)+\frac{\ell\lambda^2}{2}\right)=\sqrt{2\ell\ln(|A|)}.$$

Поделив обе части неравенства на ℓ правое неравенство из утверждения теоремы. На этом шаге также можно проследить аналогию со введённым ранее Радемахеровским процессом $\mathcal{R}_n$. Действительно, в нём мы имели дело с равномерным на $\{-1,1\}^n$ кубе распределением, представленным в виде ε_i, равных равновероятно ±1. В комбинаторной теории логичным аналогом этого процесса является ожидаемая переобученность $\mathcal{EOF}$. Заметим, что в случае конечного семейства алгоритмов значение $\mathcal{R}_n$ также имеет гарантированную асимптотику $O\left(\sqrt{\ln(|A|)}\right)$ по числу алгоритмов в семействе [10, 8]. □

Доказанная теорема явно учитывает расслоение семейства алгоритмов A по числу ошибок. Кроме этого, в ней было доказано даже большее, а именно оценена экспоненциальная функция моментов для $\max\limits_{a\in A}(\nu(a,\bar{X})-\nu(a,X))$. Грубая из оценок теоремы 2.1 даёт для $\lambda>0$

$$\mathbf{E}\exp\left(\lambda\max_{a\in A}(\nu(a,\bar{X})-\nu(a,X))\right)\leq|A|\exp\left(\frac{\lambda^2\ell}{2}\right).$$

Отсюда с помощью неравенства Маркова можно получить хорошо известное неравенство. Для $t>0$

$$\mathsf{P}\left(\max_{a\in A}(\nu(a,\bar{X})-\nu(a,X))>t\right)=$$
$$\mathsf{P}\left(\exp\left(\lambda\max_{a\in A}(\nu(a,\bar{X})-\nu(a,X))\right)>\exp(\lambda t)\right)\leq|A|\exp\left(\frac{\lambda^2\ell}{2}-\lambda t\right).$$

Оптимизируя по λ, получаем

$$\mathsf{P}\left(\max_{a\in A}(\nu(a,\bar{X})-\nu(a,X))>t\right)\leq|A|\exp\left(\frac{-t^2\ell}{2}\right).$$

Подобные оценки получались напрямую в комбинаторном подходе, и в [18]. Они, однако, очень завышены, но тем не менее указывают на принципиальный момент.

В случае, если $|A|$ растет с ростом ℓ, например, как показательная функция, то при $\ell\to\infty$ величина $\max\limits_{a\in A}(\nu(a,\bar{X})-\nu(a,X))$ сходится по вероятности к нулю. Действительно, легко показать, что верно аналогичное неравенство и для величины $\max\limits_{a\in A}(\nu(a,X)-\nu(a,\bar{X}))$.

В частности, в случае, если мы имеем дело с семейством алгоритмов конечной ёмкости V, то $|A(\mathbb{X})|\leq(2\ell+1)^V$ [18, 8]. Это означает, что в данном случае в асимптотике частота ошибок на обучении близка к ошибке на ненаблюдаемой контрольной выборке.

2.2 Оценка complete cross-validation

Теорема 2.2. *Пусть* $\ell=k=\frac{L}{2}$ *и* $\max\limits_{a\in A}m(a)\leq\frac{L}{2}$, *тогда*

$$\mathcal{CCV}_{\max}\leq$$
$$\min_{\lambda>0}\frac{1}{\lambda}\left(\ln(\cosh(\lambda))+\frac{1}{\ell}\ln\left(\sum_{s=0}^{l}\Delta_s\ {}_2F_1\left(-\ell,-s,-2\ell,-\tanh(\lambda)\right)\right)\right),$$

где Δ_s *— число алгоритмов в* s*-ом слое семейства алгоритмов.*

Доказательство. Все шаги доказательства полностью аналогичны теореме 2.1. Легко показать, что в данной теореме в выражении

$$\mathbf{E}(\hat{x}_1 \dots \hat{x}_j) = \frac{\sum\limits_{i=0}^{j} (-1)^i C_{2\ell-j}^{\ell-j+i} C_j^i}{C_{2\ell}^{\ell}}$$

нужно учитывать лишь слагаемое, соответствующее $i = 0$, что и заменит разность частот на одну частоту.

После этого вместо ограниченной функции

$${}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2} - \ell, (\tanh(\lambda))^2\right)$$

получится уже неограниченная

$${}_2F_1\left(-\ell, -m, -2\ell, -\tanh(\lambda)\right),$$

после этого шаги доказательства опять повторяются. □

2.3 Мажорирующие меры для уточнения равномерной оценки

Равномерные по семейству алгоритмов неравенства из предыдущих разделов очень не оптимально использовали неравенство Буля. Вклад в некоторой степени близких алгоритмов в ожидаемую переобученность практически одинаков и простое неравенство Буля этого не учитывает. Однако, в теории эмпирических процессов в работах Талаграна [13] разработан аппарат generic chaining, позволяющий гораздо лучше использовать геометрические свойства пространства состояний (в нашем случае совпадающим со множеством алгоритмов). Альтернативным и часто используемым названием данного подхода являются *мажорирующие меры*. Заметим, что и комбинаторный подход существенно опирается на геометрические свойства семейства алгоритмов. Однако, в нём это происходило за счёт учёта метода обучения, мажорирующие меры же работают с равномерной оценкой, поэтому и опираются на другие свойства.

Стоит также отметить, что результат ранее описанной теоремы из [5], контролирующий Радемахеровский процесс $\mathcal{R}_n$, существенно опирается именно на оценки, полученные с помощью мажорирующих мер. Говоря неформально, с точностью до

поправок и констант аппарат мажорирующих мер строит некоторую глобальную характеристику класса $\mathcal{F}$, которая полностью описывает поведение соответствующего Радемахеровского процесса.

Поэтому мы в некоторой простой форме применим некоторые очень общие результаты данной теории и получим вычислимые оценки ожидаемой переобученности, более точные чем те, что предложены в предыдущем разделе.

Для получения более сильных результатов в данном разделе мы будем работать с меньшей метрикой на семействе алгоритмов. Обозначим для пары алгоритмов a, b

$$d(a,b) = \sqrt{\rho(a,b)}.$$

Для удобства мы введём эмпирический процесс

$$Y_a = \sum_{i=1}^{L} \sigma_i I(a, x_i).$$

В качестве пространства состояний в нём выступает наше семейство алгоритмов A. Следующая лемма даёт нам условие субгауссовского приращения введённого процесса.

Лемма 2.3. *Для* $\ell = k$, $\max\limits_{a \in A} m(a) \leq \frac{\ell}{2}$, $a \neq b$ *и* $\lambda > 0$ *имеет место неравенство*

$$\mathbf{E} \exp\left(\lambda \frac{Y_a - Y_b}{d(a,b)}\right) \leq \exp\left(\lambda^2\right).$$

Доказательство. Доказательство представляет собой некоторое обобщение леммы 2.2.

Действительно, разность $Y_a - Y_b$ можно рассматривать как процесс, определенный на покомпонентной разности Y_{a-b}, где $a - b$ — вектор, состоящий из $\{-1, 1, 0\}$. Легко видеть, что число ± 1 в нём не превосходит $d(a,b)^2$. Обозначим в нём число 1 как m_1, а число -1 как m_2. Тогда $m_1 + m_2 = d(a,b)^2$. Введём $\hat{x}_i = \sigma_i \left| I(a, x_i) - I(b, x_i) \right|$. Переупорядочим генеральную выборку, так чтобы в $a - b$ сначала шли 1, затем -1, а на оставшихся позицях 0.

Теперь, используя те же выкладки, что и в лемме 2.2 и теореме 2.1, получаем с помощью неравенства Коши-Буняковского и леммы 2.1

$$\mathbf{E}\exp\left(\lambda Y_{a-b}\right) = \mathbf{E}\prod_{i=1}^{m_1}\exp(\lambda\hat{x}_i)\prod_{i=m_1+1}^{m_1+m_2}\exp(-\lambda\hat{x}_i) \leq$$

$$\sqrt{\mathbf{E}\prod_{i=1}^{m_1}\exp(2\lambda\hat{x}_i)}\sqrt{\mathbf{E}\prod_{i=m_1+1}^{m_1+m_2}\exp(-2\lambda\hat{x})} \leq$$

$$\sqrt{(\cosh(2\lambda))^{m_1}\mathbf{E}\prod_{i=1}^{m_1}\left(\hat{x}_i\tanh(2\lambda)+1\right)}\sqrt{(\cosh(2\lambda))^{m_2}\mathbf{E}\prod_{i=1}^{m_2}\left(-\hat{x}_i\tanh(2\lambda)+1\right)} \leq$$

$$\exp\left(2\lambda^2\right)^{\frac{m_1+m_2}{2}} = \exp\left(\lambda^2 d(a,b)^2\right).$$

Отсюда следует утверждение леммы. □

Обозначим a_0 — алгоритм, не ошибающийся на генеральной выборке. Для применения мажорирующих нужно ввести вероятностную меру π на множестве алгоритмов $A \cup a_0$. Она в нашем случае и будет называться *мажорирующей*. Отметим, что она никак не связана с тем, с какой вероятностью алгоритмы выбираются при обучении, оценка так и остаётся равномерной по семейству алгоритмов. Обозначим $B(a,\varepsilon)$ — множество всех алгоритмов семейства $A \cup a_0$, удаленных от a не более чем на ε по метрике d. Пусть d – диаметр семейства $A \cup a_0$. Введём понятие энтропии семейства алгоритмов

$$\mathbf{Q}(A) = \frac{1}{3}\ln\left(\frac{2}{\min\limits_{a\in A\cup a_0}\pi\left(B(a,\frac{d}{2})\right)}\right) + \frac{4}{3}\sum_{k=2}^{\infty}2^{-k}\ln\left(\frac{2}{\min\limits_{a\in A\cup a_0}\pi\left(B(a,\frac{d}{2^k})\right)}\right).$$

При таким определении энтропии оценки максимума интересующего нас процесса можно легко получить с помощью теоремы 4.1 из [12]. Она в свою очередь является обобщением техник из [4].

Теорема 2.3. *Пусть* $\ell = k = \frac{L}{2}$, $\max\limits_{a\in A} m(a) \leq \frac{\ell}{2}$, *тогда*

$$\mathbf{E}\max_{a\in A}\left(\nu(a,\bar{X}) - \nu(a,X)\right) \leq 3\sqrt{\frac{2\mathbf{Q}(A)}{\ell}}.$$

Доказательство. С помощью неравенства Йенсена для всех $\lambda > 0$ имеем

$$\frac{\lambda}{3\sqrt{\ell}}\mathbf{E}\max_{a\in A}(n(a,\bar{X}) - n(a,X)) \leq \ln\left(\mathbf{E}\exp\left(\frac{\lambda}{3\sqrt{\ell}}\max_{a\in A}(n(a,\bar{X}) - n(a,X))\right)\right).$$

Лемма 2.3 и утверждение теоремы 4.1 из [12] влекут

$$\ln\left(\mathbf{E}\exp\left(\frac{\lambda}{3\sqrt{\ell}}\max_{a\in A}(n(a,\bar{X})-n(a,X))\right)\right)\leq\frac{\lambda^2}{2}+\mathbf{Q}(A).$$

Оптимизируя по λ и разделив части полученного неравенства на ℓ, получаем утверждение теоремы. □

Результатом теоремы является тот факт, что ожидаемую переобученность контролирует не только $\ln(|A|)$, но и энтропия семейства. Интересным моментом является то, что в теореме распределение π произвольно. Но поиск оптимальной меры очень сложная задача, решенная только в некоторых частных случаях [13].

Сначала покажем, что полученная оценка не хуже, чем ранее предложенная. Действительно, возьмем в качестве π равномерную меру. Ясно, что $B(a,\varepsilon)\geq 1$. Такая оценка даёт в данном случае

$$\mathbf{Q}(A)\leq\ln\left(2(|A|+1)\right)$$

Таким образом, с точностью до постоянных данная оценка точно не хуже чем оценка теоремы 2.1, так как всегда в качестве меры π можно выбрать раномерную.

Стоит отметить, что можно ввести альтернативное и менее общее понятие энтропии $\hat{\mathbf{Q}}(A)$, свободное от мажорирующей меры и зависящее от $\mathcal{N}(A,\varepsilon)$ — размера ε–покрытий семейства алгоритмов A по метрике d [8]. При этом выводится оценка типа 2.3. В то же время возникает проблема, связанная со сложностью вычисления $\mathcal{N}(A,\varepsilon)$.

С другой стороны $\mathcal{N}(A,\varepsilon)$ могут быть оценены лишь с помощью ёмкости. Согласно [9]

Теорема 2.4 (Haussler). *Для $A(\mathbb{X})$, если $A(\mathcal{X})$ имеет ёмкость V, при $\ell=k$*

$$\mathcal{N}(A,\varepsilon)\leq e(V+1)\left(\frac{4e\ell}{\varepsilon^2}\right)^V.$$

С помощью этой теоремы доказывается, что $\hat{\mathbf{Q}}(A)=O\left(V\right)$ и не зависит от ℓ. Из этого следует, что в случае конечной ёмкости $\mathcal{EOF}_{\max}$ контролируется не $\ln(|A|)$, а ёмкостью. Тем не менее результат оценки 2.3 более общий.

2.4 Доказательство технической леммы

В этом разделе мы докажем лемму 2.1.

Для действительного параметра α многочлены $\{C_n^{(\alpha)}(x)\}_{n=0}^{\infty}$, определенные на отрезке $[-1,1]$, производящая функция которых равна

$$\frac{1}{(1-2xt+t^2)^{\alpha}} = \sum_{n=0}^{\infty} C_n^{(\alpha)}(x)t^n$$

называются *ультрасфрическими.*

В работе нам понадобится следующая рекуррентная формула [2]

$$C_0^{(\alpha)}(x) = 1,$$
$$C_1^{(\alpha)}(x) = 2\alpha x,$$
$$C_m^{(\alpha)}(x) = \frac{1}{m}(2x(m+\alpha-1)C_{m-1}^{(\alpha)}(x) - (m+2\alpha-2)C_{m-2}^{(\alpha)}(x)).$$

Доказательство. Легко видеть, что в данном случае гипергеометрическая функция является многочленом от z, причем для $z \in [0,1]$ согласно [2] имеет место равенство

$${}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2}-\ell, z\right) = (-1)^m \frac{m!}{(\frac{1}{2}-\ell)_m}\left(\frac{z}{4}\right)^{\frac{m}{2}} C_m^{(\frac{1}{2}+\ell-m)}\left(\tfrac{1}{\sqrt{z}}\right),$$

где в правой части значение в нуле определено по непрерывности, а $(x)_m$ – нижний факториал числа x. Далее нас будут интересовать лишь точки экстремумов данной функции, поэтому мы будем работать только с

$$(z)^{\frac{m}{2}} C_m^{(\frac{1}{2}+\ell-m)}\left(\tfrac{1}{\sqrt{z}}\right),$$

так как остальная часть выражения при данных соотношениях на параметры неотрицательна и не зависит от z. Обозначим $x^2 = z, x \in [-1,1]$.

Обозначим также $\hat{C}_m^{\alpha}(x) = x^m C_m^{(\alpha)}(\frac{1}{x})$. Тогда для данного многочлена легко получить рекуррентные соотношения, аналогичные тем, что имеются для ультрасферического

$$\hat{C}_0^{(\alpha)}(x) = 1,$$
$$\hat{C}_1^{(\alpha)}(x) = 2\alpha,$$
$$\hat{C}_m^{(\alpha)}(x) = \frac{1}{m}(2(m+\alpha-1)\hat{C}_{m-1}^{(\alpha)}(x) - x^2(m+2\alpha-2)\hat{C}_{m-2}^{(\alpha)}(x)).$$

Докажем по индукции, что если $\alpha \geq \frac{1}{2}$, то на всем отрезке $[-1,1]$ имеет место неравенство $\hat{C}_m^{(\alpha)}(x) \geq \hat{C}_{m-1}^{(\alpha)}(x)$.

База индукции очевидна. Для $m \geq 3$ рассмотрим разность

$$\hat{C}_m^{(\alpha)}(x) - \hat{C}_{m-1}^{(\alpha)}(x) =$$
$$\frac{1}{m}(2(m+\alpha-1)\hat{C}_{m-1}^{(\alpha)}(x) - x^2(m+2\alpha-2)\hat{C}_{m-2}^{(\alpha)}(x) - m\hat{C}_{m-1}^{(\alpha)}(x)) =$$
$$\frac{1}{m}((m+2\alpha-2)(\hat{C}_{m-1}^{(\alpha)} - x^2\hat{C}_{m-2}^{(\alpha)})).$$

Но $(m+2\alpha-2) \geq 0$, а по предположению индукции $\hat{C}_{m-1}^{(\alpha)} - \hat{C}_{m-2}^{(\alpha)} \geq 0$, но так как по индукции $\hat{C}_{m-2}^{(\alpha)} \geq 0$ и $x^2 \leq 1$, то $\hat{C}_{m-1}^{(\alpha)} - x^2\hat{C}_{m-2}^{(\alpha)} \geq 0$.

Рассмотрим теперь производные $\hat{C}_m^{\alpha}(x)$. Имеет место рекуррентное соотношение

$$\hat{C'}_0^{(\alpha)}(x) = 0,$$
$$\hat{C'}_1^{(\alpha)}(x) = 0,$$
$$\hat{C'}_m^{(\alpha)}(x) = \frac{1}{m}(2(m+\alpha-1)\hat{C'}_{m-1}^{(\alpha)}(x) - x^2(m+2\alpha-2)\hat{C'}_{m-2}^{(\alpha)}(x) -$$
$$2x(m+2\alpha-2)\hat{C}_{m-2}^{(\alpha)}(x)).$$

Из формы рекуррентных соотношений легко видеть, что $\hat{C}_m^{(\alpha)}(x)$ – чётная функция. Поэтому удобно производить анализ производных только на $[0,1]$.

Докажем по индукции, что если $\alpha \geq \frac{1}{2}$, то на всем отрезке $[0,1]$ имеет место неравенство $\hat{C'}_m^{(\alpha)}(x) \leq \hat{C'}_{m-1}^{(\alpha)}(x)$.

База индукции опять же очевидна. Аналогично предыдущему случаю:

$$\hat{C'}_m^{(\alpha)}(x) - \hat{C'}_{m-1}^{(\alpha)}(x) =$$
$$\frac{1}{m}((m+2\alpha-2)(\hat{C'}_{m-1}^{(\alpha)} - x^2\hat{C'}_{m-2}^{(\alpha)} - 2x\hat{C}_{m-2}^{(\alpha)}(x))).$$

Действительно, $(m+2\alpha-2) \geq 0$, $\hat{C'}_{m-1}^{(\alpha)} - x^2\hat{C'}_{m-2}^{(\alpha)} \leq 0$, так как $\hat{C'}_{m-1}^{(\alpha)} - \hat{C'}_{m-2}^{(\alpha)} \leq 0$, $\hat{C'}_{m-2}^{(\alpha)} \leq 0$ и $x^2 \leq 1$. Также по ранее доказанному $2x\hat{C}_{m-2}^{(\alpha)}(x) \geq 0$ на $[0,1]$.

Таким образом, с учётом чётности $\hat{C}_m^{(\alpha)}(x)$ на $[-1,1]$ не превосходит своего значения в нуле. Это соответствует тому, что на $z \in [0,1]$

$$_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2} - \ell, z\right)$$

ограничено своим значением в $z = 0$, то есть единицей.

□

3 Учёт структуры семейства алгоритмов

Оценки, полученные в предыдущем разделе, очень мало использовали структуру семейства алгоритмов. В комбинаторном подходе лучшие результаты получались именно за счёт учёта метрической структуры семейства алгоритмов. В данном разделе для предложенных оценок с помощью техник комбинаторного подхода удастся учесть метод обучения. Кроме этого, будут предложены результаты, касающиеся эффектов завышенности оценок в рамках комбинаторного подхода.

3.1 Оценка переобученности

Следующая теорема даёт оценку ожидаемой переобученности, учитывающую структуру графа расслоения–связности.

Теорема 3.1. *Пусть метод обучения μ – ПМЭР, $\ell = k = \frac{L}{2}$ и $\max\limits_{a\in A} m(a) \leq \frac{L}{2}$, тогда*

$$\mathcal{EOF}_\mu \leq$$
$$\min_{\lambda>0} \frac{1}{\lambda}\left(\ln(\cosh(\lambda)) - \frac{\ln(C_{2\ell}^{\ell})}{\ell} + \frac{1}{\ell}\ln\left(\sum_{a\in A}\phi(\ell, m(a), q(a), u(a), \lambda)\right)\right) \leq$$
$$\sqrt{\frac{2\ln(|A|)}{l}},$$

где Δ_s — число алгоритмов в s-ом слое семейства алгоритмов, а

$$\phi(\ell, m, q, u, \lambda) = \sum_{j=0}^{m-q}\sum_{i=0}^{j}(-1)^i C_j^i C_{2\ell-j-u-q}^{\ell-q+i-j} C_{m-q}^j (1+\tanh(\lambda))^q (\tanh(\lambda))^j.$$

Доказательство. Доказательство до определённого шага полностью повторяет шаги теоремы 2.1. Отличие же заключается на шаге использования неравенства Буля. За счёт учёта метода обучения этот шаг несколько уточняется. Для вектора σ подвыборка X получается выбором из $\mathbb{X}$ всех объектов, позициям которых в σ соответствуют единицы.

$$\lambda\mathbf{E}\left(\sum_{i=1}^{L}\sigma_i I(\mu X, x_i)\right) \leq$$
$$\ln\left(\mathbf{E}\left(\sum_{a\in A}[\mu X = a](\cosh(\lambda))^\ell \prod_{i=1}^{m(a)}(\hat{x}_i\tanh(\lambda)+1)\right)\right).$$

Поработаем отдельно с выражением

$$\mathbf{E}\left(\sum_{a\in A}[\mu X = a]\prod_{i=1}^{m(a)}(\hat{x}_i\tanh(\lambda)+1)\right)$$

Благодаря лемме 1.1 оно мажорируется выражением

$$\mathbf{E}\left(\sum_{a\in A}[X_a\subset X][X'_a\subset \bar{X}]\prod_{i=1}^{m(a)}(\hat{x}_i\tanh(\lambda)+1)\right)$$

Последний переход очень важен. Как и в случае равномерных оценок на этом шаге может накапливаться максимальная неточность. В следующем разделе мы дополнительно исследуем, когда данный переход является именно равенством.

Каждый алгоритм a ошибается на всех объектах X'_a. Без ограничения общности можно считать, что X'_a соответствуют последние $q(a)$ объектов. Для них в условиях $X'_a\subset\bar{X}$ соответствующие $\hat{x}_i$ тождественно равны единице. Учитывая это и раскрывая скобки, имеем

$$\mathbf{E}\left(\sum_{a\in A}[X_a\subset X][X'_a\subset \bar{X}](\tanh(\lambda)+1)^{q(a)}\prod_{i=1}^{m(a)-q(a)}(\hat{x}_i\tanh(\lambda)+1)\right)=$$

$$\mathbf{E}\Bigg(\sum_{a\in A}[X_a\subset X][X'_a\subset \bar{X}](\tanh(\lambda)+1)^{q(a)}(1+C^1_{m(a)-q(a)}\hat{x}_1\tanh(\lambda)+\ldots+$$

$$\hat{x}_1\ldots\hat{x}_{m(a)-q(a)}(\tanh(\lambda))^{m(a)-q(a)})\Bigg).$$

Теперь нужно проанализировать для каждого алгоритма a и $j\le m(a)-q(a)$

$$\mathbf{E}\left([X_a\subset X][X'_a\subset\bar{X}]\;\hat{x}_1\ldots\hat{x}_j\right).$$

Простые комбинаторные рассуждения, аналогичные приводимым ранее, приводят к выражению

$$\mathbf{E}\left([X_a\subset X][X'_a\subset\bar{X}]\;\hat{x}_1\ldots\hat{x}_j\right)=\frac{\sum\limits_{i=0}^{j}(-1)^iC^i_jC^{\ell-q+i-j}_{2\ell-j-u-q}}{C^\ell_{2\ell}}.$$

Таким образом,

$$\mathbf{E}\left(\sum_{a\in A}[X_a\subset X][X'_a\subset \bar{X}]\prod_{i=1}^{m(a)}(\hat{x}_i\tanh(\lambda)+1)\right)=$$

$$\frac{\sum\limits_{a\in A}(1+\tanh(\lambda))^q\sum\limits_{j=0}^{m-q}\sum\limits_{i=0}^{j}(-1)^iC^i_jC^{\ell-q+i-j}_{2\ell-j-u-q}C^j_{m-q}\tanh^j(\lambda)}{C^\ell_{2\ell}}.$$

Подставляя в ранее выписанные выражения полученную формулу, получаем первое неравенство теоремы. Теперь на основании того, что

$$\mathbf{E}\left(\sum_{a\in A}[X_a \subset X][X'_a \subset \bar{X}]\prod_{i=1}^{m(a)}(\hat{x}_i \tanh(\lambda)+1)\right) \leq \mathbf{E}\left(\sum_{a\in A}\prod_{i=1}^{m(a)}(\hat{x}_i \tanh(\lambda)+1)\right)$$

Получаем и второе неравенство. □

Полученная оценка вычислима по матрице ошибок семейства алгоритмов A и учитывает комбинаторные свойства этого семейства. Можно легко показать, что если обнулить в оценке все u и q, то получится та же оценка, что и в теореме 2.1. Другим важным моментом является то, что, как и в случае теоремы 2.1, здесь мы оцениваем производящую функцию моментов. Поэтому, используя неравенство Маркова и оптимизируя по параметру λ, можно получить практически точный аналог оценки расслоения–связности 1.1.

3.2 Факторы завышенности

Уже было упомянуто, что основными факторами завышенности приведённых оценок является именно использование неравенства Буля. Можно неформально считать, что все наши техники рассчитаны именно на оптимизацию использования этого неравенства. Например, предложенные оценки, использующие аппарат мажорирующих мер, лучше, чем простые оценки только за счёт оптимального использования неравенства Буля. Комбинаторный подход предлагает технику, основанную на учёте метода обучения и таким образом усреднению вклада алгоритма только по тем разбиениям, где алгоритм может реализоваться. Мы уже обращали внимание на шаг теоремы 3.1, при котором осуществлялся переход к сумме по всем алгоритмам семейства.

Напомним, там осуществлялась замена

$$\sum_{a\in A}[\mu X = a]\prod_{i=1}^{m(a)}(\hat{x}_i \tanh(\lambda)+1) \text{ на } \sum_{a\in A}[X_a \subset X][X'_a \subset \bar{X}]\prod_{i=1}^{m(a)}(\hat{x}_i \tanh(\lambda)+1).$$

С учётом того, что

$$\prod_{i=1}^{m(a)}(\hat{x}_i \tanh(\lambda)+1) > 0$$

получаем что оба выражения равны между собой тогда и только тогда, когда для всех разбиений и алгоритмов семейства

$$[\mu X = a] = [X_a \subset X][X'_a \subset \bar{X}].$$

Подобный эффект наблюдается для некоторых модельных семейств алгоритмов, о которых подробно написано в [20, 21], но в данной работе будет представлено необходимое и достаточное условие на граф расслоения–связности, при выполнении которого выполнено данное равенство.

Теорема 3.2. *Пусть μ — метод пессимистичной минимизации эмпирического риска, $\ell \geq 2 \max\limits_{a \in A} u(a)$ и $k \geq 2 \max\limits_{a \in A} q(a)$. Условие*

$$\big[\mu X = a\big] = \big[X_a \subset X\big]\big[X'_a \subset \bar{X}\big]$$

выполнено тогда и только тогда, когда для любой пары a, b различных алгоритмов из A выполнено

$$\big(X_a \cap X'_b \neq \emptyset\big) \text{ или } \big(X'_a \cap X_b \neq \emptyset\big).$$

Доказательство. Достаточность.

Рассмотрим разбиение генеральной выборки на обучение и контроль $\mathbb{X} = X \sqcup \bar{X}$. Пусть $a = \mu X$, по лемме 1.1 $X_a \subset X, X'_a \subset \bar{X}$. Теперь покажем, что для любого алгоритма b отличного от a имеет место равенство $\big[X_b \subset X\big]\big[X'_b \subset \bar{X}\big] = 0$. Действительно, предположим противное, тогда $X_b \subset X$ и $X'_b \subset \bar{X}$. Так как $X \cap \bar{X} = \emptyset$, то $X_a \cap X'_b = \emptyset$ и $X'_a \cap X_b = \emptyset$.

С другой стороны из условия теоремы хотя бы одно из множеств $X_a \cap X'_b$ и $X'_a \cap X_b$ непусто. Полученное противоречие доказывает, что $\big[X_b \subset X\big]\big[X'_b \subset \bar{X}\big] = 0$. Ввиду произвольности разбиения и выбора алгоритма b заключаем, что для всех b выполнено равенство:

$$\big[\mu X = b\big] \geq \big[X_b \subset X\big]\big[X'_b \subset \bar{X}\big].$$

Вместе с леммой 1.1 данное неравенство доказывает достаточность условия теоремы.

Необходимость. Пусть условие теоремы не выполнено, тогда существует пара различных алгоритмов a и b для которой $X_a \cap X'_b = \emptyset$ и $X'_a \cap X_b = \emptyset$, поэтому

$(X_a \cup X_b) \cap (X'_a \cup X'_b) = \emptyset$, причём из условия теоремы $|X_a \cup X_b| \leq \ell, |X'_a \cup X'_b| \leq k$. Поэтому существует разбиение генеральной выборки на обучение X и контроль $\bar{X}$ такое, что $X_a \cup X_b \subset X, X'_a \cup X'_b \subset \bar{X}$. А значит

$$\big[X_a \subset X\big]\big[X'_a \subset \bar{X}\big] = \big[X_b \subset X\big]\big[X'_b \subset \bar{X}\big] = 1.$$

Полученное противоречие доказывает необходимость условия теоремы.

□

Если не выполнено одно из условий $\ell \geq 2 \max\limits_{a \in A} u(a)$ или $k \geq 2 \max\limits_{a \in A} q(a)$, то теорема даёт только достаточное условие точности оценки.

Для заданного семейства алгоритмов условия теоремы можно проверить за квадратичное по числу алгоритмов в семействе время. Однако интересно понять метрические характеристики таких семейств алгоритмов. Проще и нагляднее всего это сделать в терминах графа расслоения–связности.

3.3 Критерий точности оценок.

Основной результат данного раздела изложен в работе [1].

Рассмотрим семейство алгоритмов A, обладающее следующими свойствами. Пусть на m_0 объектах ошибаются все алгоритмы семейства, на m_1 объектах не ошибается ни один алгоритм семейства, а на оставшихся $m = L - m_0 - m_1$ объектах реализуются все возможные варианты ошибиться. Ясно, что $|A| = 2^m$. Данное семейство алгоритмов будем называть *интервалом булева куба*.

В A существует единственный алгоритм a с $d(a) = 0$ и единственный алгоритм b с $u(b) = 0$, причём $\forall c \in A$ выполнено $a \leq c \leq b$.

Назовём алгоритм a с $d(a) = 0$ *истоком* интервала булева куба. Заметим, что задание истока и m объектов, на которых реализуются различные варианты ошибок алгоритмов семейства, полностью задаёт весь интервал булева куба. Поэтому будем говорить, что m заданных объектов *порождают* данное семейство. Если представить себе граф расслоения–связности интервала булева куба, то порождающие элементы соответствуют рёбрам, исходящим из истока.

Теорема 3.3. *Пусть* $\ell \geq \max\limits_{a \in A} 2u(a)$ *и* $k \geq \max\limits_{a \in A} 2q(a)$. *Условие*

$$\big[\mu X = a\big] = \big[X_a \subset X\big]\big[X'_a \subset \bar{X}\big]$$

выполнено тогда и только тогда, когда для A выполнены следующие условия

1. *В графе расслоения–связности семейства A существует только один сток и только один исток.*

2. *Для всех алгоритмов a семейство A содержит интервал булева куба с истоком a, порождённый с помощью элементов X_a.*

Доказателъство. Необходимость. Далее будем считать, что семейство алгоритмов A имеет более одного алгоритма, так как случай с одним алгоритмом тривиален. Пусть первое условие теоремы не выполнено. Заметим, что $|\{a \in A|\ u(a) = 0\}| > 0$ и $|\{b \in A|\ d(b) = 0\}| > 0$, так как нулевой верхней связностью обладают все алгоритмы верхнего слоя, аналогично нижняя связность нижнего слоя также равна нулю. Пусть теперь пара различных алгоритмов a_1, a_2 таких что $u(a_1) = u(a_2) = 0$, тогда $X_{a_1} \cap X'_{a_2} = \emptyset$ и $X'_{a_1} \cap X_{a_2} = \emptyset$, так как $X_{a_1} = X_{a_2} = \emptyset$, поэтому условие теоремы 3.2 не выполнено, значит существует всего один алгоритм с нулевой верхней связностью. Несколько сложнее рассматривается случай нижней связности. Пусть теперь существует пара различных алгоритмов b_1, b_2 с $d(b_1) = d(b_2) = 0$. Без ограничения общности будем считать, что $q(b_1) = 0$. Действительно, в любом семействе алгоритмов наименьший по числу ошибок алгоритм имеет как нулевую нижнюю связность, так и нулевую неоптимальность. Докажем теперь, что и $q(b_2) = 0$. Предположим противное, тогда найдётся алгоритм b_3, такой что $b_3 < b_2$. Алгоритм b_3 обладает ненулевой верхней связностью, так как он не является алгоритмом верхнего слоя. Покажем, что из b_3 исходит ребро, которому соответствует объект $x \in X'_{b_2} \setminus X'_{b_3}$. Действительно, иначе $X_{b_3} \cap X'_{b_2} = \emptyset$ и $X'_{b_3} \cap X_{b_2} = \emptyset$, так как $X'_{b_3} \subset X'_{b_2}, X'_{b_2} \cap X_{b_2} = \emptyset$. Получаем противоречие с требованием теоремы 3.2. Теперь рассматриваем алгоритм b_4, в который входит ребро соответствующее объекту из $X'_{b_2} \setminus X'_{b_3}$ и $b_4 \prec b_3$. Если он совпадает с алгоритмом b_2, то получаем противоречие с тем, что $d(b_2) = 0$. Иначе, повторяем аналогичные рассуждения с алгоритмом b_4, перейдём к алгоритму следующего слоя до тех пор пока не исчерпаем все объекты $X'_{b_2} \setminus X'_{b_3}$ и не получим на очередном шаге алгоритм b_2. Таким образом, получим, что в b_2 входит ребро, а значит $d(b_2) \neq 0$, получаем противоречие, которое говорит о том, что $q(b_2) = 0$. Аналогично случаю верхней связности (так как $X_{b_1} \cap X'_{b_2} = \emptyset$ и $X_{b_2} \cap X'_{b_1} = \emptyset$) получаем

противоречие с требованием теоремы 3.2, а значит существует всего один алгоритм с нулевой нижней связностью.

Докажем теперь необходимость второй части утверждения. Ясно, что при выполнении первого условия теоремы, каждый алгоритм лежит b в какой либо цепи, то есть наборы вида $a_1, a_2, \ldots, b, \ldots, a_m$, причём $a_1 \prec a_2 \prec \ldots \prec b \prec \ldots \prec a_m$, где a_1 и a_m соответственно исток и сток графа расслоения–связности семейства алгоритмов. Действительно, фиксируем любой алгоритм b, переходим по нему в любой алгоритм, которому он предшествует, затем из него переходим в следующий слой. Так как алгоритм с нулевой верхней связностью единственный, то именно в него мы и попадаем в конце при таких переходах. Аналогично, дополняем полученную часть цепи алгоритмами, которые предшествуют алгоритму b.

Пусть второе условие теоремы не выполнено, тогда найдётся b, такой что A не содержит интервала булева куба с истоком b, порождённого с помощью X_b. Пусть $X_b = \{x_1, \ldots, x_n\}$, тогда последнее условие эквивалентно тому, что существует подстановка $\sigma \in S_n$, такая что цепь $b = b_0 \prec b_1 \prec \ldots \prec b_n$, для которой $\forall i \in \{1, \ldots, n\}$ выполнено $I(b_{i-1}, x_{\sigma(i)}) < I(b_i, x_{\sigma(i)})$, не содержится в семействе A, где действие группы S_n определено на $\{1, \ldots, n\}$ естественным образом. Это следует из того, что в интервале булева куба содержатся все возможные цепи, которых ровно $n!$.

Пусть подцепь описанной цепи $\{b, b_1, \ldots, b_p\}, p < n$ принадлежит A, а алгоритм $b_{p+1} \notin A$, значит $x_{\sigma(p+1)} \notin X_{b_p}$. Среди алгоритмов $\{b, b_1, \ldots, b_{p-1}\}$ выберем наибольший по числу ошибок алгоритм b_j такой, что $x_{\sigma(p+1)} \in X_{b_j}$. Он существует, потому что $x_{\sigma(p+1)} \in \{x_1, \ldots, x_n\} = X_b$.

Рассмотрим теперь пару алгоритмов b_{j+1} и a, где a обладает тем свойством, что $b_j \prec a, I(b_j, x_{\sigma(p+1)}) < I(a, x_{\sigma(p+1)})$. Такой алгоритм существует, так как $x_{\sigma(p+1)} \in X_{b_j}$. Заметим, что $X'_{b_{j+1}}$ и X'_a отличаются лишь тем, что первое множество содержит $x_{\sigma(j+1)}$, а второе вместо него содержит $x_{\sigma(p+1)}$. Значит, для выполнения хотя бы одного из условий $X_{b_{j+1}} \cap X'_a \neq \emptyset$ или $X'_{b_{j+1}} \cap X_a \neq \emptyset$ необходимо, чтобы $x_{\sigma(j+1)} \in X_a$ или эквивалентное условие $x_{\sigma(p+1)} \in X_{b_{j+1}}$. Последнее возможно лишь если $b_{j+1} = b_p$, так как алгоритм b_j выбирался наибольшим по числу ошибок среди $\{b, b_1, \ldots, b_{p-1}\}$ и $b_j \prec b_{j+1}$. Но $x_{\sigma(p+1)} \notin X_{b_p}$. Полученное противоречие доказывает необходимость теоремы.

□

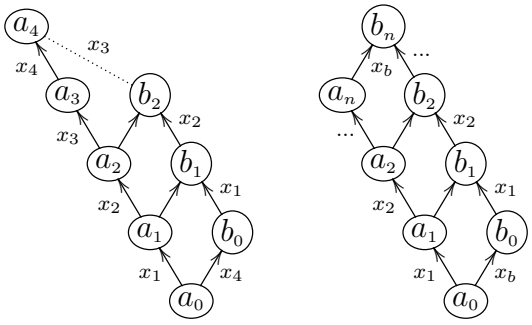
Перед доказательством достаточности докажем следующую простую лемму.

Лемма 3.1. *Пусть семейство алгоритмов таково, что для всех алгоритмов a семейство A содержит интервал булева куба с истоком a, порождённый с помощью элементов X_a. Пусть $a_0 \prec a_1 \prec \ldots \prec a_n$ некоторая цепь в A. И пусть для некоторого b_0 алгоритм $a_0 \prec b_0$. Пусть $x_b = X'_{b_0} \setminus X'_{a_0}$. Тогда*

- *Если $x_b \in X'_{a_n}$, то в A существует цепь из b_0 в a_n.*
- *Если $x_b \notin X'_{a_n}$, то $x_b \in X_{a_n}$.*

Доказательство. Обозначим $x_i = X'_{a_i} \setminus X'_{a_{i-1}}$. Пусть сначала для некоторого $p \in \{1, \ldots, n\}$ выполнено $x_b = x_p$ (это проиллюстрировано ниже для $p = 4$). Заметим, что x_1 и x_b принадлежат X_{a_0}, таким образом, в A есть алгоритм b_1, такой что $b_0 \prec b_1$ и $a_1 \prec b_1$. Повторяя это рассуждение, получаем алгоритмы $b_2, \ldots, b_{p-2}$. Очевидно, что $b_{p-2} \prec a_p$, а значит есть цепь $b_0 \prec b_1 \prec \ldots b_{p-2} \prec a_p \prec \ldots \prec a_n$.

Случай, когда $x_b \notin X'_{a_n}$ доказывается таким же способом как и первый, что проиллюстрировано ниже.



□

Доказательство. Достаточность.

Пусть условия теоремы выполнены. Как было показано ранее, из первого условия теоремы следует, что каждый алгоритм лежит хотя бы в одной цепи, начинающийся в истоке и заканчивающейся в стоке графа расслоения–связности. Зафиксируем теперь произвольную пару различных алгоритмов a, b.

Пусть выполнено $a < b$ или $b < a$, то есть пара алгоритмов находится в одной и той же цепи, тогда сразу же выполнено $X_a \cap X'_b \neq \emptyset$ или $X'_a \cap X_b \neq \emptyset$.

Иначе, если ни один алгоритм не предшествует другому, то рассмотрим наибольший по числу ошибок алгоритм c, такой что $c \leq a, c \leq b$. Заметим, что $X'_c \subset X'_a$ и $X'_c \subset X'_b$. Обозначим $\{x_1^a, \ldots, x_n^a\} = X'_a \setminus X'_{c_{ab}}$ и $\{x_1^b, \ldots, x_p^b\} = X'_b \setminus X'_{c_{ab}}$. Будем считать, что $n \leq p$, а также что $x_1^a, \ldots, x_n^a$ и $x_1^b, \ldots, x_p^b$ упорядочены в том порядке, в котором алгоритмы в цепи допускают ошибки на этих объектах.

Пусть t – наименьший индекс для которого $x_t^b \notin \{x_1^a, \ldots, x_n^a\}$. Такой элемент существует, так как $n \leq p$. В случае $n = p$ несуществование s влечёт равенство $a = b$.

Рассмотрим объекты $\{x_1^b, \ldots, x_{t-1}^b\}$. Число t выбрано таким образом, что все эти объекты есть среди $\{x_1^a, \ldots, x_n^a\}$. Применяя лемму 3.1 $t-1$ раз получаем, что в A есть цепь из b_{t-1} в a. Применяя теперь лемму 3.1 к данной цепи и алгоритму b_t получаем, что $x_t^b \in X_a$. Таким образом, $X_a \cap X'_b \neq \emptyset$, значит на основании теоремы 3.2 утверждение доказано.

□

3.4 Другие способы уточнить равномерную оценку

Зная граф расслоения–связности, можно попробовать уточнить равномерную по семейству алгоритмов оценку ожидаемой переобученности. В [19, 20, 21] предложен альтернативный подход, заключающийся во введении метода обучения, максимизирующего переобученность, другими словами метода обучения, который каждому разбиению $\mathbb{X} = X \sqcup \bar{X}$ ставит в соответствие алгоритм a, который максимизирует $\nu(a, \bar{X}) - \nu(a, X)$. Ожидаемая переобученность для такого метода в точности совпадает с равномерной по семейству ожидаемой переобученностью. Метод обучения, максимизрующий переобученность, также может быть охарактеризован порождающим и запрещающими множествами, однако запрещающие множества в данном случае включают в себя только объекты, соответствующие выходящим из алгоритма рёбрам.

В соответствие с замечаниями

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

Элементарно доказывается следующая лемма, аналогичная лемме 1.1

Лемма 3.2. *Пусть μ — метод обучения, максимизирующий переобученность, тогда для всех $a \in A$ выполнено следующее неравенство:*

$$\left[\mu X = a\right] \leq \left[X_a \subset X\right]\left[\hat{X}'_a \subset \bar{X}\right]$$

Заметим теперь, что $|\hat{X}'_a| = d(a)$. Поэтому результат теоремы 3.1 полностью переносится на случай равномерной оценки с заменой $q(a)$ на $d(a)$.

Теорема 3.4. *Пусть $\ell = k = \frac{L}{2}$ и $\max\limits_{a\in A} m(a) \leq \frac{L}{2}$, тогда*

$$\mathcal{EOF}_{\max} \leq$$
$$\min_{\lambda>0} \frac{1}{\lambda}\left(\ln(\cosh(\lambda)) - \frac{\ln(C^{\ell}_{2\ell})}{\ell} + \frac{1}{\ell}\ln\left(\sum_{a\in A}\phi(\ell, m(a), d(a), u(a), \lambda)\right)\right) \leq$$
$$\sqrt{\frac{2\ln(|A|)}{l}},$$

где Δ_s — число алгоритмов в s-ом слое семейства алгоритмов, а

$$\phi(\ell, m, d, u, \lambda) = \sum_{j=0}^{m-d}\sum_{i=0}^{j}(-1)^i C^i_j C^{\ell-d+i-j}_{2\ell-j-u-d} C^j_{m-d}(1+\tanh(\lambda))^d(\tanh(\lambda))^j.$$

У данной оценки мало преимуществ по сравнению с оценкой 2.3. Для объяснения этого эффекта нам понадобится следующая интерпретация известной теоремы

Теорема 3.5 (Haussler [9])**.** *Для $A(\mathbb{X})$, если $A(\mathcal{X})$ имеет ёмкость V*

$$\frac{\sum\limits_{a\in A}(u(a)+d(a))}{|A|} < 2V$$

Другими словами, накладывая естественное предположение о конечной ёмкости, мы получаем, что в средняя по A сумма размеров X_a и $\hat{X}'_a$ ограничена величиной, не зависящей от ℓ. Дополнительно, по индукции можно доказать, что

$$\min_{a\in A}(u(a)+d(a)) \leq V.$$

Напомним, что по лемме 3.2 размеры этих множеств ограничивают долю разбиений, на которых учитывается вклад алгоритма. В случае, когда ℓ достаточно велико, оценка фактически не отличается от 2.1.

Другой серьёзной проблемой является то, что равномерной оценке 3.4 требуются информация о всём графе расслоения–связности, в то время как энтропийные оценки могут быть оценены сверху и без обращения к полной структуре графа.

4 Концентрация меры в комбинаторном подходе

Основным техническим инструментом статистической теории обучения являются неравенства концентрации меры. И в комбинаторном подходе для получения оценок обобщающей способности, вычислимым только по наблюдаемой выборки нужны неравенства концентрации, но в специальный форме. Действительно, самые мощные неравенства существенно связаны с продакт-пространствами [15, 16, 11]. Для нас это означает, что эти неравенства существенно используют тот факт, что объекты обучающей выборки получены независимо из одного распределения. Мы же имеем дело с выборками без возвращений. Поэтому первым этапом будет обобщение варианта изопериметрического неравенства Талаграна.

Нас будет интересовать это неравенство на слоях $\{1,-1\}^L$ куба. Для полного куба теоремы была доказана в [14]. Затем в [16] этот подход был обобщён на произвольные продакт-пространства. Доказательство в нашем случае становится очень техническим и напрямую технику из оригинальной статьи [14] перенести достаточно сложно. Однако в [17] предложена несколько более удобная техника, которую мы и воспроизведём в нашем случае.

Заметим главную особенность неравенства Талаграна. За счет перехода к евклидовой норме (в отличие от случая, когда работают именно с Хэмминговым расстоянием) в [14] получить изопериметрическое неравенство, независящее от размерности пространства. Оказывается, что и в случае слоя дискретного куба можно получить такой же результат.

4.1 Концентрация на слоях дискретного куба.

Для множества $A \subset \mathbb{R}^L$ обозначим A^t множество всех точек $x \in \mathbb{R}^L$ таких, что $d(x,A) \leq t$. Причём, d — евклидова метрика.

Теорема 4.1. *Пусть A выпуклое и замкнутое множество в $\mathbb{R}^L$. Тогда для всякого целого l, такого что $0 \leq l \leq L$, имеет место неравенство:*

$$\mathsf{P}(X \in A)\,\mathsf{P}(X \notin A^t) \leq \exp\left(\frac{-t^2}{16}\right),$$

для всех $t > 0$ и вектора X, распределённого равномерно на слое дискретного куба $\{-1,1\}^L$, содержащего в точности l экземпляров -1.

Доказательство. Временно обозначим $c = \frac{1}{16}$. Ясно, что для доказательства теоремы достаточно показать, что в условиях теоремы имеет место неравенство

$$P(X \in A)\mathbf{E}\exp(cd(X,A)^2) \leq 1.$$

Действительно, с помощью неравенства Маркова мы имеем

$$\mathsf{P}(X \notin A^t) = \mathsf{P}(\exp(cd(X,A)^2) > \exp(ct^2)) \leq \frac{\mathbf{E}(\exp(cd(X,A)^2))}{\exp(ct^2)}.$$

Заменяя математическое ожидание на полученное в последнем неравенстве, получаем нужное утверждение.

Доказательство будем вести индукцией по L. В частности, база индукции очевидна, так же как и случаи $l = L$ или $l = 0$, где левая часть доказываемого в условии теоремы неравенства тождественно равна нулю. Поэтому везде далее считаем, что $l \neq 0$ и $l \neq L$.

Обозначим $X = (X', x_L)$, где $x_L = \pm 1$. Также обозначим сечение

$$A_t = \{x' \in \mathbb{R}^{L-1} : (x', t) \in A\}.$$

В случае, если A выпуклое, то и сечение A_t выпуклое.

Обозначим Y_t ближайшую в A_t точку к X'. Выберем некоторое $\lambda \in [0,1]$. Из выпуклости точка $(1-\lambda)(Y_{x_L}, x_L) + \lambda(Y_{-x_L}, -x_L)$ лежит в A. Поэтому

$$d(X,A) \leq |(1-\lambda)(Y_{x_L}, x_L) + \lambda(Y_{-x_L}, -x_L) - (X', x_L)|.$$

Сгруппируем слагаемые в правой части последнего равенства

$$|(1-\lambda)(Y_{x_L} - X', 0) + \lambda(Y_{-x_L} - X', 0) - 2\lambda(0, x_L)|.$$

Используя свойства скалярного произведения, получаем

$$d(X,A)^2 \leq 4\lambda^2 + |(1-\lambda)(X' - Y_{x_L}) + \lambda(X' - Y_{-x_L})|^2$$

Теперь воспользуемся выпуклостью фукнции x^2 и получим ограничение

$$|(1-\lambda)(X' - Y_{x_L}) + \lambda(X' - Y_{-x_L})|^2 \leq (1-\lambda)^2(X' - Y_{x_L}) + \lambda^2(X' - Y_{-x_L}).$$

В итоге получаем неравенство

$$d(X,A)^2 \leq 4\lambda^2 + (1-\lambda)d(X', A_{x_L})^2 + \lambda d(X', A_{-x_L})^2.$$

Полученное неравенство будет использовано в дальнейшем. Используем формулу полной вероятности и запишем

$$\mathsf{P}(X \in A) = \mathsf{P}(X' \in A_1 | x_L = 1)\frac{L-l}{L} + \mathsf{P}(X' \in A_{-1} | x_L = -1)\frac{l}{L}.$$

Рассмотрим сначала случай, когда

$$\mathsf{P}(X' \in A_1 | x_L = 1)(L-l) \geq \mathsf{P}(X' \in A_{-1} | x_L = -1)l.$$

Обозначим $p = \mathsf{P}(X \in A)$, тогда для некоторого $q \in [0,1]$ можно выписать

$$\mathsf{P}(X' \in A_1 | x_L = 1) = \frac{p(1+q)}{2}\frac{L}{L-l},$$
$$\mathsf{P}(X' \in A_{-1} | x_L = -1) = \frac{p(1-q)}{2}\frac{L}{l}.$$

Зафиксируем значение x_L и рассмотрим условное математическое ожидание

$$\mathbf{E}_{X'} \exp(cd(X,A)^2) \leq \exp(4c\lambda^2)\mathbf{E}_{X'} \left(\exp(cd(X', A_{x_L})^2)\right)^{1-\lambda} \left(\exp(cd(X', A_{-x_L})^2)\right)^{\lambda}.$$

Воспользуемся утверждением индукции и неравенством Гёльдера для матожидания

$$\mathbf{E}_{X'} \exp(cd(X,A)^2) \leq$$
$$\exp(4c\lambda^2) \left(\mathbf{E}_{X'} \exp(cd(X', A_{x_L})^2)\right)^{1-\lambda} \left(\mathbf{E}_{X'} \exp(cd(X', A_{-x_L})^2)\right)^{\lambda} \leq$$
$$\leq \exp(4c\lambda^2)\frac{1}{\left(\frac{p(1+x_Lq)}{2}\frac{L}{x_n(\frac{L}{2}-l)+\frac{L}{2}}\right)^{1-\lambda}\left(\frac{p(1-x_Lq)}{2}\frac{L}{-x_n(\frac{L}{2}-l)+\frac{L}{2}}\right)^{\lambda}}.$$

Теперь записываем безусловное математическое ожидание

$$\mathbf{E}_X \exp(cd(X,A)^2) \leq \frac{2\exp(4c\lambda_0^2)l^{\lambda_0}(L-l)^{2-\lambda_0}}{(1+q)^{1-\lambda_0}(1-q)^{\lambda_0}L^2p} + \frac{2\exp(4c\lambda^2)(L-l)^{\lambda}l^{2-\lambda}}{(1-q)^{1-\lambda}(1+q)^{\lambda}L^2p}.$$

Рассмотрим сначала случай $l \geq \frac{L}{2}$. В этом случае сделаем оптимальный выбор $\lambda_0 = 0$. Теперь, с учетом того, что $p = \mathsf{P}(X \in A)$ для доказательства шага индукции нам достаточно показать, что

$$\frac{2(L-l)^2}{L^2(1+q)} + \frac{2\exp(4c\lambda^2)(L-l)^{\lambda}l^{2-\lambda}}{(1-q)^{1-\lambda}(1+q)^{\lambda}L^2} \leq 1,$$

причём значение λ можно выбирать в соответствии со значением q. Теперь примем $c = \frac{1}{16}$. Выбор $\lambda = 1$ дает равномерное по q условие на выполение неравенства, а именно

$$\frac{2(L-l)^2}{L^2} + \frac{2\exp(\frac{1}{4})(L-l)l}{L^2} \leq 1,$$

что дает выполнение заданного условия при всех

$$\frac{l}{L} \geq \frac{\exp(\frac{1}{4}) - 2 + \sqrt{2 - 2\exp(\frac{1}{4}) + \exp(\frac{1}{2})}}{2(\exp(\frac{1}{4}) - 1)} = 0.5696\ldots\ .$$

С учётом того, что $\frac{2(L-l)^2}{L^2} \leq \frac{1}{2}$ и $\frac{2(L-l)l}{L^2} \leq \frac{1}{2}$ получаем, что неравенство выполнено и для всех $q \geq \exp(\frac{1}{4}) - 1 = 0.2840\ldots$. Таким образом, нас интересуют близкие к нулю значения q и близкие к $\frac{L}{2}$ значения l.

Рассмотрим на отрезке $[0, 1]$ функцию

$$\frac{2\exp(4c\lambda^2)(L-l)^\lambda l^{2-\lambda}}{L^2}.$$

С помощью дифференцирования показываем, что ее минимальное значение достигается в $\lambda^* = 2\ln(\frac{l}{L-l})$. Подставляем точку оптимума в раннее выражение при $q = 0$, обозначая при этом $y = \frac{l}{L}$

$$2(1-y)^2 + 2y^2\left(\frac{y}{1-y}\right)^{-\ln\left(\frac{y}{1-y}\right)}.$$

Проанализируем данную функцию на отрезке $[\frac{1}{2}, 1]$. При $y > \frac{e}{1+e}$ с учётом того, что функции $(1-y)^2$ и $\left(\frac{y}{1-y}\right)^{-\ln\left(\frac{y}{1-y}\right)}$ убывают на этом отрезке, подставляя данное значение получаем, что данная функция не превосходит единицу. Для $y < \frac{e}{1+e}$ доказываем разложением с помощью формулы Тейлора.

Таким образом, утверждение доказано и для $q = 0$. Теперь докажем и для оставшихся значений q. Для этого будем искать значения λ в виде $2\ln\left(\frac{y}{1-y}\right) + sq$, где s – константа, равная 1.2. Так как мы рассматриваем лишь значения $0.5 \leq y \leq 0.5696\ldots$ и $q \leq 0.2840\ldots$, то и значения λ не могут быть большими единицы при такой параметризации.

Подставляя это в исследуемое выражение, имеем

$$F(q,y) = \frac{2(1-y)^2}{(1+q)} + \frac{2\exp\left(\frac{1}{4}(2\ln(\frac{y}{1-y}) + 1.2q)^2\right)(1-y)^{2\ln\left(\frac{y}{1-y}\right)+1.2q} y^{2-2\ln\left(\frac{y}{1-y}\right)-1.2q}}{(1-q)^{1-2\ln\left(\frac{y}{1-y}\right)-1.2q}(1+q)^{2\ln\left(\frac{y}{1-y}\right)+1.2q}}.$$

Машинные вычисления показывают, что на указанном множестве значений y и q функция F ограничена единицей, что и доказывает шаг индукции. Доказательство полностью аналогично и симметрично, когда $l < \frac{L}{2}$.

Пусть теперь выполнено

$$\mathsf{P}(X' \in A_1 | x_L = 1)(L-l) < \mathsf{P}(X' \in A_{-1} | x_L = -1)l.$$

Но и в этом случае, нужно лишь заменить в предыдущем случае q на $-q$ и доказательство будет аналогичным и симметричным. Таким образом, теорема доказана. $\square$

4.2 Свойства выпуклых липшицевых функций, заданных на слое куба

Пусть функция f, определенная на $\mathbb{R}^L$ является выпуклой и 1-липшицевой по отношению к стандартной евклидовой метрике. Нас будет интересовать как зачения функции на слое дискретного $\{1,-1\}^L$ куба будут отклонятся от своей медианы и среднего, если на слое введено равномерное распределение.

Теорема 4.2. *Для функции f, определённой выше, для $\mathbf{x}$, равномерно распределённых на слое $\{1,-1\}^L$ куба, для всех $t>0$ имеют место неравенства :*

$$\mathsf{P}\left(f(\mathbf{x})-\mathbf{M}f(\mathbf{x})\geq t\right)\leq 2\exp\left(-\frac{t^2}{16}\right),$$
$$\mathsf{P}\left(|f(\mathbf{x})-\mathbf{M}f(\mathbf{x})|\geq t\right)\leq 4\exp\left(-\frac{t^2}{16}\right).$$

Доказателъство. Зафиксируем число s и рассмотрим множество

$$A(s)=\{\mathbf{x}\in\mathbb{R}^L : f(x)\leq s\}$$

Если $A(s)$ не пусто, то так как f – выпуклая функция, множество является $A(s)$ выпуклым. Рассмотрим множество $A^t(s)$. Оно состоит из всех точек $\mathbb{R}^L$, удалённых от $A(s)$ не более чем на t. Тогда с учётом липшицевости

$$\mathsf{P}(\mathbf{x}\notin A_t(s))\geq\mathsf{P}(f(\mathbf{x})>s+t).$$

Подставляем полученные результаты в неравенство теоремы 4.1.

$$\mathsf{P}(f(\mathbf{x})\leq s)\,\mathsf{P}(f(\mathbf{x})>s+t)\leq\exp\left(-\frac{t^2}{16}\right).$$

Выбираем в качестве s в первом случае медиану f, то есть $\mathbf{M}f(\mathbf{x})$, а во втором случае $\mathbf{M}f(\mathbf{x})-t$ и, используя определение медианы, получаем оба результата. $\square$

Часто будет удобнее работать с математическим ожиданием, поэтому нам понадобится следующая лемма.

Лемма 4.1. *В условиях теоремы 4.2 имеет место неравенство*

$$|\mathbf{E}f(\mathbf{x}) - \mathbf{M}f(\mathbf{x})| \leq 8\sqrt{\pi}.$$

Доказательство. Доказательство получается простой цепочкой неравенств

$$|\mathbf{E}f(\mathbf{x}) - \mathbf{M}f(\mathbf{x})| \leq \mathbf{E}|f(\mathbf{x}) - \mathbf{M}f(\mathbf{x})| =$$
$$\int_0^\infty \mathsf{P}(|f(\mathbf{x}) - \mathbf{M}f(\mathbf{x})| > t)dt \leq 4\int_0^\infty \exp\left(-\frac{t^2}{16}\right) = 8\sqrt{\pi}.$$

□

Как следствие получаем следующую теорему.

Теорема 4.3. *В условиях теоремы 4.2 для $t > 0$ имеет место неравенство*

$$\mathsf{P}(f(\mathbf{x}) \geq \mathbf{E}f(\mathbf{x}) + 8\sqrt{\pi} + t) \leq 2\exp\left(-\frac{t^2}{16}\right).$$

4.3 Применение к оценке ненаблюдаемой частоты ошибок

Теперь можно применить полученные неравенства для нашей основной задачи. Хотя полученные результаты очень общие и позволяют нам работать с произвольными ℓ и k и небинарными функциями потерь, мы применим их лишь к рассматриваемому частному случаю.

Теорема 4.4. *При $\ell = k = \frac{L}{2}$ для $t > 0$ имеет место неравенство*

$$\mathsf{P}\left(\max_{a\in A}(\nu(a,\bar{X}) - \nu(a,X)) \geq \mathcal{EOF}_{\max} + \frac{8\sqrt{\pi}\sigma}{\ell} + t\right) \leq 2\exp\left(-\frac{t^2\ell^2}{16\sigma^2}\right),$$

где $\sigma^2 = \max\limits_{a\in A} m(a)$. В частности, $\sigma \leq \sqrt{2\ell}$.

Доказательство. Сопоставим алгоритму a его L-мерный вектор ошибок $\mathbf{a}$. Рассмотрим на $\mathbb{R}^L$ функцию f такую, что

$$f(\mathbf{x}) = \max_{a\in A}(\mathbf{a}, \mathbf{x}).$$

Данная функция (как максимум выпуклых функций) является выпуклой. Проверим её липшицевость. Для любых двух точек $\mathbf{x}$ и $\mathbf{y}$ с помощью неравенства Коши-Буняковского имеем

$$|f(\mathbf{x}) - f(\mathbf{y})| = |\max_{a\in A}(\mathbf{a}, \mathbf{x}) - \max_{a\in A}(\mathbf{a}, \mathbf{y})| \leq |\max_{a\in A}(\mathbf{a}, \mathbf{x} - \mathbf{y})| \leq \sigma|\mathbf{x} - \mathbf{y}|,$$

где $\sigma^2 = \max\limits_{a\in A}\sum\limits_{i=1}^{L} a_i^2 = \max\limits_{a\in A} m(a)$.

В точках слоя $\{1,-1\}^L$ куба, содержащего ровно ℓ экземпляров -1, функция f в точности равна $\max\limits_{a\in A}(n(a,\bar{X})-n(a,X))$. Тогда применим теорему 4.3 к $\frac{f(\mathbf{x})}{\sigma}$. После этого в полученном выражении поделим все части неравенства на ℓ и после нормировки t получим утверждение теоремы. □

Обращая оценку, получаем простое и практичное следствие

Теорема 4.5. *В условиях теоремы 4.4 для $\delta > 0$ с вероятностью не меньшей $1-\delta$ имеет место неравенство*

$$\max_{a\in A}(\nu(a,\bar{X}) - \nu(a,X)) < \mathcal{EOF}_{\max} + \frac{4\sigma}{\ell}\left(2\sqrt{\pi} + \sqrt{\ln\left(\frac{2}{\delta}\right)}\right).$$

Согласно ранним замечаниям все слагаемые правой части неравенства можно оценить лишь по наблюдаемой выборке.

Чтобы лучше понять силу результата теоремы 4.4 вспомним, что ранее для $t > 0$ была получена формула

$$\mathsf{P}\left(\max_{a\in A}(\nu(a,\bar{X}) - \nu(a,X)) > t\right) \le |A|\exp\left(\frac{-t^2\ell}{2}\right).$$

Оценивая сверху $\mathcal{EOF}_{\max}$ как $\sqrt{\frac{2\ln(|A|)}{\ell}}$, получаем, что для $t > 0$

$$\mathsf{P}\left(\max_{a\in A}(\nu(a,\bar{X}) - \nu(a,X)) \ge \mathcal{EOF}_{\max} + \frac{8\sqrt{\pi}\sigma}{\ell} + t\right) \le \exp\left(-\frac{\hat{t}^2\ell}{2}\right),$$

где

$$\hat{t}^2 = t^2 + \frac{64\pi\sigma^2}{\ell^2} + 2t\sqrt{\frac{2\ln(|A|)}{\ell}} + \frac{16\sqrt{\pi}\sigma}{\ell}\left(t + \sqrt{\frac{2\ln(|A|)}{\ell}}\right).$$

В практических задачах методы обучения устроены так, что выбираются лишь алгоритмы нижних слоев семейства A. Это наводит на естественную мысль о локализации семейства [10], то есть рассмотрении только тех алгоритмов, которые допускают не более определенного некоторым порогом числа ошибок. При этом даже равномерная оценка должна давать хороший результат.

Отметим, что при уменьшении σ, то есть, контроле за максимальным числом ошибок, хвост последней формулы практически не уменьшается и сохраняет неизменным член $t^2\ell$, в то время как более мощный результат теоремы 4.4 говорит о существенном «ускорении» концентрации.

Заметим, что на этом шаге нельзя учесть метод обучения, так как при доказательстве мы существенно пользовались выпуклостью функции f. Если бы на разбиении не выбирался алгоритм, который максимизирует указанное скалярное произведение, то выпуклости бы не было. Можно показать, что условия выпуклости в теоремах типа 4.1 являются необходимыми [14].

4.4 Неравенства типа МакДиармида на слоях куба

Итак, повторим, что основным преимуществом оценок, получаемых в комбинаторном подходе, является возможность учесть метод обучения. Для минимизации эмпирического риска результаты предыдущего раздела в общем случае применены быть не могут. В данном разделе будет предложено неравенство типа МакДиармида для функций, заданных на слое куба. Затем будет введено понятие устойчивости семейства алгоритмов и для частного случая — пессимистичной минимизации эмпирического риска будет получена оценка, позволяющая оценить частоту ошибок на ненаблюдаемой контрольной выборке.

В данном разделе будет применена техника логарифмических неравенств Соболева [11]. В работах [6, 7] получено такое неравенство для слоя дискретного куба. Оно согласно [11] влечёт нужную нам концентрацию меры. В работе мы специально не будем подробно описывать результаты связанной теории, подробно изложенной в [11], а непосредственно воспользуемся основными результатами для получения необходимого нам неравенства. Отметим, что подход, связанный с использованием логарифмических неравенств Соболева даёт с точки зрения констант даже более точные неравенства концентрации, чем подход Талаграна, использованный в предыдущем разделе. Таким образом, с помощью результатов [6, 11] можно улучшить константы предыдущего раздела.

Пусть f — положительная функция, заданная на слое куба $\{1,-1\}^L$, состоящем из вершин с ℓ экземплярами -1.

Пусть $\mathbf{x},\mathbf{y}$ пара вершин слоя куба таких, что $\mathbf{y}$ может быть получена из $\mathbf{x}$ одной перестановкой пары координат. Другими словами, либо оба вектора совпадают, либо отличаются перестановкой 1 с -1. Теперь дополнительно наложим на нашу функцию

условие ограниченных приращений

$$\max_{\mathbf{x},\mathbf{y}} |f(\mathbf{x}) - f(\mathbf{y})| \leq c,$$

где c – некоторая константа.

Теорема 4.6 (Бобков [6, 7]). *Пусть на слое дискретного $\{1,-1\}^L$ куба с $L-\ell$ единицами введено равномерное распределение. Для описанной функции f выполнен следующий вариант логарифмического неравенства Соболева*

$$\mathbf{E} f(\mathbf{x}) \exp(f(\mathbf{x})) - \mathbf{E} \exp(f(\mathbf{x})) \ln\left(\mathbf{E} \exp(f(\mathbf{x}))\right) \leq \frac{\ell(L-\ell)c^2}{L+2} \mathbf{E} \exp(f(\mathbf{x})).$$

Отсюда с помощью стандартного рассуждения из [11] можно получить нужное нам неравенство концентрации.

Теорема 4.7. *Для описанной функции f и любого $t > 0$ имеет место неравенство*

$$\mathsf{P}(f(\mathbf{x}) \geq \mathbf{E} f(\mathbf{x}) + t) \leq \exp\left(\frac{-(L+2)t^2}{2\ell(L-\ell)c^2}\right).$$

Доказательство. Для числа $\lambda > 0$ рассмотрим функцию λf. Очевидно, что она обладает свойствами функции f с заменой константы условия конечных приращений c на $c\lambda$. Обозначим $C = \frac{\ell(L-\ell)c^2}{L+2}$ и $H(\lambda) = \mathbf{E} \exp(\lambda f(\mathbf{x}))$. Тогда условие теоремы 4.6 может быть выражено как

$$\lambda H'(\lambda) \leq H(\lambda) \ln\left(H(\lambda)\right) + C\lambda^2 H(\lambda).$$

Последнее влечёт

$$\frac{d}{d\lambda}\left(\frac{H(\lambda)}{\lambda}\right) \leq C.$$

Раскладывая $H(\lambda)$ в окрестности нуля по формуле Тейлора, получаем, что

$$\lim_{\lambda \to 0}\left(\frac{H(\lambda)}{\lambda}\right) = \mathbf{E} f(\mathbf{x}).$$

Интегрируя последнее неравенство, заключаем, что для $\lambda > 0$

$$\frac{H(\lambda)}{\lambda} \leq \mathbf{E} f(\mathbf{x}) + C\lambda.$$

Отсюда

$$\mathbf{E} \exp(\lambda f(\mathbf{x})) \leq \exp\left(\lambda \mathbf{E} f(\mathbf{x}) + C\lambda^2\right).$$

Отсюда как и ранее с помощью неравенства Маркова, оптимизируя по λ, получаем утверждение теоремы. □

4.5 Устойчивость метода обучения

Результаты последнего раздела могут быть применены непосредственно для оценки частоты ошибок на ненаблюдаемой контрольной выборке. Будем считать, что $\ell = k$ и число ℓ чётно, тогда для обучающей выборки X определим наблюдаемое семейство алгоритмов $A(X)$ и соответственно $\mathcal{EOF}_\mu(X)$. Для обучающей выборки X обозначим X_{ij} выборку, полученную заменой её i-го объекта на j-ый объект $\mathbb{X}$. Соответствующая контрольная выборка обозначается $\bar{X}_{ij} = \mathbb{X} \setminus X_{ij}$. Введём естественное предположение устойчивости.

ПМЭР (вместе с семейством $A(\mathbb{X})$) называется α, β*–устойчивым*, если для всех разбиений генеральной выборки на обучение X и контроль $\bar{X}$ выполнены следующие условия

1. $\mathcal{EOF}_\mu(\mathbb{X}) - \mathcal{EOF}_\mu(X) \leq \alpha$.
2. $\left|n(\mu X, X) - \max\limits_{ij} n(\mu X_{ij}, X_{ij})\right| \leq \beta$ и $\left|n(\mu X, \bar{X}) - \max\limits_{ij} n(\mu X_{ij}, \bar{X}_{ij})\right| \leq \beta$.

Обе предположения очень естественны. Первое утверждает, что ожидаемая переобученность, вычисленная по наблюдаемой обучающей выборке, слабо отличается от вычисленной на всём контроле. Отметим, что это более слабое предпложение чем близость вероятностей переобучения, вычисленных по всей генеральной выборке и обучающей выборке соответственно. Второе говорит о том, что замена одного объекта в обучении (или контроле) незначительно меняет число ошибок выбираемого алгоритма.

Теорема 4.8. *Пусть* $\ell = k = \frac{L}{2}$, μ *–* α, β*-устойчивый ПМЭР, тогда для всех* $t > 0$

$$\mathsf{P}\left(\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geq \mathcal{EOF}_\mu(X) + \alpha + t\right) \leq \exp\left(\frac{-(\ell+1)t^2}{4\beta^2}\right).$$

Доказателъство. Разбиению $X \sqcup \bar{X}$ ставим в соответствие вершину слоя $\{1, -1\}^L$ куба $\mathbf{x}$, для которой всем позициям X в генеральной выборке соответствуют -1, а позициям $\bar{X}$ соответствуют 1. Теперь с учётом условия устойчивости нужно применить теорему 4.7 для функции f, такой что $f(\mathbf{x}) = \nu(\mu X, \bar{X}) - \nu(\mu X, X)$. □

Как следствие получаем желаемую оценку, вычислимую по наблюдаемой выборке.

Теорема 4.9. *В условиях теоремы 4.8 для $\delta > 0$ с вероятность не меньшей $1 - \delta$ имеет место неравенство*

$$\nu(\mu X, \bar{X}) < \nu(\mu X, X) + \mathcal{EOF}_{\mu}(X) + \alpha + \frac{2\beta}{\sqrt{\ell + 1}} \sqrt{\ln\left(\frac{1}{\delta}\right)}.$$

Стоит напомнить, что теорема 3.1 даёт оценку $\mathcal{EOF}_{\mu}$, учитывающую структуру семейства алгоритмов. Параметры α, β могут быть оценены для конкретных семейств $A(\mathbb{X})$.

5 Выводы

- В рамках комбинаторного подхода получены новые оценки ожидаемой переобученности.
- Исследованы их асимптотики и факторы завышенности.
- Доказан общий критерий точности комбинаторных оценок обобщающей способности.
- Доказано изопериметрическое неравенство на слоях дискретного куба.
- Получены оценки частоты ошибок на ненаблюдаемой контрольной выборке, вычислимые по наблюдаемой обучающей.

Список литературы

[1] *Животовский Н. К., Воронцов К. В.* Критерий точности комбинаторных оценок вероятности переобучения // Сборник докладов 9-ой международной конференции «Интеллектуализация обработки информации». — М.: Торус Пресс, 2012. — С. 25–28.

[2] *Прудников А. П., Брычков Ю.А., Маричев О.И* Интегралы и ряды. Том 3. Специальные функции. Дополнительные главы. // М.: ФизМатЛит, 2003.

[3] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. — 2005. — No. 9. — Pp. 323–375.

[4] *Bednorz W.* A theorem of majorizing measures // Annals of probability. — 2006. — Pp. 1771–1781.

[5] *Bednorz W., Latala R.* On the boundness of Bernouilly processes // 2013. — http://arxiv.org/abs/1305.4292

[6] *Bobkov S. G., Tetali, P.* Modified logarithmic Sobolev inequalities in discrete settings // Journal of Theoretical Probability 2006. — No. 2. — Pp. 289–336.

[7] *Bobkov S. G.* Concentration of normalized sums and a central limit theorem for noncorrelated random variables // Annals of probability 2004. — No. 4. — Pp. 2884-2907.

[8] *Devroye L., Lugosi G.* Combinatorial Methods in Density Estimation // Springer Series in Statistics. Springer-Verlag, 2001.

[9] *Haussler D.* Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension // Journal of Combinatorial Theory. — 1995. — Pp. 217–232.

[10] *Koltchinskii V.* Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems // Ecole d'Etŕe de Probabilitŕes de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics. Springer-Verlag, 2011.

[11] *Ledoux M.* The Concentration of Measure Phenomenon // American Mathematical Society, 2005.

[12] *Spokoiny V.* Wilks Theorem for penalized maximum likelihood estimators // 2013. — http://arxiv.org/abs/1205.0498

[13] *Talagrand M.* The Generic Chaining. Upper and Lower Bounds of Stochastic Processes // Springer Monographs in Mathematics, 2005.

[14] *Talagrand M.* An Isoperimetric Theorem on the Cube and the Kinitchine-Kahane Inequalities // Proceedings of the American Mathematical Society 1988. — Pp. 905-909.

[15] *Talagrand M.* New concentration inequalities in product spaces // Inventiones mathematicae 1996. — Pp. 505-563.

[16] *Talagrand M.* Concentration of Measure and Isoperimetric Inequalities in Product Spaces // Publications Mathematiques de l'Institut des Hautes Etudes Scientifiques 1995. — Pp. 73-205.

[17] *Tao T.* An Epsilon of Room, II // American Mathematical Society, 2010.

[18] *Vapnik V.* Statistical Learning Theory. — John Wiley and Sons, New York, 1998.

[19] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // *Pattern Recognition and Image Analysis.* — 2008. — Vol. 18, no. 2. — Pp. 243–259.

[20] *Vorontsov K. V.* Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // *Pattern Recognition and Image Analysis.* — 2009. — Vol. 19, no. 3. — Pp. 412–420.

[21] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // *Pattern Recognition and Image Analysis.* — 2010. — Vol. 20, no. 3. — Pp. 269–285.