

# Теория статистического обучения

Н. К. Животовский

`nikita.zhivotovskiy@phystech.edu`

4 марта 2014 г.

Материал находится в стадии разработки, может содержать ошибки и неточности. Автор будет благодарен за любые замечания и предложения, направленные по указанному адресу

## 1 Введение

Цель данного курса лекций заключается в доступном изложении основных результатов теории статистического обучения (Statistical learning theory – ‘SLT’). Систематическое исследование теоретических основ вопросов машинного обучения привело к созданию теории статистического обучения и началось около 30 лет назад с работ Вапника и Червоненкиса. Одним из основных преимуществ разработанной ими теории была независимость основных результатов от того, по какому закону распределены данные. Таким образом, был осуществлен переход от подхода, ориентированного на модель данных (статистический подход), к подходу, заключающемуся в анализе в первую очередь методов обучения. Вторым важным шагом было получение необходимых и достаточных условий для равномерной по классу гипотез сходимости частот к вероятностям. Теперь процесс обучения можно контролировать вне зависимости от распределения данных и даже сложной процедуры выбора алгоритма из семейства. Общность подхода, конечно, имела очевидные недостатки, многие из которых были ликвидированы в последнее десятилетие. Катализатором исследований были два вероятностных раздела – теория эмпирических процессов и неравенства концентрации меры. Перейдем теперь к постановке задачи.

### §1.1 Постановка задачи

Начнем с так называемого обучения с учителем (supervised learning). Предположим, что существует множество объектов  $\mathcal{X}$  (объекты принято отождествлять с их признаковыми описаниями) и множество ответов  $\mathcal{Y}$ . Последнее, например, в случае задачи классификации на два класса может состоять всего из двух элементов (классы 1 и  $-1$ ) или в случае задачи регрессии совпадать со множеством действительных чисел. Далее предполагается, что нам дана *обучающая* выборка из  $n$  пар  $(X, Y)$  из  $\mathcal{X} \times \mathcal{Y}$ .

Говоря неформально, цель статистического обучения заключается в том чтобы на основании имеющейся обучающей выборки построить некоторое правило, которое бы смогло предсказать ответ  $Y$  на основании нового объекта  $X$ . Тем не менее

какое-то предположение о природе данных должно существовать. В данной теории считается, что на  $\mathcal{X} \times \mathcal{Y}$  существует некоторая неизвестная вероятностная мера  $\mathbf{P}$ . И все пары  $(X, Y)$  из обучающей выборки получены независимо согласно этой мере (вероятностному распределению). Второе предположение заключается в том, что любая новая пара  $(X, Y)$  получается согласно тому же самому распределению.

Предположим, что на основании обучающей выборки нам удалось построить функцию  $\hat{f} : \mathcal{X} \rightarrow \mathcal{Y}$ . Заметим, что наличие взаимосвязи между  $X$  и  $Y$  как-то характеризуется самой вероятностной мерой  $\mathbf{P}$ . Для того чтобы делать какие-то предсказания логично предположить, что  $\mathbf{P}$  не является произведением мер по  $X$  и  $Y$ , то есть объекты и, например, их классы вовсе не независимые случайные величины. Одновременно слишком сильное предположение заключается и в существовании строгой функциональной зависимости между  $X$  и  $Y$ . Поэтому  $\mathbf{P}$  такова, что предполагается существование достаточно хорошей (в некотором смысле) связи между объектами и ответами. Для того чтобы формализовать эту идею нужно ввести *функцию ошибки*. Это некоторая функция  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ , которая характеризует потери при отношении объекта  $X$  к ответу  $\hat{f}$  в сравнении с его реальным ответом  $Y$ . Удобно определить функцию  $\ell$  на парах  $(X, Y)$  следующим образом:

$$\ell(\hat{f}, X, Y) := \ell(\hat{f}(X), Y)$$

Типичные примеры:

- В случае задачи классификации бинарные потери  $\ell(\hat{f}, X, Y) = \mathbf{I}\{\hat{f}(X) \neq Y\}$ .
- В задачах регрессии  $\ell(\hat{f}, X, Y) = (\hat{f}(X) - Y)^2$ .
- или  $\ell(\hat{f}, X, Y) = |\hat{f}(X) - Y|^2$ .

Разумной характеристикой решающего правила была бы его ожидаемая ошибка по отношению к обучающей выборке, на основании которого оно построено

$$\mathbf{E} \left[ \ell(\hat{f}, X, Y) \mid (X^n, Y^n) \right]$$

Важно понимать, что математическое ожидание берется по новому объекту  $(X, Y)$ , в то время как само решающее правило  $\hat{f}$  само строится по случайной выборке  $(X^n, Y^n)$ . Для того чтобы избавиться от зависимости от случайной реализации определим уже неслучайную величину, называемую средним риском:

$$\mathbf{E} \left[ \ell(\hat{f}, X, Y) \right] := \mathbf{E} \left[ \mathbf{E} \left[ \ell(\hat{f}, X, Y) \mid (X^n, Y^n) \right] \right]$$

Средний риск зависит теперь только от меры  $\mathbf{P}$  и способа выбора  $\hat{f}$ . Среди всех решающих правил ищется то, которое доставляет минимальный средний риск. Это соответствует так называемому Probably approximately correct learning [13], заключающемуся в выборе решающего правила, которая хорошо приближает ненаблюдаемые данные. Опять же напомним, что в общем случае  $\hat{f}$  – это случайная функция, которая строится на основании обучающей выборки.

Если бы  $\mathbf{P}$  была известна, то задача поиска оптимального  $\hat{f}$  была бы лишь задачей оптимизации.

**Пример 1.1.** Пусть мы имеем дело с задачей классификации  $\mathcal{Y} = \{1, -1\}$  с бинарной функцией потерь. В этом случае ожидаемый риск равен  $P(\hat{f}(X) \neq Y)$ . Среди всевозможных выборов  $\hat{f}$  его минимизирует так называемое байесовское решающее правило  $g(x) = \text{sgn}(\mathbf{E}(Y|X = x))$ . Отметим, что байесовское решающее правило зависит не от обучающей выборки, а от неизвестной меры  $\mathbf{P}$ , поэтому одним из способов приближенного построения байесовского решающего правил являются так называемые plug-in rules, основанные на построении по наблюдаемой выборке эмпирического аналога  $g(x)$ .

**Упр. 1.1.** Для случая классификации докажите оптимальность байесовского решающего правила.

Одной из самых подробно изученных моделей в статистической теории является линейная. Попробуем кратко продемонстрировать разницу между постановками задач статистической теории и теории статистического обучения.

**Пример 1.2 (Линейная регрессия (математическая статистика)).** В качестве функции потерь принято считать квадратичное отклонение  $\ell(f, x, y) = \|f(x) - y\|^2$ . Считая  $x$  —  $d$  мерным вектором ( $n \times d$  матрицей), а  $y$  числом ( $n$  вектором), в случае линейной модели  $f(x)$  можно рассматривать как значение матричного умножения некоторого вектора  $f$  на вектор (матрицу)  $x$ .

Пусть мы пронаблюдали векторы  $x_1, \dots, x_n$  (которые, в простой статистической постановке считаются неслучайными). Наблюдению  $x_t$  соответствует ответ  $Y_t$ , причем для некоторой функции  $g^*$  имеет место равенство

$$Y_t = g^* x_t + \varepsilon_t,$$

где  $\varepsilon_t$  независимые в совокупности случайные величины с нулевым средним и дисперсией  $\sigma^2$ . Стандартное матричное представление записывается

$$Y = Xg^* + \varepsilon,$$

где строки матрицы  $X$  являются векторами  $x_t$ . Определим  $\hat{\Sigma} = \frac{1}{n} \sum_{t=1}^n x_t x_t^T$ . Задача уменьшения среднего риска сводится теперь к поиску  $\hat{g}$ , минимизирующему

$$\mathbf{E} \|\hat{g} - g^*\|_{\hat{\Sigma}}^2 = \mathbf{E} \left( \frac{1}{n} \sum_{t=1}^n (\hat{g}(x_t) - g^*(x_t))^2 \right).$$

Считая, что  $d \leq n$  и матрицы  $\hat{\Sigma}$  обратима, построим по наблюдениям оценку наименьших квадратов

$$\hat{g} = \arg \min_{g \in \mathbf{R}^d} \frac{1}{n} \sum_{t=1}^n (Y_t - g x_t)^2.$$

Хорошо известная явная формула для оценки наименьших квадратов выглядит

$$\hat{g} = \hat{\Sigma}^{-1} \left( \frac{1}{n} \sum_{t=1}^n Y_t x_t \right) = \frac{1}{n} \hat{\Sigma}^{-1} X^T Y.$$

Из  $Y_t = g^* x_t + \varepsilon_t$  легко получить, что

$$g^* = \hat{\Sigma}^{-1} \left( \frac{1}{n} \sum_{t=1}^n (Y_t - \varepsilon_t) x_t \right)$$

Подставляя полученные выражения в  $\mathbf{E}\|\hat{g} - g^*\|_{\hat{\Sigma}}^2$  получаем с учётом того, что матрица  $\frac{1}{n}X\hat{\Sigma}^{-1}X^T$  является проектором на  $Im(X)$

$$\mathbf{E}\|\hat{g} - g^*\|_{\hat{\Sigma}}^2 = \frac{1}{n}\mathbf{E}\left(\varepsilon^T\left(\frac{1}{n}X\hat{\Sigma}^{-1}X^T\right)\varepsilon\right) = \frac{\sigma^2}{n}\text{tr}\left(\frac{1}{n}X\hat{\Sigma}^{-1}X^T\right) \leq \frac{\sigma^2 d}{n}.$$

Такую же скорость сходимости можно получить и в случае, если матрица  $X$  случайна (random design), но при некоторых ограничениях на распределение ее элементов [6].

В теории статистического обучения не принято задавать модель данных в явном виде или предполагать зависимость между  $X$  и  $Y$ . Наше априорное знание о задаче должно быть представлено не ограничением на меру  $\mathbf{P}$ , а априорно заданным семейством отображений  $\mathcal{F}$ , каждое из которых отображает  $X$  в  $Y$ . В литературе, однако, часто и семейство решающих правил  $\mathcal{F}$  называется моделью, а выбор оптимального для задачи  $\mathcal{F}$  — называется задачей выбора модели. В качестве семейства решающих правил могут выступать, например, гиперплоскости в случае линейных классификаторов.

Теперь для некоторого построенного решающего правила  $\hat{f}$  разумной мерой качества будет так называемый ожидаемый избыточный риск (expected excess risk) [9]

$$\mathbf{E}\ell(\hat{f}, X, Y) - \inf_{f \in \mathcal{F}} \mathbf{E}\ell(f, X, Y)$$

Он показывает насколько построенное правило хуже чем лучшее в среднем правило в семействе  $\mathcal{F}$ . Заметим важную особенность: в нашем определении усреднение берется также и по обучающей выборке, то есть, мы работаем с детерминированной величиной и можем давать верхние оценки. Некоторые авторы [7] предпочитают работать с более общей величиной избыточного риска (excess risk), в которой усреднение в  $\mathbf{E}\ell(\hat{f}, X, Y)$  берется только по паре  $X, Y$  и, таким образом, избыточный риск является случайной величиной и все утверждения про него носят вероятностный характер.

**Пример 1.3 (Линейная регрессия(статистическое обучение)).** Пусть  $\mathcal{F}$  представляет собой векторы  $d$ -мерного шара единичного радиуса. Функция потерь также остаётся квадратичной. Теперь пары из обучающей выборки  $(x_t, Y_t)$  получены независимо согласно неизвестной мере  $\mathbf{P}$ . В отличие от статистической постановки мы не предполагаем никаких соотношений между  $X$  и  $Y$ . Более того нам не нужно даже чтобы байесовское решающее правило ( $f(x) = \mathbf{E}(Y|X = x)$ ) принадлежало семейству  $\mathcal{F}$ .

Мы сможем показать, что в случае, если  $\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n (Y_t - f x_t)^2$ , то для некоторой абсолютной константы  $C$

$$\mathbf{E}\ell(\hat{f}, X, Y) - \inf_{f \in \mathcal{F}} \mathbf{E}\ell(f, X, Y) \leq C \left(\frac{d}{n}\right).$$

Таким образом, без ограничений на распределение, пользуясь лишь структурой  $\mathcal{F}$ , можно получить приближение оценки наименьших квадратов к лучшему линейному приближению с порядком  $O(\frac{d}{n})$ .

## 2 Оценки обобщающей способности

### §2.1 Минимизация эмпирического риска

Для того чтобы получить хотя бы один содержательный результат нам нужно получить простое неравенство концентрации меры [3].

**Лемма 2.1 (лемма Хеффдинга).** Пусть  $X$  – случайная величина, такая что почти наверное  $X \in [a, b]$  и  $\mathbf{E}X = 0$ . Тогда для всех  $\lambda > 0$

$$\mathbf{E} \exp(\lambda X) \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

**Доказательство.**

На лекции. ■

**Теорема 2.2 (неравенство Хеффдинга).** Пусть  $Z_1, \dots, Z_n$  – независимые случайные величины, такие что почти наверное  $Z_i \in [a_i, b_i]$ . Обозначим  $S_n = \sum_{i=1}^n Z_i$ , тогда для любого  $t > 0$  имеют место неравенства

$$\mathbf{P}\{S_n - \mathbf{E}S_n \geq t\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

и

$$\mathbf{P}\{S_n - \mathbf{E}S_n \leq -t\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

**Доказательство.**

Является следствием применения леммы 2.1 к случайной величине  $S_n$ , а затем использования варианта Неравенства Маркова (метода Чернова):

$$\mathbf{P}\{S_n - \mathbf{E}S_n \geq t\} \leq \inf_{\lambda > 0} \left(\frac{\mathbf{E} \exp(\lambda(S_n - \mathbf{E}S_n))}{\exp(\lambda t)}\right).$$

■

В данном разделе для удобства обозначим  $L(f) = \mathbf{E}\ell(f, X, Y)$ . Для классификатора  $f$  введем понятие эмпирического риска  $L_n(f) = \frac{1}{n} \sum_{i=1}^n \ell(f, X_i, Y_i)$ . Разумно выбирать такой классификатор, который минимизирует эмпирический риск, то есть  $\hat{f} = \arg \min_{f \in \mathcal{F}} L_n(f)$ . Методы обучения, основанные на этой идее будем называть методами *минимизации эмпирического риска*.

Отметим, что  $L(f) - L_n(f)$  является случайной величиной даже если  $f$  не зависит от обучающей выборки. Оценкой обобщающей способности называется всякая верхняя оценка на вероятность уклонений среднего риска от эмпирического, то есть для  $t > 0$  оценка на  $\mathbf{P}\{L(f) - L_n(f) \geq t\}$ . Далее будем для простоты считать, что функция ошибок  $\ell$  равномерно ограничена единицей. Отсюда можно вывести простое соотношение

**Теорема 2.3.** Пусть  $\mathcal{F} = \{f\}$ , то есть  $\hat{f} = f$ . Тогда для любого  $\delta > 0$  с вероятностью не меньшей  $1 - \delta$  выполнено

$$L(\hat{f}) \leq L_n(\hat{f}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

**Упр. 2.1.** Доказать теорему.

Хотелось бы обобщить приведенный результат на случай, когда  $\mathcal{F}$  содержит более одного решающего правила. Проблема заключается в том, что в оценках сложно учесть специфику, связанную с тем, что мы выбираем именно лучшее на обучающей выборке решающее правило. Поэтому начнем с классического подхода, который заключается в получении равномерных оценок, то есть исследований

$$\sup_{f \in \mathcal{F}} |L(f) - L_n(f)|,$$

или

$$\sup_{f \in \mathcal{F}} (L(f) - L_n(f)).$$

**Теорема 2.4 (Конечный класс функций).** Пусть  $\mathcal{F} = \{f_1, \dots, f_n\}$ . Тогда для любого  $\delta > 0$  одновременно с вероятностью не меньше  $1 - \delta$  выполнено

$$\forall f \in \mathcal{F} : L(f) \leq L_n(f) + \sqrt{\frac{\log(N) + \log \frac{1}{\delta}}{2n}}.$$

**Доказательство.**

С помощью неравенств Хеффдинга и Буля имеем

$$\mathbb{P}\{\exists f \in \mathcal{F} : L(f) - L_n(f) > \varepsilon\} \leq \sum_{i=1}^n \mathbb{P}\{L(f_i) - L_n(f_i) > \varepsilon\} \leq N \exp(-2n\varepsilon^2).$$

Отсюда

$$\mathbb{P}\{\forall f \in \mathcal{F} : L(f) - L_n(f) \leq \varepsilon\} \geq 1 - N \exp(-2n\varepsilon^2).$$

Обращая оценку (что и нужно было научиться делать в упражнении), получаем утверждение теоремы. ■

## §2.2 Теория Вапника-Червоненкиса

В данном разделе будем считать, что имеем дело с теорией классификации:  $\mathcal{Y} = \{1, 0\}$ . В этом случае  $\ell(f, X, Y) = \mathbf{I}\{f(X) \neq Y\}$ . Для фиксированной обучающей выборки  $(X_i, Y_i)_{i=1}^n$  можно определить проекцию  $\mathcal{F}$  на эту выборку, как множество различных булевых векторов:

$$\mathcal{F}_{(X_i, Y_i)_{i=1}^n} = \{(\mathbf{I}\{f(X_1) \neq Y_1\}, \dots, \mathbf{I}\{f(X_n) \neq Y_n\}) : f \in \mathcal{F}\}.$$

*Функцией роста* назовем верхнюю грань по всевозможным выборкам мощности построенной проекции:

$$S_{\mathcal{F}}(n) = \sup_{(X_i, Y_i)_{i=1}^n} |\mathcal{F}_{(X_i, Y_i)_{i=1}^n}|.$$

Очевидно, что если  $|\mathcal{F}| = N$ , то  $S_{\mathcal{F}}(n) \leq N$ . Некоторые свойства функции роста:

- $S_{\mathcal{F}}(n) \leq 2^n$ .
- $S_{\mathcal{F}}(n + m) \leq S_{\mathcal{F}}(n)S_{\mathcal{F}}(m)$ .
- если  $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$ , то  $S_{\mathcal{F}}(n) \leq S_{\mathcal{F}_1}(n) + S_{\mathcal{F}_2}(n)$ .

Размерностью Вапника-Червоненкиса семейства  $\mathcal{F}$  назовем наибольшее натуральное число  $V$ , при котором

$$S_{\mathcal{F}}(V) = 2^V.$$

В случае, если для данного семейства классификаторов такого числа не существует, то считаем, что  $V = \infty$ .

**Пример 2.1.** Одномерное семейство пороговых решающих правил

$$\mathcal{F} = \{f_{\theta}(x) = \mathbf{I}\{x \leq \theta\} : \theta \in [0, 1]\}$$

имеет размерность Вапника-Червоненкиса, равную единице.

**Пример 2.2.** Семейство классификаторов, представляющее собой семейство разделяющих  $d$ -мерных гиперплоскостей имеет размерность Вапника-Червоненкиса, равную  $d + 1$ .

**Пример 2.3.** Семейство классификаторов

$$\{\text{sgn}(\sin(tx)) : t \in \mathbb{R}\}$$

имеет размерность равную  $\infty$ , даже несмотря на то, что параметризуется лишь одним параметром.

Семейство классификаторов, обладающее конечной ёмкостью обладает замечательным свойством:

**Лемма 2.5 (Зауэр, Вапник-Червоненкис).** Для любого семейства классификаторов с размерностью Вапника-Червоненкиса  $V$  для  $n \geq V$ :

$$S_{\mathcal{F}}(n) \leq \sum_{i=0}^V C_n^i$$

**Доказательство.**

Зафиксируем некоторую выборку  $(X_i, Y_i)_{i=1}^n$ , на которой достигается супремум в определении функции роста. Пусть  $\mathcal{F}_0 = \mathcal{F}_{(X_i, Y_i)_{i=1}^n}$  – соответствующая проекция. Будем говорить, что множество булевых векторов  $\mathcal{F}_i$  *разбивает* множество индексов  $S = \{s_1, \dots, s_m\}$ , если ограничение  $\mathcal{F}_i$  на эти индексы реализует полный  $m$ -мерный булев куб.

Пронумеруем векторы в  $\mathcal{F}_0$ . Зафиксируем множество первых компонент этих векторов. Последовательно для каждой 1-чной компоненты заменим 1 на 0 в том случае, если данная процедура не создаст повторных векторов в  $\mathcal{F}_0$ . С нулевыми компонентами не сделаем никаких изменений. После осуществления всех возможных таких замен для первого столбца получаем некоторое множество векторов  $\mathcal{F}_1$ .

Оно совпадает по мощности со множеством  $\mathcal{F}_0$  и обладает следующим замечательным свойством: каждое множество  $S$ , разбиваемое  $\mathcal{F}_1$ , разбивается и  $\mathcal{F}_0$ . Затем по аналогии для второго столбца строим из  $\mathcal{F}_1$  множество  $\mathcal{F}_2$ . И так далее по всем столбцам до множества  $\mathcal{F}_n$ .

Множество  $\mathcal{F}_n$  имеет ту же мощность, что и  $\mathcal{F}_0$  и не разбивает ни одного множества мощностью больше чем  $V$ . Более того, если  $\mathbf{b} \in \mathcal{F}_n$ , то для любого  $\mathbf{b}' \in \{0, 1\}^n$  такого, что  $\mathbf{b}'_i \leq \mathbf{b}_i$  имеет место включение  $\mathbf{b}'_i \in \mathcal{F}_n$ . Таким образом, в  $\mathcal{F}_n$  могут быть только векторы, которые содержат не более  $n$  единичных компонент, так как иначе  $\mathcal{F}_n$  разбило бы некоторое множество, состоящее более чем из  $V$  индексов. Максимальная мощность множества булевых векторов с не более чем  $V$  единицами равна  $\sum_{i=0}^V C_n^i$ , что и доказывает утверждение леммы. ■

С помощью леммы Зауера можно получить верхнюю полиномиальную верхнюю оценку на функцию роста:

$$S_{\mathcal{F}}(n) \leq (n+1)^V$$

Следующей важной идеей является так называемая симметризация. Обозначим  $(X'_i, Y'_i)_{i=1}^n$  независимую копию выборки  $(X_i, Y_i)_{i=1}^n$ . Соответствующая выборка дает эмпирический риск, обозначаемый  $L'_n(f)$ .

**Лемма 2.6.** Для  $\varepsilon > 0$ ,  $n\varepsilon^2 \geq 2$ :

$$\mathbf{P} \left\{ \sup_{f \in \mathcal{F}} (L(f) - L_n(f)) \geq \varepsilon \right\} \leq 2 \mathbf{P} \left\{ \sup_{f \in \mathcal{F}} (L'_n(f) - L_n(f)) \geq \frac{\varepsilon}{2} \right\}.$$

**Доказательство.**

Доказана на лекции. ■

Приведенная лемма позволяет работать лишь с конечными выборками для анализа равномерной по классу классификаторов вероятности отклонения средней ошибки от эмпирической.

**Теорема 2.7 (Вапника–Червоненкиса).** С вероятностью не меньше  $1 - \delta$

$$\forall f \in \mathcal{F} : L(f) \leq L_n(f) + 2 \sqrt{2 \frac{\log(S_{\mathcal{F}}(2n)) + \log(\frac{2}{\delta})}{n}}.$$

Заметим, что кроме констант оценка ничем не уступает оценке для конечного набора функций.



**Доказательство.**

Идея заключается в использовании леммы о симметризации и переходе за счет этого к супремуму по конечному числу функций:

$$\begin{aligned}
& \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} (L(f) - L_n(f)) \geq \varepsilon \right\} \leq \\
& 2 \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} (L'_n(f) - L_n(f)) \geq \frac{\varepsilon}{2} \right\} = \\
& 2 \mathbb{P} \left\{ \sup_{f \in \mathcal{F}_{(x_1, y_1), \dots, (x'_1, y'_1), \dots, (x'_n, y'_n)}} (L'_n(f) - L_n(f)) \geq \frac{\varepsilon}{2} \right\} \leq \\
& 2S_{\mathcal{F}}(2n) \mathbb{P} \left\{ (L'_n(f) - L_n(f)) \geq \frac{\varepsilon}{2} \right\} \leq \\
& 4S_{\mathcal{F}}(2n) \exp \left( \frac{-nt^2}{8} \right).
\end{aligned}$$

■

Теперь в случае конечной ёмкости Вапника–Червоненкиса, используя, например, неравенство  $S_{\mathcal{F}}(n) \leq (n+1)^V$ , можно при фиксированном  $\delta$  получить отклонение среднего риска от эмпирического порядка  $O(\sqrt{\frac{V \log(n)}{n}})$ .

## §2.3 Радемахеровская сложность

### 2.3.1 Необходимые утверждения

Докажем один очень полезный результат: так называемое неравенство ограниченных разностей. Пусть функция  $g : \mathcal{X}^n \rightarrow \mathbb{R}$  удовлетворяет *условию ограниченных разностей*:

$$\sup_{x_1, \dots, x_n, x'} |g(x_1, \dots, x_n) - g(x_1, \dots, x_{i-1}, x', x_{i+1}, \dots, x_n)| \leq c_i, 1 \leq i \leq n.$$

**Лемма 2.8 (лемма Хеффдинга).** Пусть  $X$  – случайная величина, а  $Z$  – случайный вектор, такие что почти наверное  $\mathbf{E}(X|Z) = 0$  и для некоторой неотрицательной функции  $h$  также почти наверное имеет место неравенство:

$$h(Z) \leq V \leq h(Z) + c$$

Тогда для  $\lambda > 0$ :

$$\mathbf{E}[\exp(\lambda V)|Z] \leq \exp \left( \frac{\lambda^2 c^2}{8} \right).$$

**Доказательство.**

Повторяет доказательство леммы 2.1. ■

**Теорема 2.9.** Если  $Z_1, \dots, Z_n$  - независимые случайные величины, а функция  $g$  обладает свойством ограниченных разностей, тогда для  $t \geq 0$ :

$$\begin{aligned} \mathbb{P}\{g(Z_1, \dots, Z_n) - \mathbf{E}g(Z_1, \dots, Z_n) \geq t\} &\leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right), \\ \mathbb{P}\{\mathbf{E}g(Z_1, \dots, Z_n) - g(Z_1, \dots, Z_n) \geq t\} &\leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right). \end{aligned}$$

**Доказательство.**

Введем случайную величину  $V = g - \mathbf{E}g$  и определим

$$V_i = \mathbf{E}\{g|X_1, \dots, X_i\} - \mathbf{E}\{g|X_1, \dots, X_{i-1}\}, \quad i = 1, \dots, n.$$

Легко видеть, что  $\sum V_i = V$ . Введем также случайные величины

$$H_i(X_1, \dots, X_i) = \mathbf{E}\{g(X_1, \dots, X_n)|X_1, \dots, X_i\}.$$

Если  $X_i$  распределена согласно  $F_i$ , то

$$V_i = H(X_1, \dots, X_i) - \int H_i(X_1, \dots, X_{i-1}, x)F_i(dx).$$

Определим случайные величины

$$W_i = \sup_u \left( H(X_1, \dots, X_{i-1}, u) - \int H_i(X_1, \dots, X_{i-1}, x)F_i(dx) \right)$$

и

$$Z_i = \inf_v \left( H(X_1, \dots, X_{i-1}, v) - \int H_i(X_1, \dots, X_{i-1}, x)F_i(dx) \right)$$

Из условия ограниченных разностей

$$W_i - Z_i \leq c_i$$

Таким образом, с помощью леммы 2.8 для всех  $i \in \{1, \dots, n\}$ :

$$\mathbf{E} \exp(\lambda V_i | X_1, \dots, X_{i-1}) \leq \exp\left(\frac{\lambda^2 c_i^2}{8}\right).$$

И почти наверное

$$Z_i \leq V_i \leq W_i.$$

Далее используем метод Чернова и равенство  $\mathbf{E}\{XY\} = \mathbf{E}\{Y\mathbf{E}\{X|Y\}\}$  для  $\lambda > 0$ :

$$\begin{aligned} & \mathbf{P}\{g - \mathbf{E}g \geq t\} \leq \\ & \frac{\mathbf{E} \exp \lambda \sum_{i=1}^n V_i}{\exp(\lambda t)} = \\ & \frac{\mathbf{E} \exp \left( \lambda \sum_{i=1}^{n-1} V_i \mathbf{E}\{\exp(\lambda V_n) | X_1, \dots, X_{n-1}\} \right)}{\exp(\lambda t)} \leq \\ & \exp(-\lambda t) \exp \left( \frac{\lambda^2 \sum_{i=1}^n c_i^2}{8} \right). \end{aligned}$$

Оптимизируя по  $\lambda$ , получаем условие теоремы. ■

Заметим, что неравенство ограниченных разностей является обобщением неравенства Хеффдинга для практически произвольных функций ( с ограниченными разностями ) от независимых случайных величин. Заметим, что  $S_n = \sum_{i=1}^n X_i$  является функцией с ограниченными равенствами  $c_i = \frac{b_i - a_i}{n}$ , коль скоро  $X_i \in [a_i, b_i]$ .

## §2.4 Радмехеровский процесс

### Список литературы

- [1] *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.
- [2] *Boucheron S., Bousquet O., Lugosi G.* Introduction to Statistical Learning Theory // 2004. — Pp. 169-207.
- [3] *Boucheron S., Lugosi G., Massart P.* Concentration Inequalities: A Nonasymptotic Theory of Independence // 2013. —
- [4] *Devroye L., Lugosi G.* Combinatorial Methods in Density Estimation // Springer Series in Statistics. Springer-Verlag, 2001.
- [5] *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.
- [6] *Hsu D., Kakade S. M, Zhang T.* An Analysis of Random Design Linear Regression // . — 2011 <http://arxiv.org/pdf/1106.2363v1.pdf>
- [7] *Koltchinskii V.* Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems // Ecole d'Été de Probabilités de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics. Springer-Verlag, 2011.

- 
- [8] *Ledoux M.* The Concentration of Measure Phenomenon // American Mathematical Society, 2005.
  - [9] *Rakhlin A.* Statistical Learning Theory and Sequential Prediction // Lecture notes, 2014, <http://www-stat.wharton.upenn.edu/~rakhlin/>
  - [10] *Talagrand M.* The Generic Chaining. Upper and Lower Bounds of Stochastic Processes // Springer Monographs in Mathematics, 2005.
  - [11] *Talagrand M.* New concentration inequalities in product spaces // *Inventiones mathematicae* 1996. — Pp. 505-563.
  - [12] *Vapnik V.* Statistical Learning Theory. — John Wiley and Sons, New York, 1998.
  - [13] *L. G. Valiant* A theory of the Learnable. — *Communications of the ACM*, 27, 1984.
  - [14] *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.