Теория статистического обучения: оценки обобщающей способности и избыточного риска.

Толстихин Илья
iliya.tolstikhin@gmail.com

Апрель, 2012

1 Лекция 1: проблемы старых подходов

Мы начнем с рассмотрения общей **вероятностной** постановкой задачи обучения по прецедентам, принятой в теории статистического обучения (Statistical Learning Theory, SLT).

Пусть $\mathcal{X}$ — пространство объектов, $\mathcal{Y}$ — пространство ответов (или классов). Например, в случае задачи бинарной классификации $\mathcal{Y} = \{-1, +1\}$, в случае восстановления регрессии — $\mathcal{Y} = \mathbb{R}$. Пусть на Декартовом произведении $\mathcal{X} \times \mathcal{Y}$ задано **неизвестное нам** вероятностное распределение P. Мы не будем сосредотачиваться на формальном и строгом введении вероятностного пространства $(\mathcal{X} \times \mathcal{Y}, \mathcal{A}, \mathsf{P})$ в этом докладе и будем считать, что «все введено за нас». Нам также дана обучающая выборка $Z = \{Z_1, \dots, Z_i\} = \{(X_i, Y_i)\}_{i=1}^n$ — последовательность независимых одинаково распределенных величин из распределения P на множестве $\mathcal{X} \times \mathcal{Y}$.

В качестве примера использования вероятностной постановки можно привести нормальный дискриминантный анализ — при этом предполагается, что $\mathsf{P}(X, Y) = P(Y)P(X|Y)$, где плотность класса $P(X|Y) \sim \mathcal{N}(\mu_Y, \Sigma_Y)$. Однако, в общем случае мы не знаем распределения P. Большáя часть машинного обучения занимается оценкой этого распределения для последующего применения этих оценок при построении классификаторов.

Задача обучения по прецедентам обычно состоит в поиске измеримого отображения $g\colon \mathcal{X} \to \mathcal{Y}$. *Об измеримости заходит речь, поскольку мы будем работать с различными вероятностными характеристиками случайных величин. Эти вопросы тоже рассматриваться не будут — будем считать, что все функции «хорошие», и мы можем спокойно интегрировать их по заданному распределению.* Каков критерий выбора отображения g? Один из понятных вариантов — минимизировать вероятность ошибки алгоритма g: $\mathsf{P}\{g(X) \neq Y\} \to \min$. *Всюду далее понятия «алгоритм», «классификатор», «отображение $\mathcal{X} \to \mathcal{Y}$» будем использовать взаимозаменяемо.*

В случае задачи восстановления регрессии ($\mathcal{Y} = \mathbb{R}$) этот критерий становится необоснованно жестким. Поэтому чаще всего вводится неотрицательная **функция потерь** (loss function) $\ell\colon \mathcal{Y}^2 \to \mathbb{R}$, при этом $\ell(y, \hat{y})$ выражает величину потерь, связанных с отнесением объекта класса y к классу $\hat{y}$. Часто используется индикатор ошибки $\ell(y, \hat{y}) = I(y \neq \hat{y})$ в случае задачи классификации и квадратичный риск $\ell(y, \hat{y}) = (y - \hat{y})^2$ в задачах восстановления регрессии. Тогда задача обучения по прецедентам сводится к поиску отображения g с малым значением матожидания потерь, или функционала **среднего риска**:

$$\mathsf{E}\,\ell\big(g(X), Y\big) \to \min_g .$$

всюду далее матожидания будут браться относительно распределения P.

Рассмотрим частный случай бинарной классификации $\mathcal{Y} = \{+1, -1\}$ с функцией потерь $\ell(y, \hat{y}) = I(y \neq \hat{y})$. В этом случае средний риск $\mathsf{E}\,\ell\big(g(X), Y\big)$ превращается просто в вероятность ошибки классификатора g. Введем **функцию регрессии** (regression function) $\eta(x) = \mathsf{E}\{Y|X = x\} = 2\,\mathsf{P}\{Y = +1|X = x\} - 1$. Известно, что в этом случае отображение $b(x) = \operatorname{sign}\big(\eta(x)\big)$ минимизирует функционал среднего риска. Отображение b называется **байесовским классификатором**, а его средний риск $\mathsf{P}\{b(X) \neq Y\} = \min_g \mathsf{P}\{g(X) \neq Y\}$ называют **байесовским риском**.

Тот факт, что условное матожидание $\mathsf{E}\{Y|X = x\}$ не вырождено — не равно $+1$ или -1 — указывает на наличие «шума» в ответе Y. Это вполне реалистичная картина — представим задачу классификации с 2-мя перекрывающимися гауссианами. Тогда на границе перекрытия нет

возможности точно сказать, какая из двух «шапок» породила ее. Более того, любая точка пространства в этом случае имеет **ненулевую** *вероятность принадлежности к каждому из классов.*

В реальных ситуациях отображение g ищут в неком ограниченном классе функций $\mathcal{G}$. Это делается из разных соображений, о которых мы упомянем позже. Вследствие этого, нет никаких гарантий, что байесовский классификатор содержится в множестве поиска $\mathcal{G}$. Вопросы выбора семейства (или модели) алгоритмов $\mathcal{G}$ представляет широкую область теории статистического обучения и известен под названием "model selection". Этой темы мы не будем касаться. Поэтому будем считать, что множество $\mathcal{G}$ заранее фиксировано неким образом.

Введем ряд удобных обозначений. Как только фиксирована функция потерь ℓ, мы можем ввести множество

$$\mathcal{F} = \{f(X,Y) = \ell\big(g(X), Y\big),\ g \in \mathcal{G}\}$$

— **класс потерь** (loss class), ассоциированных с семейством алгоритмов $\mathcal{G}$. Еще одно обозначение

$$Pf \equiv \mathsf{E}f = \int_{\mathcal{X}\times\mathcal{Y}} f d\,\mathsf{P}$$

— средний риск алгоритма g, соответствующего f.

Во-первых, всюду далее мы будем оперировать классом потерь и забудем о структуре пространства $\mathcal{X}\times\mathcal{Y}$, обозначив его $\mathcal{Z}$, а его элементы — $Z = (X,Y)$. Таким образом, мы переходим к рассмотрению абстрактной постановки задачи с множеством $\mathcal{Z}$ и распределением P на нем. Во-вторых, из технических соображений, мы будем предполагать, что функции класса потерь $\mathcal{F}$ **равномерно ограничены** — иными словами, существует такое неотрицательное число U, что $\sup_{f\in\mathcal{F}}|f| \leqslant U$. Для удобства будем полагать, что $\ell\colon \mathcal{Y}^2 \to [0,1]$, то есть $U = 1$.

Наша задача сводится к

$$Pf \to \min_{f\in\mathcal{F}}. \tag{1}$$

Проблема, конечно, в том, что нам неизвестно распределение P. Зато мы знаем поведение класса $\mathcal{F}$ на обучающей выборке $\{Z_1,\ldots,Z_n\} = \{(X_i,Y_i)\}_{i=1}^n$. Используя **эмпирическое распределение** $\mathsf{P}_n(Z) = \sum_{i=1}^n I(Z = Z_i)$ (дискретное распределение с вероятностями $1/n$ в точках обучающей выборки), введем понятие **эмпирического риска**:

$$P_n f \equiv \int_{\mathcal{Z}} f d\,\mathsf{P}_n = \frac{1}{n}\sum_{i=1}^n f(Z_i).$$

Закон больших чисел дает нам основания приближать неизвестное распределение P эмпирическим распределением P_n. Поэтому вместо решения задачи (1) мы будем решать следующую задачу **минимизации эмпирического риска**:

$$P_n f \to \min_{f\in\mathcal{F}}. \tag{2}$$

Решение задачи (2) обозначим $\hat{f}_n$. Отображение $\hat{f}_n$ представляется нам хорошим кандидатом решения исходной задачи (1).

Скажем несколько слов о выборке семейства алгоритмов $\mathcal{G}$:

- Если не ограничивать $\mathcal{G}$ вообще, то возможна ситуация очевидного переобучения: $P_n f = 0$, $Pf = 1$. Алгоритм не ошибается на обучающей выборке и ошибается на всех прочих объектах.
- Если $\mathcal{G}$ очень мало, то $\min_{f\in\mathcal{F}} Pf$ будет очень далек от байесовского риска.
- Если $\mathcal{G}$ сильно увеличить, то, во-первых, снова велик шанс переобучиться, и, во-вторых, становится сложным решить задачу (2).

Так или иначе, найдя $\hat{f}_n$, мы хотим оценить, насколько хорошо это отображение справляется с исходной задачей (1). Для этого можно ввести несколько разных характеристик решения $\hat{f}_n$:

1. Оценка **избыточный риска** функции f (excess risk bound):

$$\mathcal{E}(f) \equiv \mathcal{E}_P(f) = Pf - \inf_{g\in\mathcal{F}} Pg \leqslant B_1(n,\mathcal{F}). \tag{3}$$

2. Оценка **обобщающей способности** функции f (error bound):

$$Pf - P_n f \leqslant B_2(n, \mathcal{F}). \tag{4}$$

Таким образом, избыточный риск $\mathcal{E}(\hat{f}_n)$ показывает, «насколько близко мы подобрались» к лучшей функции в семействе $\mathcal{F}$, а величина $P\hat{f}_n - P_n\hat{f}_n$ — насколько точно эмпирический риск функции приближает её реальный риск. Обе величины представляют для нас интерес.

Для дальнейших рассуждений крайне важно отметить, что $P(\hat{f}_n)$ — **случайная величина**, поскольку функция $\hat{f}_n$ выбирается на основе **случайной** выборки $\{Z_1, \ldots, Z_n\}$. Она может быть формально записана в виде условного матожидания: $P\hat{f}_n = \mathsf{E}\{\hat{f}_n | (Z_1, \ldots, Z_n)\}$ — тогда становится очевидна ее зависимость от обучающей выборки. Поэтому неравенства вида (3) и (4) для функции $\hat{f}_n$ будут всегда иметь вероятностный характер, например:

$$\mathbb{P}\{\mathcal{E}(\hat{f}_n) \geqslant \delta\} \leqslant \varepsilon(\delta, n, \mathcal{F}) \quad \text{или} \quad \mathbb{P}\{P\hat{f}_n - P_n\hat{f}_n \geqslant \delta\} \leqslant \varepsilon(\delta, n, \mathcal{F}), \tag{5}$$

где распределение $\mathbb{P}$ — декартово произведение $\mathsf{P}^{\otimes n}$ исходного распределения P — распределение на случайных простых выборках длиной n из распределения P.

Бóльшая часть теории статистического обучения посвящена построению как можно более точных оценок вида (5), учитывающих «структуру» класса функций $\mathcal{F}$. Причем построению как ненаблюдаемых и зависящих от распределения P оценок (distribution dependant bounds), так и оценок, вычислимых по обучающей выборке (data dependant bounds). Очевидно, второй класс оценок представляется нам наиболее полезным.

Для начала мы займемся получением простейших оценок обобщающей способности (4).

1.1 Оценки обобщающей способности (error bounds).

Покажем, где в этой задаче подводные камни и почему она не может быть решена «влоб». Пристально присмотримся к интересующей нас величине $Pf - P_n f$, забыв для начала, что нас интересует случай $f = \hat{f}_n$. Перед нами разность матожидания случайной величины $f(Z)$ и ее среднего выборочного значения: $Pf - P_n f = \mathsf{E}f - \frac{1}{n}\sum_{i=1}^n f(Z_i)$. Закон больших чисел утверждает, что

$$\mathbb{P}\Big\{\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathsf{E}f(Z) = 0\Big\} = 1.$$

То есть при достаточно большом размере обучающей выборки среднее выборочное отлично приближает искомое матожидание. В теории вероятностей существует неассимптотический количественный аналог закона больших чисел на случай, когда случайные величины ограничены. Это неравенство Хевдинга:

Теорема 1.1 (Неравенство Хевдинга) *Пусть $Z_1, \ldots, Z_n$ — независимые одинаково распределенные согласно P случайные величины и $f(Z) \in [a, b]$. Тогда для всех $\varepsilon > 0$ справедливо:*

$$\mathsf{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathsf{E}f(Z)\Big| > \varepsilon\Big\} \leqslant 2\exp\Big(-\frac{2n\varepsilon^2}{(b-a)^2}\Big).$$

Обозначив правую часть δ, фиксировав конкретную функцию f и используя наши обозначения, мы получаем: с вероятностью не меньше $1 - \delta$ справедливо:

$$|Pf - P_n f| \leqslant \sqrt{\frac{\log\frac{2}{\delta}}{2n}} \tag{6}$$

— исходя из того, что $\ell(\cdot, \cdot) \in [0, 1]$.

Здесь очень важно понимать следующее: результат справедлив **для фиксированной функции** и вероятность рассматривается относительно повторного выбора случайной обучающей выборки $\{Z_1, \ldots, Z_n\}$. Если функция f зависит от обучающей выборки (как в нашем случае $\hat{f}_n$), то этот результат **неприменим**. Также обратим внимание, что единственная характеристика случайной величины, участвующая в правой части — $U = b - a$, равномерно ограничивающая значение случайной величины.

Ограниченность этого результата состоит в следующем: утверждается, что для каждой функции f в классе $\mathcal{F}$ существует событие S_f, реализуемое с большой вероятностью, на котором $|Pf - P_n f|$ мала. Однако ничего не утверждается о связи событий S_f для разных функций класса $\mathcal{F}$ — они могут оказаться совершенно разными для разных функций. Значит, при конкретной реализации обучающей выборки $\{Z_1, \ldots, Z_n\}$ неравенство (6) будет справедливо только для заранее неизвестного **подмножества** класса $\mathcal{F}$. Если посмотреть на рисунок 1, то можно лучше понять картину происходящего (здесь R и R_n обозначают наши P и P_n соответственно). Кривая R обозначает средний риск и она фиксирована. Кривая R_n — эмпирический риск, и она меняется вместе с обучающей выборкой. Неравенство Хевдинга описывает колебание точек $R_n(g)$ фиксированной функции g вокруг значения $R(g)$. Если класс функций достаточно большой, то для концретной обучающей выборки найдутся функции, для которых $|Pf - P_n f|$ велико.

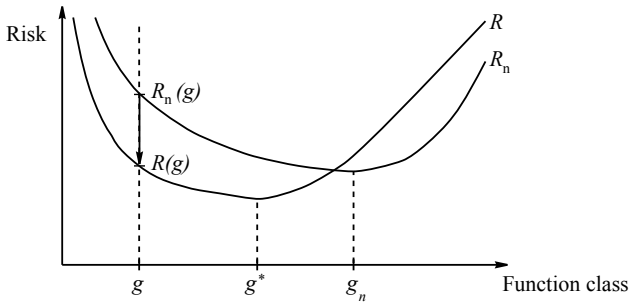

Рис. 1: Эмпирический и средний риск класса функций.

Как же нам оценить волнующую нас величину $|P\hat{f}_n - P_n\hat{f}_n|$? Ниже приведено уже классическое неравенство, в котором берут свое начало многие подходы в теории статистического обучения:

$$(P - P_n)\hat{f}_n \leqslant \sup_{f \in \mathcal{F}} (P - P_n) f. \tag{7}$$

Идея простая: ограничив максимальное по классу отклонение, мы тем самым ограничим его и для минимизатора эмпирического риска $\hat{f}_n$. Это очень простая идея и ей пошли с самого начала. Рассмотрим несколько примеров оценок правой части неравенства в различных ситуациях.

Конечный класс функций $\mathcal{F} = \{f_1, \ldots, f_N\}$. Покажем, как в этом случае получить равномерный по классу функций $\mathcal{F}$ аналог оценки Хевдинга. Используя неравенство Буля (union bound) вместе с неравенством Хевдинга, мы получаем:

$$\begin{aligned}
&\mathbb{P}\Big\{\forall f \in \{f_1, \ldots, f_N\} \colon Pf - P_n f \leqslant \varepsilon\Big\} = \\
&= 1 - \mathbb{P}\Big\{\exists f \in \{f_1, \ldots, f_N\} \colon Pf - P_n f > \varepsilon\Big\} \geqslant \\
&\geqslant 1 - \sum_{i=1}^{N} \mathbb{P}\{Pf_i - P_n f_i > \varepsilon\} \geqslant 1 - N\exp(-2n\varepsilon^2).
\end{aligned}$$

Обозначив $\delta = N\exp(-2n\varepsilon^2)$, мы получили: для любых $\delta > 0$ с вероятностью не меньше $1 - \delta$ справедливо:

$$\forall f \in \mathcal{F}, \;\; Pf - P_n f \leqslant \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2n}}. \tag{8}$$

Обратим внимание, что единственное отличие этого результата от (6) — присутствие величины $\log N$ в числителе дроби. Это наша плата за требование равномерного по классу $\mathcal{F}$ контроля уклонения эмпирического риска от равномерного и мы еще ни раз столкнемся с похожими выражениями.

Бесконечный но счетный класс функций $\mathcal{F}$. В этом случае работает такая же логика, как и в прошлом пункте. Для каждой отдельно взятой функции $f_i \in \mathcal{F}$ мы можем записать

$$\mathbb{P}\left\{ Pf_i - P_n f_i > \sqrt{\frac{\log \frac{1}{\delta(f_i)}}{2n}} \right\} \leqslant \delta(f_i).$$

Мы распорядимся свободой выбора величины $\delta(f_i) > 0$ для каждой отдельной функции из $\mathcal{F}$ и положим $\delta(f_i) = \delta \cdot q(f_i)$, так что $\sum_{f_i \in \mathcal{F}} q(f_i) = 1$ и $\delta > 0$. Тогда, снова применив неравенство Буля, мы получим:

$$\mathbb{P}\left\{ \exists f \in \mathcal{F} \colon Pf - P_n f > \sqrt{\frac{\log \frac{1}{\delta(f)}}{2n}} \right\} \leqslant \sum_{f \in \mathcal{F}} \delta(f) = \delta.$$

Обратив вероятность, получаем утверждение: с вероятностью не меньше $1 - \delta$ справедливо:

$$\forall f \in \mathcal{F}, \;\; Pf - P_n f \leqslant \sqrt{\frac{\log \frac{1}{q(f)} + \log \frac{1}{\delta}}{2n}}. \tag{9}$$

Эта оценка в литературе известна под названием «бритва Оккама». Интересно отметить, что если класс функций $\mathcal{F}$ конечен, то положив в качестве q равномерное распределение на его элементах, мы получаем в точности прошлую обобщенную оценку типа Хевдинга (8).

Еще один важный момент: эта оценка дает нам возможность использовать некие «дополнительные» знания о задаче и семействе $\mathcal{F}$. Если вся величина $q(\cdot)$ была бы сконцентрирована на минимизаторе эмпирического риска $\hat{f}_n$, то мы бы получили полный аналог оценки типа Хевдинга в случая единственной функции в семействе: $\mathcal{F} = \{f_0\}$ ((C) «в идеале хотелось бы предсказывать вероятность ошибки с той же точностью, с какой закон больших чисел предсказывает частоту выпадения орла...»). Однако, дело здесь в том, что мы должны выбрать $q(\cdot)$ **до того, как увидим обучающую выборку**. А значит, мы не можем предугадать, какая из функций будет минимизировать эмпирический риск. Тем не менее хороший выбор величины $q(\cdot)$ позволяет улучшать оценку.

Бесконечный несчетный класс функций $\mathcal{F}$. Рассмотрим случай, когда класс потерь $\mathcal{F}$ — бесконечный несчетный. В этом случае мы ограничим рассмотрение случаем $\ell(y, y^*) = I(y \neq y^*)$ — бинарной функцией потерь.

В этом случае логика прошлых доказательств уже не работает — например, потому что мы не можем применить неравенство Буля. Нам поможет понятие **функции роста** $S_{\mathcal{F}}(n)$:

$$S_{\mathcal{F}}(n) = \sup_{z_1,\ldots,z_n} \left| \mathcal{F}_{z_1,\ldots,z_n} \right| = \sup_{z_1,\ldots,z_n} \left| \left\{ \big(f(z_1), \ldots, f(z_n)\big) \colon f \in \mathcal{F} \right\} \right|.$$

Справедливо элементарное неравенство $S_{\mathcal{F}}(n) \leqslant 2^n$.

Оказывается, справедлива следующая оценка

Теорема 1.2 (оценка Вапника–Червоненкиса) *Для всех $\delta > 0$ с вероятностью не меньше $1 - \delta$ справедливо:*

$$\forall f \in \mathcal{F}, \;\; Pf - P_n f \leqslant 2\sqrt{2\frac{\log S_{\mathcal{F}}(2n) + \log \frac{2}{\delta}}{n}}. \tag{10}$$

Крайне полезно и важно обсудить доказательство теоремы Вапника–Червоненкиса. Основной «пружиной» в доказательстве является следующий результат, известный в литературе как **лемма симметризации**. Сформулируем ее без доказательства и кратко обсудим ее смысл.

Лемма 1.1 (Симметризация) *Для всех $t > 0$, таких что $nt^2 > 2$,*

$$\mathbb{P}\Big\{ \sup_{f \in \mathcal{F}} (Pf - P_n f) \geqslant t \Big\} \leqslant 2\mathbb{P}\Big\{ \sup_{f \in \mathcal{F}} (P'_n f - P_n f) \geqslant t/2 \Big\}. \tag{11}$$

Здесь P'_n — эмпирическое распределение относительно независимой «призрачной» (ghost) выборки $\{Z'_1, \ldots, Z'_n\}$ — независимой копии обучающей выборки. Таким образом симметризация позволяет заменить неизвестный средний риск функции ее средним выборочным значением на еще одной

независимой выборке. Подобное введение дополнительной **рандомизации** — частый прием в теории вероятностей. В результате правая часть неравенства (11) зависит лишь от проекции класса $\mathcal{F}$ на двойную выборку $\{Z_1, \dots, Z_n, Z'_1, \dots, Z'_n\}$ длины $2n$. Поскольку эта проекция **конечна**, мы можем снова использовать неравенство Буля вместе со слегка модифицированной версией неравенства Хевдинга:

$$\mathbb{P}\{P_n f - P'_n f > t\} \leqslant \exp\Big(-\frac{nt^2}{2}\Big).$$

Собрав все кубики вместе, получим:

$$\begin{aligned}
\mathbb{P}\Big\{\sup_{f\in\mathcal{F}}(P-P_n)f \geqslant t\Big\} &\leqslant 2\mathbb{P}\Big\{\sup_{f\in\mathcal{F}}(P'_n-P_n)f \geqslant t/2\Big\} = \\
&= 2\mathbb{P}\Big\{\sup\nolimits_{f\in\mathcal{F}_{Z_1,\dots,Z_n,Z'_1,\dots,Z'_n}}(P'_n-P_n)f \geqslant t/2\Big\} \leqslant \\
&\leqslant 2S_{\mathcal{F}}(2n)\mathbb{P}\{(P'_n-P_n)f > t/2\} \leqslant 4S_{\mathcal{F}}(2n)\exp\Big(-\frac{nt^2}{8}\Big).
\end{aligned}$$

□. Далее короткое обсуждение леммы Сауера, VC-классов и сходимости $\sup|P-Pn|$ к нулю.

Пришло время подытожить, что нам удалось достичь. Оценки (6), (9) и (10) имеют очень похожий вид. Все они имеют порядок $O(n^{-1/2})$. Все они в качестве основного ингредиента используют три шага: а) использование $\sup_{f\in\mathcal{F}}|(P-P_n)f|$; б) использование неравенства Буля и в) использование неравенства типа Хевдинга для отдельно взятой функции. Перечислим вкратце причины завышенности перечисленных оценок:

- При оценке величины $(P-P_n)\hat{f}_n$ сверху $\sup_{f\in\mathcal{F}}|(P-P_n)f|$ мы рассматриваем «худший случай». Действительно, эти оценки справедливы для **всех** отображений класса $\mathcal{F}$. Рисунок 1 продемонстрировал нам, что в достаточно больших для решения прикладных задач классов функций с высокой вероятностью найдется функция f с **большим** значением $Pf-P_nf$. Таким образом, оценки становятся завышенными.

- Неравенство Буля становится точной оценкой в том случае, когда все функции в классе независимы. Это требование чаще всего не выполняется, что ведет к потере точности.

- Неравенство Хевдинга использует только информацию о величине U, равномерно ограничивающей функции из класса $\mathcal{F}$, и не использует дисперсии функций. Это во многих случаях существенно ухудшает оценки.

Самым критическим местом считается оценка с помощью супремума по всему семейству функций. Дальше в докладе мы будем рассматривать оценки избыточного риска, поэтому неравенство будет ограничивать именно его:

$$\begin{aligned}
\mathcal{E}(\hat{f}_n) = P\hat{f}_n - \inf_{f\in\mathcal{F}} Pf = P\hat{f}_n - P\bar{f} = \\
= P_n\hat{f}_n - P_n\bar{f} + (P-P_n)(\hat{f}_n-\bar{f}) \leqslant \\
\leqslant \sup_{f,g\in\mathcal{F}}|(P-P_n)(f-g)| \leqslant 2\sup_{f\in\mathcal{F}}|(P-P_n)f| \equiv 2\|P-P_n\|_{\mathcal{F}}.
\end{aligned} \tag{12}$$

(Обобщающая способность $(P-P_n)\hat{f}_n$, рассматривавшаяся до сих пор, оценивается сверху той же величиной без множитея 2).

Случайные величины $\{(P-P_n)f\}_{f\in\mathcal{F}}$ в литературе называются **эмпирическим процессом, индексированным классом** $\mathcal{F}$. Эти случайные процессы изучаются в теории эмпирических процессов. Одной из наиболее полно изученных характеристик, связанных с эмпирическими процессами, является какраз их норма $\|P-P_n\|_{\mathcal{F}}$. Этот факт указывает на чрезвычайную полезность теории эмпирических процессов в приложении к статистическому обучению и зачастую большинство результатов, используемых в последнее время при выводе оценок, заимствованы оттуда.

Все оценки, которые мы видели до сих пор, утверждали, что величина $\|P-P_n\|_{\mathcal{F}}$ имеет порядок $O(n^{-1/2})$. Более того, в случае конечного класса, это просто следствие из центральной предельной теоремы. В случае бесконечного класса $\mathcal{F}$ этот порядок остается справедливым, если $\mathcal{F}$

имеет «не слишком большую» сложность — этот результат очень подробно описывается в теории эмпирических процессов. Он имеет отношение к задаче Гливенко–Кантелли и к таким понятиям, как Донскеровские классы функций.

Вернемся к цепочке неравенств (12). Раз правая часть мала (порядка $O(n^{-1/2})$), то и избыточный риск функции $\hat{f}_n$ также мал. А значит нам не нужен супремум по всему классу $\mathcal{F}$ — нам достаточно брать супремум по подмножеству функций с достаточно малыми значениями избыточного риска: $\|P - P_n\|_{\mathcal{F}' \subset \mathcal{F}}$. Эта идея нашла свою реализацию в современных подходах SLT, которые мы затронем в конце доклада.

Так или иначе, даже после локализации супремума, нам снова придется иметь дело с оценками тех же эмпирических процессов $\|P - P_n\|_{\mathcal{F}'}$. Поэтому дальше мы сконцентрируемся на получении оценок нормы эмпирического процесса.

2 Лекция 2: рademахеровский процесс и вычислимые оценки избыточного риска

2.1 Рademахеровский процесс

Итак, нас интересует величина $P_n f - Pf$. Мы хотим контролировать размер этой величины равномерно по всему классу $\mathcal{F}$ и для этого рассматриваем $\|P_n - P\|_{\mathcal{F}}$. Далее мы обсудим, зависит ли последняя величина от «структуры» класса $\mathcal{F}$ и распределения P и как. И главный на данный момент вопрос — можем ли мы оценивать эту величину с помощью данных по крайней мере с точностью до константы.

При изучении подобных величин важную роль играет другой случайный процесс — **рademахеровский случайный процесс**, который определяется так:

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(Z_i),\ f \in \mathcal{F},$$

где ε_i — **рademахеровские случайные величины**, конечно, независимые от обучающей выборки. Глобальная **рademахеровская сложность** определяется следующим образом:

$$\sup_{f \in \mathcal{F}} |R_n(f)| \equiv \|R_n\|_{\mathcal{F}}.$$

У этой величины есть очень понятная интуитивная интерпретация — корреляция со случайным шумом. Если рademахеровская сложность велика, это означает, что в нашем классе есть функция, хорошо приближающая случайный шум. Это верный признак «слишком большой сложности» семейства и, как следствие, большого шанса переобучиться. Если класс функций состоит из одной константной функции, то мы получим $\|R_n(f)\|_{\mathcal{F}} \sim O(n^{-1/2})$. Такой порядок величины мы будем считать «малым». Если в классе есть все функции, то получим $\|R_n(f)\|_{\mathcal{F}} = 1$ и $O(1)$ будем считать «большим» значением.

Рademахеровский процесс в нашей задаче играет важную роль по следующей причине:

Лемма 2.1 (снова Симметризация)

$$\frac{1}{2}\mathsf{E}\mathsf{E}_\varepsilon \|R_n\|_{\mathcal{F}_c} \leqslant \mathsf{E}\|P - P_n\|_{\mathcal{F}} \leqslant 2\mathsf{E}\mathsf{E}_\varepsilon \|R_n\|_{\mathcal{F}},$$
$$\mathcal{F}_c = \{f - Pf \colon f \in \mathcal{F}\}.$$

Причем неравенство останется справедливым, если правую часть заменить на $2\mathsf{E}\|R_n\|_{\mathcal{F}_c}$.

Доказательство. Докажем правую часть неравенства.

$$\mathsf{E}\|P - P_n\|_{\mathcal{F}} = \mathsf{E}\sup_{f\in\mathcal{F}}|Pf - P_nf| = \tag{13}$$

$$= \mathsf{E}\sup_{f\in\mathcal{F}}|P_nf - \mathsf{E}'P'_nf| = \mathsf{E}\sup_{f\in\mathcal{F}}|\mathsf{E}'\{P_nf - P'_nf\}| \leqslant$$

$$\leqslant \mathsf{E}\,\mathsf{E}'\sup_{f\in\mathcal{F}}|P_nf - P'_nf| = \mathsf{E}\,\mathsf{E}'\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(Z_i) - f(Z'_i)\big)\Big| = \tag{14}$$

$$= \mathsf{E}_\varepsilon\mathsf{E}\,\mathsf{E}'\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big(f(Z_i) - f(Z'_i)\big)\Big| \leqslant$$

$$\leqslant \mathsf{E}_\varepsilon\mathsf{E}\,\mathsf{E}'\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Z_i)\Big| + \mathsf{E}_\varepsilon\mathsf{E}\,\mathsf{E}'\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Z'_i)\Big| = 2\mathsf{E}\mathsf{E}_\varepsilon\|R_n\|_{\mathcal{F}}$$

■

Левая часть неравенства доказывается аналогично. Итак, размеры радемахеровского и эмпирического процессов совпадают с точностью до константы. Более того, вместо радемахеровской случайной величины можно использовать любую величину с симметричным распределением — например, нормальную.

Зачем нам сводить задачу к рассмотрению радемахеровского процесса? Conditioning! Если фиксировать обучающую выборку, то у нас остается набор очень простых и хорошо изученных случайный величин. Остается открыть классический учебник и воспользоваться необходимыми оценками оттуда. В особенности, если вместо радемахеровских случайных величин мы используем нормальные. После этого остается вспомнить, что нам еще надо взять матожидание по обучающей выборке — это обычно не представляет особого труда.

Важно обсудить аналогию с прошлой симметризацией: за счет введения новой рандомизации мы избавились от Pf.

Далее перечислим ряд полезных свойств радемахеровской сложности.

Лемма 2.1 (Неравенство сжатия, contraction inequality (Talagrand)) *Пусть функции из $\mathcal{F}$ принимают значения в $[-1,1]$. Пусть задано липшицево отображение $\varphi\colon [-1,1]\to\mathbb{R}$ с константой 1, такое что $\varphi(0)=0$. Тогда для $\varphi\circ\mathcal{F} = \{\varphi\circ f\colon f\in\mathcal{F}\}$ выполнено:*

$$\mathsf{E}\|R_n\|_{\varphi\circ\mathcal{F}} \leqslant 2\mathsf{E}\|R_n\|_{\mathcal{F}}.$$

Во-первых, очевидно, что если φ липшицева с константой $L\neq 1$, то в правой части просто появится множитель L. Во-вторых, зачем это может понадобиться? Здесь надо вспомнить, что наш класс потерь — это суперпозиция множества классификаторов и функции потерь. Например, для случая регрессии с квадратичной функцией потерь, если множество ответов классификаторов и $\mathcal{Y}\subset\mathbb{R}$ ограничены, то $\ell(\cdot,\cdot)$ — липшицева. Неравенство сжатия позволяет избавиться от функции потерь и перейти к супремуму просто по множеству классификаторов.

Лемма 2.2 *Пусть множество функций $\mathcal{F} = \{f_1,\ldots,f_N\}$ — конечно. Введем обозначение:*

$$\mathrm{ac}(\mathcal{F}) = \Big\{\textstyle\sum_{i=1}^{N} c_jf_j(Z)\colon \sum_{j=1}^{N}|c_j|\leqslant 1, f_j\in\mathcal{F}\Big\}.$$

Тогда справедливо следующее равенство:

$$\|R_n\|_{\mathrm{ac}(\mathcal{F})} = \|R_n\|_{\mathcal{F}}.$$

Доказательство. Обозначим $C^* = \{\boldsymbol{c}\in\mathbb{R}^N\colon \sum_{i=1}^{N}|c_i|\leqslant 1\}$. Тогда:

$$\|R_n\|_{\mathrm{ac}(\mathcal{F})} = \sup_{f\in\mathrm{ac}(\mathcal{F})}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Z_i)\right| = \sup_{c\in C^*}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Big(\textstyle\sum_{j=1}^{N}c_jf_j(Z_i)\Big)\right| = \sup_{c\in C^*}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\varepsilon_ic_jf_j(Z_i)\right| =$$

$$= \sup_{c\in C^*}\left|\sum_{j=1}^{N}c_j\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_if_j(Z_i)\right)\right| = \sup_{c\in C^*}\left|\sum_{j=1}^{N}c_j\,\alpha_j\right|.$$

Супремум в последнем выражении достигается на векторе c с единицей в позиции k, где $k = \arg\max_j |\alpha_j|$ (и нулями в остальных позициях). Таким образом

$$\|R_n\|_{\mathrm{ac}(\mathcal{F})} = \sup_{j\in\{1,\dots,N\}} |\alpha_j| = \sup_{j\in\{1,\dots,N\}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_j(Z_i)\right| = \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Z_i)\right| = \|R_n\|_{\mathcal{F}}$$

■

Мы только что установили, что выпуклое замыкание конечного класса функций не увеличивает радемахеровскую сложность! Этот фундаментальный результат вместе с неравенством сжатия уже позволяет нам в ряде случаев при использовании выпуклых композиций алгоритмов с липшицевой функцией потерь оценивать радемахеровскую сложность: а) с помощью неравенства сжатия мы переходим от радемахеровской сложности класса потерь $\|R_n\|_{\mathcal{F}}$ к сложности самой выпуклой оболочки классификаторов $\|R_n\|_{\mathcal{G}}$; б) с помощью последнего результата переходим к сложности конечного класса функций $\|R_n\|_{\{g_1,\dots,g_N\}}$.

Для конечных классов функций справедлива следующая оценка:

Лемма 2.3 (конечный класс $\mathcal{F}$) *Пусть $\mathcal{F} = \{f_1,\dots,f_N\}$ — конечный класс функций, равномерно ограниченных константой $U > 0$. Обозначим $\sigma^2 = \sup_{f\in\mathcal{F}} Pf^2$. Тогда существует константа $K > 0$, такая что*

$$\mathsf{E}\|R_n\|_{\mathcal{F}} \leqslant K \max\left[\sigma\sqrt{\frac{\log N}{n}}, U\frac{\log N}{n}\right].$$

Крайне полезно привести здесь доказательство. Оно продемонстрирует ряд типичных техник работы с радемахеровским процессом, которые часто применяются в выводе подобных рехультатов.

Для доказательства леммы нам понадобится следующее определение. Будем говорить, что случайная величина X — **субгауссовская** с параметром σ^2, или $Y \in SG(\sigma^2)$, если для любой $\lambda \in \mathbb{R}$ справедливо:

$$\mathsf{E}e^{\lambda X} \leqslant e^{\lambda^2\sigma^2/2}. \tag{15}$$

В частности, нормальная случайная величина X с нулевым матожиданием и дисперсией σ^2 принадлежит классу $SG(\sigma^2)$. Более того, для нормальных случайных величин неравенство (15) превращается в равенство.

Оказывается, для субгауссовских случайных величин выполняется следующее важное свойство:

Утверждение 2.1 *Для любых случайных величин $X_1,\dots,X_N$, таких что $X_i \in SG(\sigma_i^2)$, $j = 1,\dots,N$, справедливо:*

$$\mathsf{E}\max_{i=1,\dots,N}|X_i| \leqslant C\max_{i=1,\dots,N}\sigma_i\sqrt{\log N},$$

где C — положительная константа.

Доказательство. Докажем лемму для случая $\mathsf{E}X_i = 0$, $i = 1,\dots,n$ и симметричнх плотностей распределений. Пусть $X \in SG(\sigma^2)$. Тогда для любого $t \in \mathbb{R}$

$$\mathsf{E}e^{tX} = \int_{-\infty}^0 e^{tx}p_X(x)\,dx + \frac{\mathbf{1}}{\mathbf{2}}\mathbf{E}\,\boldsymbol{e}^{\boldsymbol{t|X|}} \leqslant e^{t^2\sigma^2/2}. \quad \textbf{симметричя}$$

Следовательно

$$\mathsf{E}\,e^{t|X|} \leqslant 2e^{t^2\sigma^2/2} - 2\int_{-\infty}^0 e^{tx}p_X(x)\,dx \leqslant 2e^{t^2\sigma^2/2}.$$

Теперь

$$\exp\Big\{t\mathsf{E}\max_{i=1,\dots,n}|X_i|\Big\} \leqslant \mathsf{E}\exp\Big\{t\max_{i=1,\dots,n}|X_i|\Big\} = \mathsf{E}\max_{i=1,\dots,n}\exp\Big\{t|X_i|\Big\} \leqslant \sum_{i=1}^N \mathsf{E}\exp\Big\{t|X_i|\Big\} \leqslant$$
$$\leqslant 2N\max_{i=1,\dots,N}\exp\Big\{t^2\sigma_i^2/2\Big\} = 2N\exp\Big\{t^2\max_{i=1,\dots,N}\sigma_i^2/2\Big\}.$$

Логарифмируя обе части неравенства, для $t > 0$ получим:

$$\mathsf{E} \max_{i=1,\dots,n} |X_i| \leqslant \frac{\log 2N}{t} + \frac{t}{2} \max_{i=1,\dots,N} \sigma_i^2.$$

Возьмем $t = \frac{\sqrt{\log 2N}}{\max_{i=1,\dots,N} \sigma}$, что завершает доказательство утверждения. ■

Следующее утверждение показывает, что рademахеровский процесс при фиксированной выборке $X_1, \dots, X_n$ — субгауссовская случайная величина:

Утверждение 2.2 *Случайная величина*

$$\sqrt{n} R_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(X_i),\ f \in \mathcal{F}$$

принадлежит классу $SG(\sigma_f^2)$ *с параметром* $\sigma_f^2 = \|f\|^2_{L_2(P_n)} = \frac{1}{n}\sum_{i=1}^n f^2(X_i)$.

Доказательство. Обозначим $\xi = \sqrt{n} R_n(f)$. Тогда для любой $t \in \mathbb{R}$:

$$\mathsf{E} e^{t\xi} = \mathsf{E} \exp\Big\{ \frac{t}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(X_i) \Big\} = \prod_{i=1}^{n} \mathsf{E} \exp\Big\{ \frac{t}{\sqrt{n}} \varepsilon_i f(X_i) \Big\} \leqslant \prod_{i=1}^{n} \exp\Big\{ \frac{t^2 f^2(X_i)}{2n} \Big\} = \exp\Big\{ \frac{t^2}{2n} \sum_{i=1}^{n} f^2(X_i) \Big\}.$$

Отсюда вытекает утверждение. В цепочке неравенств мы воспользовались легко проверяемым фактом $\varepsilon_i \in SG(1)$. ■

Перейдем к доказательству леммы 2.3.

Доказательство. [Лемма 2.3] Из утверждение 2.1 и 2.2 следует, что

$$\mathsf{E}_\varepsilon \|R_n\|_{\mathcal{F}} \leqslant K \max_{f \in \mathcal{F}} \|f\|_{L_2(P_n)} \sqrt{\frac{\log N}{n}}. \tag{16}$$

Обозначим $\mathcal{F}^2 = \{f^2 \colon f \in \mathcal{F}\}$. Заметим, что

$$\sup_{f\in\mathcal{F}} \|f\|_{L_2(P_n)} = \sup_{f\in\mathcal{F}} \sqrt{\frac{1}{n}\sum_{i=1}^n f^2(X_i)} \leqslant \sup_{f\in\mathcal{F}} \sqrt{\left|\frac{1}{n}\sum_{i=1}^n f^2(X_i) - \mathsf{E}f^2\right| + \mathsf{E}f^2} \leqslant$$
$$\leqslant \sup_{f\in\mathcal{F}} \sqrt{\mathsf{E}f^2} + \sup_{f\in\mathcal{F}} \sqrt{|(P_n - P)f^2|} = \sup_{f\in\mathcal{F}} \|f\|_{L_2(P)} + \sqrt{\|P - P_n\|_{\mathcal{F}^2}}.$$

Взяв матожидания обеих частей неравенства, получим:

$$\mathsf{E} \max_{f\in\mathcal{F}} \|f\|_{L_2(P_n)} \leqslant \sigma + \mathsf{E}\sqrt{\|P - P_n\|_{\mathcal{F}^2}} \leqslant \sigma + \sqrt{\mathsf{E}\|P - P_n\|_{\mathcal{F}^2}}.$$

Неравенства симметризации и сжатия дают нам:

$$\mathsf{E}\|P - P_n\|_{\mathcal{F}^2} \leqslant 2\mathsf{E}\|R_n\|_{\mathcal{F}} \leqslant 8U\mathsf{E}\|R_n\|_{\mathcal{F}}.$$

С учетом (16), мы получаем:

$$\mathsf{E}\|R_n\|_{\mathcal{F}} \leqslant K\mathsf{E} \max_{f\in\mathcal{F}} \|f\|_{L_2(P_n)} \sqrt{\frac{\log N}{n}} \leqslant K\left(\sigma + \sqrt{8U\mathsf{E}\|R_n\|_{\mathcal{F}}}\right)\sqrt{\frac{\log N}{n}}.$$

Решая последнее неравенство относительно $\mathsf{E}\|R_n\|_{\mathcal{F}}$, получим:

$$\mathsf{E}\|R_n\|_{\mathcal{F}} \leqslant K_1\sigma\sqrt{\frac{\log N}{n}} + K_2 U \frac{\log N}{n}$$

для положительных констант K_1 и K_2. ■

Рассмотрим задачу классификации с бинарной функцией потерь. Приведем в этом случае без доказательства еще одну оценку величины $\mathsf{E}\|R_n\|_{\mathcal{F}}$ для бесконечного класса функций $\mathcal{F}$ в терминах знакомого нам **коэффициента разнообразия** (shattering number):

$$\Delta^{\mathcal{F}}(Z_n, \dots, Z_n) = \left|\left\{\big(f(Z_1), \dots, f(Z_n)\big) \colon f \in \mathcal{F}\right\}\right|.$$

Лемма 2.4 *Пусть $\mathcal{F}$ — бесконечный класс равномерно ограниченных функций, соответствующий бинарной функции потерь. Обозначим $\sigma^2 = \sup_{f\in\mathcal{F}} Pf^2$. Тогда существует константа $K > 0$, такая что*

$$\mathsf{E}\|R_n\|_{\mathcal{F}} \leqslant K \max\left[\sigma\mathsf{E}\sqrt{\frac{\log \Delta^{\mathcal{F}}(Z_1,\ldots,Z_n)}{n}}, \mathsf{E}\frac{\log \Delta^{\mathcal{F}}(Z_1,\ldots,Z_n)}{n}\right].$$

2.2 Неравенства концентрации меры

Предположим, математическое ожидание радемахеровской сложности $\|R_n\|_{\mathcal{F}}$, а значит и супнормы эмпирического процесса мы научились оценивать. Как пользуясь этими оценками контролировать саму норму эмпирического процесса $\|P - P_n\|_{\mathcal{F}}$? Оказывается, связь значения случайной величины с ее матожиданием может быть описана достаточно универсальным способом, не использующим «структурные» характеристики класса функций $\mathcal{F}$. Сложностные характеристики класса потерь — будь то размерность Вапника–Червоненкиса или что-то еще — содержатся в оценках $\mathsf{E}\|R_n\|_{\mathcal{F}}$. В теории вероятностей существуют инструменты, позволяющие достаточно точно и с высокой вероятностью оценивать отклонение случайной величины от ее матожидания. Это делается с помощью неравенств, называемых **неравенствами концентрации меры** (concentration inequalities). Простейшим примером таких неравенств являются неравенства Хевдинга и Бернштейна, оценивающие с вероятностью не менее $1-\delta$ сверху $\mathsf{E}Z - \frac{1}{n}\sum_{i=1}^n Z_i$ величинами:

$$\sqrt{\frac{2\log\frac{1}{\delta}}{n}} \text{ (н-во Хевдинга);}$$

$$\sqrt{\frac{2\mathrm{Var}Z\log\frac{1}{\delta}}{n}} + \frac{2\log\frac{1}{\delta}}{3n} \text{ (н-во Бернштейна).}$$

Эти оценки справедливы для суммы независимых случайных величин, принимающих значения из ограниченных интервалов.

Оказывается, похожие результаты могут быть получены и для более общих функций случайных выборок. Одно из них — **неравенство МакДиармида** или неравенство ограниченных разностей. Будем говорить, что функция $f\colon \mathcal{Z}^n \to \mathbb{R}$ имеет **ограниченные разности** с параметрами $c_1,\ldots,c_n$, если

$$\sup_{\substack{Z_1,\ldots,Z_n,\\ Z_i'\in\mathcal{Z}}} |f(Z_1,\ldots,Z_n) - f(Z_1,\ldots,Z_{i-1},Z_i',Z_{i+1},\ldots,Z_n)| \leqslant c_i,\ i=1,\ldots,n.$$

Оказывается, справедливо следующее утверждение, обобщающее неравенство Хевдинга на класс функций с ограниченными разностями:

Теорема 2.1 (неравенство ограниченных разностей, McDiarmid) *Пусть функция f имеет ограниченные разности с параметрами $c_1,\ldots,c_n$, а $\{Z_1,\ldots,Z_n\}$ — независимые случайные величины. Тогда для случайной величины $Z = f(Z_1,\ldots,Z_n)$ выполнено:*

$$\mathsf{P}\{|Z - \mathsf{E}Z| > t\} \leqslant 2e^{-t^2/C},$$

где $C = \sum_{i=1}^n c_i^2$.

Сумма очевидно обладает свойством ограниченных разностей. Заметим, что полученная оценка имеет в точности такой же вид, как для случая одной случайной величины.

Теперь заметим, что случайная величина $\|P - P_n\|_{\mathcal{F}}$ в случае равномерно ограниченного константой U класса функций имеет ограниченные разности с параметрами $c_i = \frac{2U}{n}$, $i = 1,\ldots,n$. Таким образом справедливо следующее утверждение:

$$\mathsf{P}\Big\{\Big|\|P - P_n\|_{\mathcal{F}} - \mathsf{E}\|P - P_n\|_{\mathcal{F}}\Big| \geqslant t\Big\} \leqslant 2\exp\Big(-\frac{t^2 n}{4U^2}\Big). \tag{17}$$

Радемахеровская сложность $\|R_n\|_{\mathcal{F}}$ удовлетворяет тем же требованиям, что и $\|P - P_n\|_{\mathcal{F}}$. С учетом неравенства симметризации и оценки (17) мы можем получить следующее утверждение:

Лемма 2.5 *Для всех $t > 0$ справедливо:*

$$\mathsf{P}\Big\{\|P - P_n\|_{\mathcal{F}} \geqslant 2\|R_n\|_{\mathcal{F}} + \frac{3tU}{\sqrt{n}}\Big\} \leqslant e^{-t^2/2}. \tag{18}$$

Доказательство этого утверждения может быть получено применением неравенства ограниченных сумм для случайной величины $Z = \|P_n - P\|_{\mathcal{F}} - 2\|R_n\|_{\mathcal{F}}$ и неравенства $\mathsf{E}Z \leqslant 0$, следующего из симметризации.

Итак, мы получили нашу первую оценку, вычислимую по данным — действительно, величина $\|R_n\|_{\mathcal{F}}$ зависит только от обучающей выборки и последовательности случайных рademахеровских величин $\varepsilon_1, \ldots, \varepsilon_n$ — все что нам нужно сделать, это n раз подкинуть монету, тем самым получив ε_i, и вычислить максимальную корреляцию функций из класса $\mathcal{F}$ с этой последовательностью. Используя эту оценку, например, в (7), мы уже можем получить вычислимую по данным оценку обобщающей способности (правда, все еще грубую из-за супремума по всему классу функций).

В середине 90х французский математик Мишель Талагран (Talagrand) доказал следующее обобщение неравенства Бернштейна на эмпирические процессы:

Теорема 2.2 (неравенство Талаграна) *Пусть $Z_1, \ldots, Z_n$ — независимые случайные величины. Для любого равномерно ограниченного константой U класса функций $\mathcal{F}$ и для всех $t > 0$ справедливо:*

$$\mathsf{P}\Big\{n\Big|\|P_n - P\|_{\mathcal{F}} - \mathsf{E}\|P - P_n\|_{\mathcal{F}}\Big| \geqslant t\Big\} \leqslant K\exp\Big\{-\frac{1}{K}\frac{t}{U}\log\Big(1 + \frac{tU}{V}\Big)\Big\}$$

где $V = \mathsf{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n f^2(Z_i)$.

Для сравнения — неравенство Бернштейна:

$$\mathsf{P}\{Pf - P_n f \geqslant t\} \leqslant \exp\Big(-\frac{nt^2}{2\mathrm{Var}f - 2t/3}\Big).$$

Итак, при малых значениях $t \ll \frac{\mathrm{Var}}{U}$ мы имеем субгауссовское поведение, в то время как для больших значений — субэкспоненциальное. Таким образом мы контролируем отклонения в терминах смеси гауссовских и экспоненциальных хвостов. Неравенства Хевдинга и МакДиармида оценивают эти хвосты одним гауссовским.

Этот результат привел к активному развитию новых подходов в теории статистического обучения, позволивших изящно избавляться от необходимости брать супремум по всему классу функций при оценке избыточного риска и получать более точные оценки.

Заметим, что в отличие от неравенства МакДиармида, неравенство Талаграна позволяет нам учитывать дисперсии функций в классе $\mathcal{F}$. Чтобы сделать эту зависимость явной, воспользуемся следующим неравенством, полученным нами в доказательстве леммы 2.3:

$$\mathsf{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n f^2(X_i) \leqslant \sigma^2(\mathcal{F}) + 8U\|R_n\|_{\mathcal{F}},$$

где мы положили $\sigma^2(\mathcal{F}) = \sup_{f\in\mathcal{F}} Pf^2$. Отсюда следует, что с вероятностью не меньше $1 - e^{-t}$ справедливо:

$$\Big|\|P - P_n\|_{\mathcal{F}} - \mathsf{E}\|P - P_n\|_{\mathcal{F}}\Big| \leqslant K\Big(\sqrt{\frac{t}{n}\big(\sigma^2(\mathcal{F}) + U\mathsf{E}\|R_n\|_{\mathcal{F}}\big)} + \frac{tU}{n}\Big).$$

В общем случае с произвольным распределением P и функцией потерь $\ell(\cdot,\cdot)$, учет дисперсии может не привести к улучшению получаемых оценок. Однако, представим, что нам удалось некоторым образом ограничить дисперсии (нецентрированную) Pf^2 функций класса $\mathcal{F}$ сверху их избыточными рисками $\mathcal{E}f$. Тогда при использовании локализации $\sup_{f\in\mathcal{F}(\delta)}|(P - P_n)f|$, о которой мы сказали в конце прошлой лекции, мы ограничиваем множество, по которому берется супремум, функциями с малыми значениями избыточного риска $\mathcal{E}(f)$. А значит, дисперсии функций из множества поиска тоже малые! Это приведет к улучшению оценки именно за счет учета дисперсий класса функций в неравенстве Талаграна. Заметим, что использование неравенств типа Хевдинга принципиально не позволило бы получить подобные результаты.

3 Лекция 3: локализация и оценки избыточного риска

Итак, в конце первой лекции мы коротко описали один из способов улучшения точности оценок — переход от супремума эмпирического процесса по всему семейству $\sup_{f,g\in\mathcal{F}}|(P-P_n)(f-g)|$ к его локальной версии $\sup_{f,g\in\mathcal{F}(\delta)}|(P-P_n)(f-g)|$ по подмножеству функций с малыми избыточными рисками. Локальная супнорма в свою очередь с точностью до константы совпадает с локальной радемахеровской сложностью $\|R_n\|_{\mathcal{F}(\delta)}$, о чем свидетельствует неравенство симметризации. Способы оценки радемахеровских сложностей и их матожиданий мы рассмотрели на прошлой лекции, закончив формулировкой неравенства Талаграна и кратко отметив пользу учета равномерного аналога дисперсии $\sup_{f\in\mathcal{F}}Pf^2$.

3.1 Неформальное описание техники

Мы наконец можем подробно изложить методику построения оценок для избыточного риска минимизатора эмпирического риска $\mathcal{E}(\hat{f}_n)$, в большинстве случаев оказывающихся правильного порядка малости. Мы приведем несколько эвристическое описание техники, оставив формальные доказательства всех шагов в стороне.

Всюду далее будем считать, что функции из класса $\mathcal{F}$ принимают значения в $[0,1]$.

Начнем с введения определений. Как мы уже говорили, мы будем рассматривать δ-**минимальное** подмножество класса $\mathcal{F}$, определяемое

$$\mathcal{F}(\delta)=\{f\in\mathcal{F}\colon \mathcal{E}(f)\leqslant\delta\}.$$

Для нас будет важен $L_2(P)$-диаметр δ-минимального множества

$$D(\delta)\equiv D_P(\mathcal{F},\delta)=\sup_{f,g\in\mathcal{F}(\delta)}\sqrt{P(f-g)^2},$$

а также матожидание супнормы следующего эмпирического процесса:

$$\varphi_n(\delta)=\mathsf{E}\sup_{f,g\in\mathcal{F}(\delta)}|(P-P_n)(f-g)|,$$

которое отражает точность аппроксимации распределения P эмпирическим распределением P_n. Наконец, определим с помощью этих величин с первого взгляда сложную функцию

$$U_n(\delta,t)=2\Big(\varphi_n(\delta)+D(\delta)\sqrt{\frac{t}{n}}+\frac{t}{n}\Big). \tag{19}$$

Напомним, что мы будем пользоваться следующей формой неравенством Талаграна (форма Bousquet)

$$\mathsf{P}\Big\{\|P-P_n\|_{\mathcal{F}}\geqslant\mathsf{E}\|P-P_n\|_{\mathcal{F}}+\sqrt{2\frac{t}{n}\left(\sup_{f\in\mathcal{F}}\big(Pf^2-(Pf)^2\big)+2\mathsf{E}\|P-P_n\|_{\mathcal{F}}\right)}+\frac{t}{3n}\Big\}\leqslant e^{-t}.$$

С учетом элементарных неравенств $\sqrt{a+b}\leqslant\sqrt{a}+\sqrt{b}$ и $2\sqrt{ab}\leqslant a+b$ неравенство Талаграна дает нам следующую оценку:

$$\mathsf{P}\Big\{\sup_{f,g\in\mathcal{F}(\delta)}|(P-P_n)(f-g)|\geqslant U_n(\delta,t)\Big\}\leqslant e^{-t}. \tag{20}$$

Итак, вид функции (19) был выбран неслучайно — он продиктован неравенством Талаграна.

Коротко обсудим на неформальном уровне важность всех понятий. Мы занимаемся оценкой случайной величины $\sup_{f,g\in\mathcal{F}(\delta)}|(P-P_n)(f-g)|$. Следуя второй лекции мы обращаемся к ее матожиданию $\mathsf{E}\sup_{f,g\in\mathcal{F}(\delta)}|(P-P_n)(f-g)|$. Величина, стоящая под супремумом, в типичных ситуациях имеет порядок малости $\frac{1}{\sqrt{n}}$. Предположим теперь, что с уменьшением δ $L_2(P)$-диаметр $D(\delta)$ δ-минимального множества убывает к нулю (это может происходить, например, когда Pf, $f\in\mathcal{F}$, имеет уникальный глобальный минимум — в этом случае диаметр δ-минимального множества стремится к нулю с уменьшением δ). Поскольку это означает, что дисперсия функции $f-g$ убывает к нулю,

величина под матожиданием становится еще меньшего порядка при одновременном стремлении $n \to \infty$ и $\delta \to 0$, чем $\frac{1}{\sqrt{n}}$.

Таким образом оценка (19) во многих случаях имеет порядок меньший, чем $\frac{1}{\sqrt{n}}$. В случае, когда минимумов риска несколько, или диаметр δ-минимального множества не убывает к нулю по другим причинам, оценка остается справедливой — однако в этом случае она дает порядок $\frac{1}{\sqrt{n}}$.

Теперь возникает вопрос — какое δ выбирать для получения оценок избыточного риска $\mathcal{E}(\hat{f}_n)$? Следующие рассуждения отвечают на этот вопрос. Вспомним нашу простую цепочку неравенств

$$\mathcal{E}(\hat{f}_n) \leqslant \sup_{f,g\in\mathcal{F}} |(P - P_n)(f - g)|.$$

Положим $\delta_n^{(0)} = 1$. Учитывая тождество $\mathcal{F}\big(\delta_n^{(0)}\big) = \mathcal{F}(1) = \mathcal{F}$ и оценку (20), с большой вероятностью, описываемой неравенством Талаграна, мы получаем

$$\mathcal{E}(\hat{f}_n) \leqslant \sup_{f,g\in\mathcal{F}} |(P - P_n)(f - g)| \leqslant U_n(\delta_n^{(0)}, t),$$

где неравенство Талаграна мы используем во втором знаке неравенства. Положим $\delta_n^{(1)} = U_n(\delta_n^{(0)}, t)$. В типичных случаях величина $\delta_n^{(1)}$ имеет порядок $O(\frac{1}{\sqrt{n}})$. Последнее неравенство дает нам возможность уточнить оценку и записать

$$\mathcal{E}(\hat{f}_n) \leqslant \sup_{f,g\in\mathcal{F}(\delta_n^{(1)})} |(P - P_n)(f - g)| \leqslant U_n(\delta_n^{(1)}, t).$$

Важно отметить, что при совершении каждой такой итерации мы пользуемся неравенством Талаграна, а значит каждый раз немного возрастает вероятность, что оценка не будет справедлива. Мы снова обозначаем $\delta_n^{(2)} = U_n(\delta_n^{(1)}, t)$. Эта величина уже будет иметь порядок меньший, чем $\frac{1}{\sqrt{n}}$, если $D(\delta) \to 0$ при $\delta \to 0$. Затем мы снова повторяем итерации

$$\mathcal{E}(\hat{f}_n) \leqslant \sup_{f,g\in\mathcal{F}(\delta_n^{(2)})} |(P - P_n)(f - g)| \leqslant U_n(\delta_n^{(1)}, t) \equiv \delta_n^{(3)}$$

и так далее, получая убывающую последовательность оценок $\{\delta_n^{(i)}\}_i$, которая сходится при $i \to \infty$ к неподвижной точке функции $U_n(\delta, t)$:

$$\delta_n^*(t) = U_n(\delta_n^*(t), t).$$

Величина $\delta_n^*(t)$ также является оценкой избыточного риска $\mathcal{E}(\hat{f}_n)$, и, оказывается, что в большинстве рассматриваемых в теории случаев эта оценка имеет оптимальный порядок малости.

Важным вопросом является следующий. На каждой итерации описанного процесса мы теряем небольшую вероятность в результате применения неравенства Талаграна. Оказывается, справедлив результат, показывающий, что итерации сходятся к неподвижной точке настолько быстро, что накопленная вероятность остается достаточно малой.

Введем следующие $\flat$ и $\sharp$-преобразования функции $\psi\colon \mathbb{R}^+ \to \mathbb{R}^+$:

$$\psi^\flat(\delta) = \sup_{\sigma\geqslant\delta} \frac{\psi(\sigma)}{\sigma};$$
$$\psi^\sharp(\varepsilon) = \inf\big\{\delta\colon \psi^\flat(\delta) \leqslant \varepsilon\big\}.$$

Смысл введенных величин становится понятен в случае, когда $\psi(\delta)$ выпукла (а именно для таких функций в большинстве случаев применяются описываемые результаты). В этом случае $\psi^\flat(\delta) = \psi(\delta)/\delta$, а $\psi^\sharp(1)$ будет в точности решением уравнения $\psi(\delta) = \delta$. В ряде случаев вместо поиска решения уравнения $U_n(\delta, t) = \delta$ приходится решать уравнения более общего вида $U_n(\delta, t) = c\delta$, где c — константа. Более того, рассматриваемые функции не всегда оказываются выпуклыми по δ и приходится оценивать их сверху выпуклыми мажорантами. Именно для таких случаев мы ввели эти обозначения.

Определим $\bar{\delta}_n(t) = U_n^\sharp(2^{-1}, t) \equiv U_{n,t}^\sharp(2^{-1})$. Грубо говоря, это решение уравнения $U_n(\delta, t) = \frac{1}{2}\delta$. Следующая теорема — результат, обосновывающий рассуждения из этого раздела и показывающий, что величина $\bar{\delta}_n(t)$ является естественным порогом в рассматриваемой задаче:

Теорема 3.1 *Для любых $\delta > \bar{\delta}_n(t)$ справедливо:*

$$\mathsf{P}\big\{\mathcal{E}(\hat{f}_n) \geqslant \delta\big\} \leqslant \log_2(2/\delta)\, e^{-t}.$$

3.2 Доказательства и некоторые частные случаи

Мы докажем результат, отличающийся в незначительных деталях от 3.1. Предположим, что $\{\delta_j\}_j$ — убывающая последовательность положительных чисел, $\delta_0 = 1$. Пусть $\{t_j\}_j$ — последовательность положительных чисел. Для $\delta \in (\delta_{j+1}, \delta_j]$ положим

$$U_n(\delta) = \varphi_n(\delta_j) + \sqrt{2\frac{t_j}{n}\big(D^2(\delta_j) + 2\varphi_n(\delta_j)\big)} + \frac{t_j}{2n},$$

где по-прежнему

$$\varphi_n(\delta) = \mathsf{E} \sup_{g,f\in\mathcal{F}(\delta_j)} |(P - P_n)(f - g)|;$$
$$D(\delta) = \mathsf{E} \sup_{g,f\in\mathcal{F}(\delta_j)} P(f - g)^2.$$

Обозначим $\delta_n(\mathcal{F}, P) = \sup\{\delta \in (0, 1]\colon \delta \leqslant U_n(\delta)\}$. Тогда справедлива следующая теорема:

Теорема 3.2 *Для всех $\delta \geqslant \delta_n(\mathcal{F}, P)$ выполнено*

$$\mathsf{P}\{\mathcal{E}(\hat{f}_n) > \delta\} \leqslant \sum_{\delta_j \geqslant \delta} e^{-t_j}.$$

Заметим, что этот результат в точности повторяет рассуждение прошлого раздела. Он утверждает, что любая величина δ превосходящая естественный порог $\delta_n(\mathcal{F}, P)$ является оценкой избыточного риска $\mathcal{E}(\hat{f}_n)$ с вероятностью, уменьшающейся с каждым применением неравенства Талаграна, потребовавшимися нам для «достижения» значения δ. Полезно нарисовать на доске, что утверждается в теореме: обратная функция распределения $F(\delta)$ мажорируется ступенчатой функцией от δ.

Доказательство. Предположим, что $\delta > \delta_n(\mathcal{F}, P)$ (в противном случае результат будет справедлив вследствие непрерывности справа функций распределения. Обозначим $\hat{\delta} \equiv \mathcal{E}(\hat{f}_n)$. Тогда если $\hat{\delta} \geqslant \delta \geqslant \varepsilon > 0$ для некоторой ε и если $g \in \mathcal{F}(\varepsilon)$, то

$$\hat{\delta} = P\hat{f}_n - \inf_{f\in\mathcal{F}} \leqslant P(\hat{f}_n - g) + \varepsilon \leqslant$$
$$\leqslant P_n(\hat{f}_n - g) + (P - P_n)(\hat{f}_n - g) + \varepsilon \leqslant \sup_{f,g\in\mathcal{F}(\delta)} |(P - P_n)(f - g)| + \varepsilon.$$

Обозначим для краткости $\|P - P_n\|_{\mathcal{F}'(\delta)} \equiv \sup_{f,g\in\mathcal{F}(\delta)} |(P - P_n)(f - g)|$, имея ввиду, что $\mathcal{F}'(\delta) = \{f - g\colon f, g \in \mathcal{F}(\delta)\}$. Устремив $\varepsilon \to 0$, приходим к выводу:

$$\hat{\delta} \leqslant \|P - P_n\|_{\mathcal{F}'(\hat{\delta})}. \tag{21}$$

Введем следующее событие:

$$E_{n,j} \equiv \left\{\|P - P_n\|_{\mathcal{F}'(\delta_j)} \leqslant U_n(\delta_j)\right\},$$

для которого из неравенства Талаграна следует $\mathsf{P}\{E_{n,j}\} \geqslant 1 - e^{-t_j}$. Нас интересует пересечение событий:

$$E_n \equiv \bigcap_{\delta_j \geqslant \delta} E_{n,j},$$

для которого очевидно справедливо

$$\mathsf{P}\{E_n\} \geqslant 1 - \sum_{\delta_j \geqslant \delta} e^{-t_j}.$$

Из определения функции $U_n(\delta)$ и монотонности функции $\delta \to \|P - P_n\|_{\mathcal{F}'(\delta)}$ следует, что на событии E_n для всех $\sigma \geqslant \delta$ выполнено $\|P - P_n\|_{\mathcal{F}'(\sigma)} \leqslant U_n(\sigma)$. Следовательно, на событии $E_n \cap \{\hat{\delta} \geqslant \delta\}$ с учетом

$$\hat{\delta} \leqslant \|P - P_n\|_{\mathcal{F}'(\hat{\delta})} \leqslant U_n(\hat{\delta})$$

мы получаем $\delta \leqslant \hat{\delta} \leqslant \delta_n(\mathcal{F}, P)$ и приходим к противоречию. Таким образом событие $\{\hat{\delta} \geqslant \delta\}$ должно быть вложено в дополнение к E_n, откуда следует утверждение теоремы. ■

При использовании последовательности $\{\delta_j\}_j$ следующего вида: $\delta_j = 2^{-j}$ и $t_j \equiv t$, $j \geqslant 0$ мы получим результат, аналогичный приведенному в 3.1.

Регрессия с квадратичными потерями. Продемонстрируем на примере задачи восстановления регрессии с квадратичной функцией потерь, как можно ограничивать $L_2(P)$-диаметр δ-минимального множества. Сформулируем следующую теорему:

Теорема 3.3 *Пусть множество ответов Y и выпуклое множество функций $\mathcal{G}$ принимают значения в $[0,1]$. Тогда справедливо:*

$$D(\delta) \equiv \sup_{f,g \in \mathcal{F}(\delta)} \sqrt{P(f-g)^2} \leqslant 4\sqrt{2\delta}.$$

Предположим, что множество ответов Y принимает значения в $[0,1]$ и $\ell(y,y^*) = (y-y^*)^2$. Минимум риска

$$P(\ell \bullet g) = \mathsf{E}\big(Y - g(X)\big)^2$$

достигается на функции регрессии

$$\eta(x) = \mathsf{E}(Y|X=x).$$

Пусть $\mathcal{G}$ — класс функций, принимающих значения в $[0,1]$, и $\eta \in \mathcal{F}$. Обозначим маргинальное распределение на пространстве объектов X буквой Π. Тогда справедливо следующее утверждение:

Лемма 3.1

$$\mathcal{E}(\ell \bullet g) = \|g-\eta\|^2_{L_2(\Pi)} = \int_X \big(g(X)-\eta(X)\big)^2 d\Pi \equiv \mathsf{E}_X\big(g(X)-\eta(X)\big)^2,$$

Доказательство.

$$\mathcal{E}(\ell \bullet g) = \mathsf{E}_{X,Y}\big(Y-g(X)\big)^2 - \mathsf{E}_{X,Y}\big(Y-\mathsf{E}(Y|X)\big)^2 =$$
$$= \mathsf{E}_{X,Y}\big(Y-g(X)\big)^2 - \mathsf{E}_{X,Y}\big(Y-\mathsf{E}(Y|X)\big)^2 - \mathsf{E}_X\big(\mathsf{E}(Y|X)-g(X)\big)^2 + \mathsf{E}_X\big(\mathsf{E}(Y|X)-g(X)\big)^2.$$

Теперь заметим, что

$$\mathsf{E}_{X,Y}\big(Y-g(X)\big)^2 - \mathsf{E}_{X,Y}\big(Y-\mathsf{E}(Y|X)\big)^2 - \mathsf{E}_X\big(\mathsf{E}(Y|X)-g(X)\big)^2 =$$
$$= 2\mathsf{E}_{X,Y}\big(\mathsf{E}(Y|X)g - Yg\big) + 2\mathsf{E}_{X,Y}\big(\mathsf{E}(Y|X)Y - \mathsf{E}(Y|X)\mathsf{E}(Y|X)\big) =$$
$$= 2\mathsf{E}_X\mathsf{E}_{Y|X}\big(\mathsf{E}(Y|X)g - Yg\big) + 2\mathsf{E}_X\mathsf{E}_{Y|X}\big(\mathsf{E}(Y|X)Y - \mathsf{E}(Y|X)\mathsf{E}(Y|X)\big) = 0.$$

■

Легко проверить, что в общем случае $\eta \notin \mathcal{G}$, избыточный риск функции g записывается в следующем виде:

$$\mathcal{E}(\ell \bullet g) = \|g-\eta\|^2_{L_2(\Pi)} - \inf_{h \in \mathcal{G}} \|h-\eta\|^2_{L_2(\Pi)}.$$

Рассмотрим случай **выпуклого** класса $\mathcal{G}$. Обозначим $\bar{g} = \arg\min_{g\in\mathcal{G}} \|g-\eta\|^2_{L_2(\Pi)}$. Тогда справедливо следующее утверждение об избыточном риске функции g:

Лемма 3.2 *Если $\mathcal{G}$ — выпуклый, то*

$$2\,\mathcal{E}(\ell \bullet g) \geqslant \|g-\bar{g}\|^2_{L_2(\Pi)}.$$

Доказательство. Воспользуемся следующим тождеством:

$$\frac{u^2+v^2}{2} - \left(\frac{u+v}{2}\right)^2 = \frac{(u-v)^2}{4}$$

и получим

$$\frac{(g-\eta)^2 + (\bar{g}-\eta)^2}{2} = \left(\frac{g+\bar{g}}{2} - \eta\right)^2 + \frac{(g-\bar{g})^2}{4}.$$

Проинтегрируем по Pi обе части равенства:

$$\frac{\|g-\eta\|^2_{L_2(\Pi)} + \|\bar{g}-\eta\|^2_{L_2(\Pi)}}{2} = \left\|\frac{g+\bar{g}}{2} - \eta\right\|^2_{L_2(\Pi)} + \frac{\|g-\bar{g}\|^2_{L_2(\Pi)}}{4}.$$

Учитывая определение $\bar g$ и выпуклость класса $\mathcal{G}$, откуда следует $\frac{g|\bar g}{2} \in \mathcal{G}$, получим:

$$\left\| \frac{g+\bar g}{2} - \eta \right\|^2_{L_2(\Pi)} \geqslant \|\bar g - \eta\|^2_{L_2(\Pi)},$$

откуда следует утверждение теоремы. ■

Как и раньше обозначим $\mathcal{F} = \{\ell \bullet g \colon g \in \mathcal{G}\}$. Тогда из леммы следует, что

$$\mathcal{F}(\delta) \subset \{\ell \bullet g \colon \|g - \bar g\|^2_{L_2(\Pi)} \leqslant 2\delta\}.$$

Также для любых $g_1, g_2 \in \mathcal{G}$ выполнено

$$|(\ell \bullet g_1)(x,y) - (\ell \bullet g_2)(x,y)| = \left|\big(y - g_1(x)\big)^2 - \big(y - g_2(x)\big)^2\right| =$$
$$= |g_1(x) - g_2(x)||2y - g_1(x) - g_2(x)| \leqslant 2|g_1(x) - g_2(x)|.$$

интегрируя, получим

$$P\big(\ell \bullet g_1 - \ell \bullet g_2\big)^2 \leqslant 4\|g_1 - g_2\|^2_{L_2(\Pi)}.$$

Окончательно, мы получаем утверждение теоремы 3.3:

$$D(\delta) = \sup_{f,g \in \mathcal{F}(\delta)} \sqrt{P(f-g)^2} \leqslant$$
$$\leqslant 2\sup\Big\{\|g_1 - g_2\|_{L_2(\Pi)} \colon \|g_1 - \bar g\|^2_{L_2(\Pi)} \leqslant 2\delta, \|g_2 - \bar g\|^2_{L_2(\Pi)} \leqslant 2\delta\Big\} \leqslant 4\sqrt{2\delta}.$$

3.3 Дальнейшие результаты, открытые вопросы и обсуждение

- Оказывается, возможно получение вычислимого по данным (data dependant) аналога оценки 3.1. Он основан на использовании симметризации и переходу к рademахеровской сложности, которая зависит исключительно от данных и оценке δ-минимальных подмножеств $\mathcal{F}(\delta)$ их эмпирическим вариантом $\mathcal{F}_n(\delta) = \{f \in \mathcal{F} \colon P_n f - \inf_{g \in \mathcal{F}} P_n g \leqslant \delta\}$.
- В случае нескольких минимумов риска Pf есть результаты, позволяющие несмотря на $D(\delta) \not\to 0$ при $\delta \to 0$ получать оценки, зависящие от распределения (distribution dependant), аналогично 3.1 дающие порядок малости $o(1/\sqrt{n})$. Однако в этом случае пока что не найден способ получать вычислимый по данным аналог таких оценок.
- Открытым вопросом является уточнение всех констант, использовавшихся в вычислениях и теоремах. В ряде случаев уже получены оптимальные константы.
- Подход, основанный на локализации, в рамках комбинаторной теории переобучения.

Материал подготовлен на основе лекций В. И. Колчинского [4], лекций и обзоров S. Boucheron, O. Bousquet и G. Lugosi [1], [3], [2] и книги [5].

Список литературы

[1] *Boucheron S., Bousquet O., Lugosi G.* Classification: a survey of recent advances // ESAIM: Probability and Statistics, 9. — 2005. — Pp. 323–375.

[2] *Boucheron S., Lugosi G., Bousquet O.* Concentration inequalities // Advanced Lectures in Machine Learning. — Springer, 2004. — Pp. 208–240.

[3] *Bousquet O., Boucheron S., Lugosi G.* Introduction to statistical learning theory // Advanced Lectures in Machine Learning. — Springer, 2004. — Pp. 169–207.

[4] *Koltchinskii V.* Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics. — Springer, 2011.

[5] *van der Vaart A., Wellner J.* Weak Convergence and Empirical Processes: With Applications to Statistics (Springer Series in Statistics). — Springer, 2000.