Теория статистического обучения

Н. К. Животовский
nikita.zhivotovskiy@phystech.edu

8 апреля 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 семейтсва. Общность подхода, конечно, имела очевидные недостатки, многие из которых были ликвидированы в последнее десятилетие. Катализатором исследлований 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 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 $\hat{f} : \mathcal{X} \to \mathcal{Y}$. Заметим, что наличие взаимосвязи между X и Y как-то характеризуется самой вероятностной мерой P. Для того чтобы делать какие-то предсказания логично предположить, что P не является произведением мер по X и Y, то есть объекты и, например, их классы вовсе не независмые случайные величины. Одновременно слишком сильное предположение заключается и в существовании строгой функциональной зависимости между X и Y. Поэтому 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 $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, которая характеризует потери при относении объекта X к ответу $\hat{f}$ в сравнении с его реальным ответом Y. Удобно определить функцию ℓ 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 регресии $\ell(\hat{f}, X, Y) = (\hat{f}(X) - Y)^2$.
- 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 P и способа выбора $\hat{f}$. Среди всех решающих правил ищется то, которое доставляет минимальный средний риск. Это соответствует так назвываемому Probably approximately correct learning [17], заключающемуся в выбоке 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 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 баесовское решающее правило $g(x) = \operatorname{sgn}(\mathbf{E}(Y|X = x))$. Отметим, что баесовское решающее правило зависит не от обучающей выборки, а от неизвестной меры P, поэтому одним из способов приближенного построения баесовского 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 баесовского 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 векотором), в случае линеной модели $f(x)$ можно рассматривать как значение матричного умножения некоторого вектора f на вектор (матрицу) x.

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

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

где ε_t независимые в совокупности случайные величины с нулевым средним и дисперсией σ^2. Стандартное матричное представление записывается

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

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

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

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

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

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

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

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

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

Такую же скорость сходимости можно получить и в случае, если матрица X случайна (random design), но при некоторых ограничениях на распределение ее элементов [8].

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 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) [12]

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

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 [9] 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 велчиной 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 P. В отличие от статистической постановки мы не предполагаем никаких соотношений между X и Y. Более того нам не нужно даже чтобы баесовское решающее правило $(f(x) = \mathbf{E}(Y|X = x))$ принадлежало семейству $\mathcal{F}$.

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

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

Таким образом, без ограничений на распределние, пользуясь лишь структурой $\mathcal{F}$, можно получить приближение оценки наименьших квадратов к лучшему линейному приблежению 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 [5].

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 $Z_1, \ldots, Z_n$ - независимые случайные величины, такие что почти наверное $Z_i \in [a_i, b_i]$. Обозначим $S_n = \sum\limits_{i=1}^{n} Z_i$, тогда для любого $t > 0$ имеют место неравенства*

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

и

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

Доказательство.

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

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

∎

В данном разделе для удобства обозначим $L(f) = \mathbf{E}\ell(f, X, Y)$. Для классификатора f введем понятие эмпирического риска $L_n(f) = \frac{1}{n}\sum\limits_{i=1}^{n}\ell(f, X_i, Y_i)$. 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 величной 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 $\mathsf{P}\{L(f) - L_n(f) \geq t\}$. Далее будем для простоты считать, что функция ошибок ℓ 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 исследовании

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

или

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

Теорема 2.4 (Конечный класс функций). *Пусть $\mathcal{F} = \{f_1, \ldots, f_n\}$. 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

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

Отсюда

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

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:

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

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 разделюящих d-мерных гиперплоскостей имеет размерность Вапника–Червоненкиса, равную $d+1$.

Пример 2.3. Семейство классификаторов

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

имеет размерность равную ∞, 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 $\mathcal{F}_0 = \mathcal{F}_{(X_i,Y_i)_{i=1}^n}$ – соответсвующая проекция. Будем говорить, что множество булевых векторов $\mathcal{F}_i$ *разбивает* множество индексов $S = \{s_1, \ldots, s_m\}$, 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 $(X_i, Y_i)_{i=1}^n$. Соответствуюшая выборка дает эмпирический риск, обозначаемый $L'_n(f)$.

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

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

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:

$$\mathsf{P}\left\{\sup_{f\in\mathcal{F}}(L(f)-L_n(f))\geq\varepsilon\right\}\leq$$
$$2\,\mathsf{P}\left\{\sup_{f\in\mathcal{F}}(L'_n(f)-L_n(f))\geq\frac{\varepsilon}{2}\right\}=$$
$$2\,\mathsf{P}\left\{\sup_{f\in\mathcal{F}_{(X_1,Y_1),\ldots,(X'_1,Y'_1),\ldots,(X'_n,Y'_n)}}(L'_n(f)-L_n(f))\geq\frac{\varepsilon}{2}\right\}\leq$$
$$2S_{\mathcal{F}}(2n)\,\mathsf{P}\left\{(L'_n(f)-L_n(f))\geq\frac{\varepsilon}{2}\right\}\leq$$
$$4S_{\mathcal{F}}(2n)\exp\left(\frac{-nt^2}{8}\right).$$

■

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

§2.3 Рademахеровская сложность

2.3.1 Необходимые утверждения

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

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

Лемма 2.8 (лемма Хеффдинга). *Пусть X – случайная величина, а Z– случаный 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 $\lambda>0$:

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

Доказательство.

Повторяет доказательство леммы 2.1. ■

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

$$\mathsf{P}\{g(Z_1, \ldots, Z_n) - \mathbf{E}g(Z_1, \ldots, Z_n) \geq t\} \leq \exp\left(-\frac{2t^2}{\sum\limits_{i=1}^{n} c_i^2}\right),$$

$$\mathsf{P}\{\mathbf{E}g(Z_1, \ldots, Z_n) - g(Z_1, \ldots, Z_n) \geq t\} \leq \exp\left(-\frac{2t^2}{\sum\limits_{i=1}^{n} c_i^2}\right).$$

Доказательство.

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

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

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

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

Если X_i распределена согласно F_i, то

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

Определим случайные величины

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

и

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

Из условия ограниченных разностей

$$W_i - Z_i \leq c_i$$

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

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

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 $\lambda > 0$:

$$\mathsf{P}\{g - \mathbf{E}g \geq t\} \leq$$
$$\frac{\mathbf{E}\exp\lambda\sum\limits_{i=1}^{n} V_i}{\exp(\lambda t)} =$$
$$\frac{\mathbf{E}\exp\left(\lambda\sum\limits_{i=1}^{n-1} V_i\ \mathbf{E}\{\exp(\lambda V_n)|X_1,\ldots,X_{n-1}\}\right)}{\exp(\lambda t)} \leq$$
$$\exp(-\lambda t)\exp\left(\frac{\lambda^2\sum\limits_{i=1}^{n} c_i^2}{8}\right).$$

Оптимизируя по λ, получаем условие теоремы. ■

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

§2.4 Рademахеровский процесс

Рассмотрим математическое ожидание интересующего нас функционала: равномерного по классу решающих правил отклонения среднего риска от эмпирического

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

Введем как и ранее, $L'_n(f)$ – эмпирическое среднее по независимой копии обучающей выборки. Соответсвующее ей математическое ожидание будем обозначать $\mathbf{E}'$. Имеем,

$$\mathbf{E}\sup_{f\in\mathcal{F}}|L(f) - L_n(f)| =$$
$$\mathbf{E}\sup_{f\in\mathcal{F}}|\mathbf{E}'L'_n(f) - L_n(f)| \leq$$
$$\mathbf{E}\mathbf{E}'\sup_{f\in\mathcal{F}}|L'_n(f) - L_n(f)| =$$
$$\frac{1}{n}\ \mathbf{E}\mathbf{E}'\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}(\ell(f, X'_i, Y'_i) - \ell(f, X_i, Y_i))\right|.$$

Введем *Рademахеровские случайные величины*, то есть независимые в савокупности (и от X_i, Y_i) случайные величины σ_i, принимающие равновероятно значения 1 и -1. Легко видеть, что для всех i распределения случайных величин $(\ell(f, X'_i, Y'_i) - -\ell(f, X_i, Y_i))$ и $\sigma_i(\ell(f, X'_i, Y'_i) - \ell(f, X_i, Y_i))$ одинаковы. Поэтому, обозначая матема-

тическое ожидание по σ_i как $\mathbf{E}_\sigma$, получаем:

$$\frac{1}{n}\ \mathbf{E}\mathbf{E}' \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}(\ell(f,X_i',Y_i') - \ell(f,X_i,Y_i))\right| =$$
$$\frac{1}{n}\ \mathbf{E}\mathbf{E}'\mathbf{E}_\sigma \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i(\ell(f,X_i',Y_i') - \ell(f,X_i,Y_i))\right| \leq$$
$$\frac{2}{n}\ \mathbf{E}\mathbf{E}_\sigma \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i\ell(f,X_i,Y_i)\right|$$

Введем для фиксированной выборки $(X_i,Y_i)_{i=1}^n$ *условную Радемахеровскую сложность*:

$$\mathcal{R}_n(\mathcal{F}) = \frac{1}{n}\ \mathbf{E}_\sigma \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i\ell(f,X_i,Y_i)\right|$$

и просто *Радемахеровскую сложность*

$$\mathcal{R}(\mathcal{F}) = \mathbf{E}\mathcal{R}_n(\mathcal{F}) = \frac{1}{n}\ \mathbf{E}_\sigma \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i\ell(f,X_i,Y_i)\right|.$$

Таким образом, мы получили, что

$$\mathbf{E}\sup_{f\in\mathcal{F}}|L(f) - L_n(f)| \leq 2\mathcal{R}(\mathcal{F}).$$

Радемахеровскую сложность можно рассматривать как величину, описывающую сложность класса решающих правил. Чем больше Радемахеровская сложность, тем лучше ошибки $\mathcal{F}$ могут кореллировать со случайным шумом σ_i. Как только мы зафиксировали выборку $(X_i,Y_i)_{i=1}^n$ условную Радемахеровскую сложность можно рассматривать как *Радемахеровское среднее*, связанное со множеством $A\subset\mathbf{R}^n$:

$$\mathcal{R}_n(A) = \frac{1}{n}\ \mathbf{E}_\sigma \sup_{a\in A}\left|\sum_{i=1}^{n}\sigma_i a_i\right|,$$

где множество A является множеством векторов ошибок $\mathcal{F}$ на $(X_i,Y_i)_{i=1}^n$.

Рассмотрим простые свойства Радемахеровских средних. Если A, B – ограниченные множества в $\mathbb{R}^n$, $c\in\mathbb{R}$.

- $\mathcal{R}_n(A\cup B) \leq \mathcal{R}_n(A) + \mathcal{R}_n(B)$.
- $\mathcal{R}_n(cA) = |c|\mathcal{R}_n(A)$.
- $\mathcal{R}_n(A\oplus B) \leq \mathcal{R}_n(A) + \mathcal{R}_n(B)$.
- Если $A = \{a^{(1)},\ldots,a^{(N)}\}$, то $\mathcal{R}_n(A) \leq \max\limits_j ||a^{(j)}||_2\ \sqrt{\frac{2\log(2N)}{n}}$.
- (Contraction inequality [9]) Если $\varphi:\mathbb{R}\to\mathbb{R}$ Липшицева с константой L, причем $\varphi(0)=0$, то $\mathcal{R}_n(\varphi(A)) \leq L\mathcal{R}_n(A)$, где φ действует на векторы A покомпонентно.

- $\mathcal{R}_n(A) = \mathcal{R}_n(conv(A))$.

Доказательства первых трёх пунктов являются простыми упражнениями. Разберёмся подробно с 4-ым пунктом.

Случайная величина Y называется сабгауссовской с параметром σ^2 ($Y \in SG(\sigma^2)$), если для $\lambda > 0$:

$$\mathbf{E}\exp(\lambda Y) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

Можно легко показать, что, если $(Y_i)_{i=1}^n$ независимые случайные величины, такие, что $Y_i \in SG(\sigma_i^2)$, то $\sum\limits_{i=1}^n Y_i \in SG\left(\sum\limits_{i=1}^n \sigma_i^2\right)$.

Более того, все так определенные случайные величины имеют нулевое математическое ожидание.

Лемма 2.10. *Пусть $Y_i \in SG(\sigma_i^2)$, $i = 1, \ldots, N$. Тогда*

$$\mathbf{E}\max_i |Y_i| \leq \max \sigma_i \sqrt{2\ln(2n)}.$$

Доказательство.

$$\exp\left(\lambda \mathbf{E}\left\{\max_{i=1,\ldots,N} Y_i\right\}\right) \leq$$
$$\mathbf{E}\left\{\exp(\lambda \max_{i=1,\ldots,N} Y_i)\right\} =$$
$$\mathbf{E}\left\{\max_{i=1,\ldots,N} \exp(\lambda Y_i)\right\} \leq$$
$$N\exp\left(\frac{\lambda^2 \max\limits_{i=1,\ldots,N} \sigma_i^2}{2}\right).$$

Логарифимруя обе части и оптимизируя по λ, получаем, что

$$\mathbf{E}\max_i Y_i \leq \max \sigma_i \sqrt{2\ln(n)}.$$

Для получения утверждения леммы нужно применить полученное оценку к набору $Y_1, \ldots, Y_n, -Y_1, \ldots, Y_n$. ∎

Важно заметить, что в доказательстве нигде не используется условие на независимость случайных величин. С помощью предложенной леммы можно легко доказать 4-ое свойство Радемахеровского среднего. Для этого нужно показать, что супремум берется по конечному числу модулей субгауссовских случаных величин.

Пусть мы имеем дело с задачей классификации с бинарной функцией потерь. Тогда 4-ое свойство можно переписать в виде

$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2\log(2S_{\mathcal{F}}(n))}{n}}.$$

где неравенство выполнены почти наверное. Что в случае конечной размерности Вапника–Червоненкиса даёт порядок $O\left(\sqrt{\frac{V\log(n)}{n}}\right)$.

Особенность Радемахеровского процесса заключается, что его можно анализировать с помощью горадо более мощных средств теории эмпирических процессов. Действительно, можно рассматривать $\left|\sum_{i=1}^{n}\sigma_i a_i\right|$ как эмпирический процесс со множеством состояний A. В этом случае Радемахеровское среднее есть ни что иное, как ожидаемый супремум этого процесса. Теория эмпирических процессов показывает, что во многих случаях поведение процесса зависит от 'геометрии' пространства состояний. В нашем случае – это метрические свойства множества A.

Рассмотрим задачу классификации. В этом случае условно по обучающей выборке множество $A = A\left((X_i, Y_i)_{i=1}^n\right)$ можно представить себе как набор не более чем $S_{\mathcal{F}}(n)$ различных булевых векторов. Введем на на паре векторов метрику ρ:

$$\rho(a,b) = \sqrt{\frac{1}{n}d_H(a,b)},$$

где d_H – метрика Хэмминга.

Будем говорить, что множества $B \subset \{0,1\}^n$ является ε-покрытием множества A, если объединение замкнутых ε-шаров (по введенной метрике) с центрами в точках B содержат A.

Обозначим $N(\varepsilon, A)$ – *число покрытия*, равное мощности минимального ε-покрытия множества A.

Теорема 2.11. *Для задачи классификации почти наверное*

$$\mathcal{R}_n(\mathcal{F}) \le \frac{12}{\sqrt{n}} \sup_{(X_i,Y_i)_{i=1}^n} \int_0^1 \sqrt{\log(2N(\varepsilon, A))}d\varepsilon,$$

где $A = A\left((X_i, Y_i)_{i=1}^n\right)$.

Доказательство.

Зафиксируем конечное множество различных n мерных булевых векторов A. Зафиксируем $B^{(0)} = \{(0,\ldots,0)\}$ – множество состоящее из нулевого вектора, а $B_1,\ldots,B_M$ подмножества $\{0,1\}^n$, являющиеся минимальными 2^{-k}-покрытиями множества A, а $M = \lfloor\log_2(\sqrt{n})\rfloor + 1$.

Пусть для конкретной реализации σ_i вектор $b^* \in A$ доставляет максимум выражения $\left|\sum_{i=1}^{n}\sigma_i b_i\right|$, среди всех векторов A. Обозначим $b^{(k)}$–ближайший к нему вектор в B_k. Из неравенства треугольника так как $\rho(b^{(k)}, b^*) \le 2^{-k}$ мы имеем

$$\rho(b^{(k)}, b^{(k-1)}) \le 2^{-k} + 2^{-k+1} = 3\times 2^{-k}.$$

Тогда

$$\sum_{i=1}^{n}\sigma_i b_i^* = \sum_{k=1}^{M}\sum_{i=1}^{n}\sigma_i(b_i^{(k)} - b_i^{(k-1)}).$$

Тогда

$$\mathbf{E}\left\{\max_{b\in A}\left|\sum_{i=1}^{n}\sigma_i b_i\right|\right\} =$$
$$\mathbf{E}\left\{\left|\sum_{k=1}^{M}\sum_{i=1}^{n}\sigma_i(b_i^{(k)}-b_i^{(k-1)})\right|\right\} \leq$$
$$\sum_{k=1}^{M}\mathbf{E}\left\{\left|\sum_{i=1}^{n}\sigma_i(b_i^{(k)}-b_i^{(k-1)})\right|\right\} \leq$$
$$\sum_{k=1}^{M}\mathbf{E}\left\{\max_{b\in B_k, c\in B_{k-1}, \rho(b,c)\leq\frac{3}{2^k}}\left|\sum_{i=1}^{n}\sigma_i(b_i^{(k)}-b_i^{(k-1)})\right|\right\}.$$

Математическое ожидание под суммой можно представить как математическое ожидание максимума модулей $|B_k||B_{k-1}| \leq N(2^{-k},A)^2$ экземпляров субгауссовских случайных величин с параметром $\sigma^2 = n(3/2^k)^2$. Условия на параметр σ^2 получаются из независимости σ_i и леммы 2.1. Применяя теперь лемму 2.10 получаем

$$\mathbf{E}\left\{\max_{b\in B_k, c\in B_{k-1}, \rho(b,c)\leq\frac{3}{2^k}}\left|\sum_{i=1}^{n}\sigma_i(b_i^{(k)}-b_i^{(k-1)})\right|\right\} \leq 3\sqrt{n}2^{-k}\sqrt{2\log(2N(2^{-k},A)^2)}.$$

А значит

$$\mathbf{E}\left\{\max_{b\in A}\left|\sum_{i=1}^{n}\sigma_i b_i\right|\right\} =$$
$$3\sqrt{n}\sum_{k=1}^{M}2^{-k}\sqrt{2\log(2N(2^{-k},A)^2)} \leq$$
$$12\sqrt{n}\sum_{k=1}^{\infty}2^{-k-1}\sqrt{\log(2N(2^{-k},A))} \leq$$
$$12\sqrt{n}\int_0^1\sqrt{\log(2N(\varepsilon,A))}d\varepsilon.$$

■

Полученная теорема говорит, что Радемахеровское среднее контролируется не логарифмом мощности множества A, а некоторой величиной, которая существенно учитывает структуру A. Будем называть величину $\int_0^1\sqrt{\log(2N(\varepsilon,A))}d\varepsilon$ *метрической энтропией* множества A.

Важность полученного результата связна с использованием следующей теоремы

Теорема 2.12 (Haussler [7]). *Если множество булевых векторов A состоит из различных векторов ошибок семейства классификаторов с размерностью Вапника-Червоненкиса равной V, то для $0 \leq \varepsilon \leq 1$:*

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

Применяя данную теорему можно получить, что для некоторой абсолютной константы C для задачи классификации почти наверное

$$\mathcal{R}_n(\mathcal{F}) \leq C\sqrt{\frac{V}{n}}.$$

Пример 2.4 (Теорема Дворецкого-Кифера-Вольфовитца). С помощью данного результата можно получить усиление теоремы Гливенко-Кантелли о равномерной сходимости эмпирической функции распределения к настоящей функции распределения. Пусть $F(x)$ — функция распределения, а $F_n(x)$ — эмпирическая функция распределения. Можно считать, что $x \in \mathbb{R}$ индексирует некоторые классификаторы, которые ошибаются на всех объектах (X, Y) тогда и только тогда, когда $X \leq x$, то есть $\ell(f(x), X, Y) = \mathbf{I}\{X \leq x\}$. Такие классификаторы обладают единичной размерностью Вапника–Червоненкиса. Таким образом, для некоторой $C > 0$

$$\mathbf{E}\left\{\sup_x |F_n(x) - F(x)|\right\} \leq \frac{C}{\sqrt{n}}$$

Более общий вариант теоремы даёт неулучшаемую явную константу $C = 1$ [11], а также задает хвосты распределения $\sup\limits_x |F_n(x) - F(x)|$.

Рассмотрим $\sup\limits_{f\in\mathcal{F}} |L(f) - L_n(f)|$ как функцию от наблюдаемой выборки $(X_i, Y_i)_{i=1}^n$. Простым упражнением является доказательство того факта, что, если функция потерь равномерно ограничена единицей, то введенная функция является функцией с ограниченными приращениями с $c_i = \frac{2}{n}$. Аналогично, рассматривая условную Радемахеровску сложность как функцию от наблюдаемой выборки, доказываем для нее, что она также является функцией с ограниченными разностями с $c_i = \frac{2}{n}$.

Теорема 2.13. *С вероятностью не меньшей* $1 - \delta$

$$\sup_{f\in\mathcal{F}} |L(f) - L_n(f)| \leq 2\mathcal{R}(\mathcal{F}) + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}}.$$

Также

$$\sup_{f\in\mathcal{F}} |L(f) - L_n(f)| \leq 2\mathcal{R}_n(\mathcal{F}) + 3\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}$$

Доказательство.

Доказательство первого неравенства заключается в применении неравенства ограниченных разностей 2.9 для $\sup\limits_{f\in\mathcal{F}} |L(f) - L_n(f)|$ и замене $\mathbf{E}\{\sup\limits_{f\in\mathcal{F}} |L(f) - L_n(f)|\}$ на $2\mathcal{R}(\mathcal{F})$.

Доказательство второго неравенства заключается в использовании неравенства ограниченных разностей для $\mathcal{R}_n(\mathcal{F})$ и учёте того, что $\mathbf{E}\mathcal{R}_n(\mathcal{F}) = \mathcal{R}(\mathcal{F})$. ∎

Заметим, что правая часть второго неравенства является полностью вычислимой по наблюдаемой выборке. Это наша первая так называемая data-dependent оценка. Подобные результаты во многом оправдывают, использование условной Радемахеровской сложности как меры сложности класса функций на практике, в частности, в задачах выбора модели.

§2.5 Локализованная Радемахеровская сложность

Вспомним вариант самой простой оценки в случае одной функции:

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

В случае, например, задачи классификации данное неравенство даёт скорость сходимости среднего к математическому ожиданию в схеме испытаний Бернулли. Заметим, что данный порядок точен тогда и только тогда, когда $p = \frac{1}{2}$. При этом достигается максимальная дисперсия. Таким образом, для уточнения оценок нужно существенно учитывать дисперсию значений функции ошибок.

Вторым важным моментом является то, что sup берется по всему большому множеству решающих правил, в то время как реальные минимизаторы эмпирического риска выбирают из узкого класса достаточно хороших решающих правил. Это и приводит к идее локализации: множество решающих правил специальным образом расслаивается на компоненты с разными дисперсиями, затем для каждого слоя можно считать уже собственные Радемахеровские сложности.

В предыдущем разделе в случае задачи классификации нами были получены порядки равномерной сходимости частот к вероятностям порядка $O(\frac{1}{\sqrt{n}})$. В реальности можно добиться чтобы риск минимизатора эмпирического риска стремился к минимальному в классе среднему риску с порядками вплоть до $O(\frac{1}{n})$ [15].

В данном разделе мы продемонстрируем один из сопособов получения оценок с порядком, лучшим чем $O(\frac{1}{\sqrt{n}})$.

Функцию $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ будем называть *подкоренной* (subroot function) , если она обладет следующими свойствами:

- Не убывает.
- А функция $r \to \frac{\psi(r)}{\sqrt{r}}$ не возрастает.
- Не является тождественным нулем.

Из определния следует непрерывность на области определения, а также существование и единственность решения уравнения $\psi(r) = r$.

Сформулируем некоторое очень общее утверждение:

Теорема 2.14. *Пусть функция ошибок ℓ равномерно ограничена единицей и существует некоторый функционал $T : \mathcal{F} \to \mathbb{R}^+$ и константа $B > 0$, такая что*

$$\mathsf{D}\ell(f, X, Y) \leq T(f) \leq B\mathbf{E}\ell(f, X, Y),$$

где символ D обозначает дисперсию. Пусть для некоторой подкоренной функция ψ для всех $r \geq r^*$, где $r^* = \psi(r^*)$, выполнено

$$\psi(r) \geq B\mathbf{E}\mathcal{R}_n(f \in \mathcal{F} : T(f) \leq r).$$

Тогда одновременно для всех $f \in \mathcal{F}$ с вероятностью не меньшей $1 - \delta$ и для всех $K > 1$

$$L(f) \leq \frac{K}{K-1} L_n(f) + \frac{704K}{B} r^* + \frac{\log(\frac{1}{\delta})(11 + 26BK)}{n}.$$

Абстрагируясь пока от деталей и условий, сразу заметим, что для достаточно больших K мы имеем $L(f) - L_n(f) \approx L(f) - \frac{K}{K-1} L_n(f) = O(\max(\frac{1}{n}, r^*))$. Если удастся показать удачную асимптотику для станционарной точки r^*, то есть решения $\psi(r) = = r$, то можно будет существенно улучшить асимптотику в наших оценках.

До обсуждения доказательства и получения следствий рассмотрим некоторые общие вопросы:

Выбор функции ψ. Действительно, в утверждении фигурирует некоторая подкоренная функция ψ, которая ограничивает сверху Радемахеровскую сложность класса функций с ограниченным дисперсиями. Первый вариант заключается в построении верхних оценок через метрическую энтропию, о которой уже шла речь в теореме 2.11. При этом могут возникнуть проблемы, связанные с доказательством того, что полученная функция будет подкоренной. Поэтому будет использоваться другая конструкция, описанная, например в [1]. Введем на решающих правилах из $\mathcal{F}$ операцию умножения на действительное число. По определению αf есть некоторое решающее правило со свойством для всех (X, Y):

$$\ell(\alpha f, (X, Y)) = \alpha \ell(f, (X, Y)).$$

Аналогично, с помощью функции ошибок, можно определить и сумму решающих правил. Важно, что операция умножения и суммы могут не сохранять свойств неотрицательности или равномерной ограниченности с той же константой функции ошибок.

Семейство $\mathcal{F}^*$ будем называть *звездным замыканием* $\mathcal{F}$ с центров в f_0, если

$$\mathcal{F}^* = \{f_0 + \alpha(f - f_0) : f \in \mathcal{F}, \alpha \in [0, 1]\}.$$

Рецепт построения нужной нам подкоренной функции строиться теперь из двух лемм:

Лемма 2.15. *Пусть $\mathcal{F}^*$ является звездным замыканием с центром $\hat{f}$ (здесь можно считать, что $\hat{f}$ является случайным и зависит от выборки). Пусть функционал $T : \mathcal{F} \to \mathbb{R}_+$ удовлетворяет $T(\alpha f) \le \alpha^2 T(f)$, для всех $f \in \mathcal{F}^*$ и $\alpha \in [0, 1]$. Тогда функции ψ, ψ_1, определяемые*

1. $\psi(r) = \mathcal{R}_n(f \in \mathcal{F}^* : T(f - \hat{f}) \le r)$ *(при этом ψ – случайная функция);*
2. $\psi_1(r) = \mathbf{E}\psi(r)$,

являются подкоренными.

Лемма 2.16. *Пусть $\mathcal{F}^*$ является звездным замыканием с центром $\hat{f}$ семейства $\mathcal{F}$, тогда в условиях предыдущей леммы для всех $r \ge 0$*

$$\mathcal{R}_n(f \in \mathcal{F}^* : T(f - \hat{f}) \le r) \ge \mathcal{R}_n(f \in \mathcal{F} : T(f - \hat{f}) \le r).$$

Отсюда получается конструкция для подкоренных функций, обладающих условиями из теоремы 2.14. Нужно для данного семейства строить звездное замыкание с произвольным центром и использовать соответствующую Радемахеровскую сложность (как функцию от r).

Одним из ключевых моментов доказательства теоремы 2.14 является использование неравенства Талаграна [14] в форме Буске [4]:

Теорема 2.17. *Пусть $c > 0$, X_i — независимые случайные величины, распределенные согласно неизвестному распределению P, заданному на $\mathcal{X}$. $\mathcal{F}$ — счетный класс функций, отображающих $\mathcal{X} \to \mathbb{R}$ (конфликт обозначений, не путать с классом решающих правил). Пусть для всех $f \in \mathcal{F}$ выполнено $\mathbf{E}f(X_i) = 0$ и $||f||_\infty \leq c$. Пусть $\sigma^2 \geq \sup_{f \in \mathcal{F}} \mathsf{D}f(X_i)$, тогда для всех $x > 0$*

$$\mathsf{P}\{Z \geq \mathbf{E}Z + x\} \leq \exp\left(-vh\left(\frac{x}{cv}\right)\right),$$

где $Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i)$, $h(x) = (1+x)\log(1+x) - x$, $v = n\sigma^2 + 2c\mathbf{E}Z$.

Используя элементарное неравенство $h(x) \geq \frac{x^2}{2+\frac{2x}{3}}$, получаем с вероятностью не меньшей $1 - \exp(-x)$:

$$Z \leq \mathbf{E}Z + \sqrt{2xv} + \frac{cx}{3}.$$

Прямым следствием данного утверждения будет следующая теорема:

Теорема 2.18. *Пусть $\mathcal{F}$ — класс решающих правил, а функции ошибок ℓ принимают значения в $[a, b]$. Пусть для $r > 0$ для всех $f \in \mathcal{F}$ выполнено $\mathsf{D}\ell(f, X, Y) \leq r$. Тогда с вероятность не меньшей $1 - \delta$:*

$$\sup_{f \in \mathcal{F}}(L(f) - L_n(f)) \leq \inf_{\alpha > 0}\left(2(1+\alpha)\mathcal{R}(\mathcal{F}) + \sqrt{\frac{2r\log(\frac{1}{\delta})}{n}} + (b-a)\left(\frac{1}{3} + \frac{1}{\alpha}\right)\frac{\log(\frac{1}{\delta})}{n}\right),$$

и с вероятностью не меньшей $1 - 2\delta$

$$\sup_{f \in \mathcal{F}}(L(f) - L_n(f)) \leq$$
$$\inf_{\alpha \in (0,1)}\left(2\frac{1+\alpha}{1-\alpha}\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{2r\log(\frac{1}{\delta})}{n}} + (b-a)\left(\frac{1}{3} + \frac{1}{\alpha} + \frac{1+\alpha}{2\alpha(1-\alpha)}\right)\frac{\log(\frac{1}{\delta})}{n}\right).$$

Более того, такие же неравенства выполнены и для величины $\sup_{f \in \mathcal{F}}(L_n(f) - L(f))$.

Заметим, что второе неравенство даёт оценку, полностью вычислимую по наблюдаемой выборке и является аналогом неравенства из теоремы 2.13. Однако на прямую выигрыша от нового неравенства получить не удастся. Хотя член порядка $\frac{1}{\sqrt{n}}$ содержит множитель $\sqrt{r}$ в виде дисперсии, в общем случае нельзя гарантировать его малость, так как в большом классе функций максимальная дисперсия ошибки может быть большой. Более хитрый прием заключается в расслоении всех классификаторов по их дисперсиям и применении теоремы 2.18.

Опишем подробнее шаги доказательства теоремы 2.14:

- На первом шаге вводится некоторый функционал $\omega(f)$, значения которого порядка $\mathsf{D}\ell(f, X, Y)$. С помощью введенной нами 'алгебры' на решающих правилах для $r > 0$ переходим к классу с малыми дисперсиями $\{\frac{rf}{\omega(f)} : f \in \mathcal{F}\}$.

- На втором шаге для введенного класса применяется первое из неравенств теоремы 2.18.

- С помощью введенного функционала T, по порядку равного дисперсии функции ошибок, построенный класс специальным образом разбивается на области, в зависимости от значений $T(f)$.

- Используя подкоренную функцию из условия теоремы 2.14, можно показать, что введенная на втором шаге локальная Радемахеровская сложность оганичена величиной, пропорциональной этой подкоренной функции. Выбирая параметр r как стационарную точку уравнения $\psi(r) = r$ и возвращаясь к изначальному классу, получаем утверждение теоремы. Важным моментом здесь является то, что выбор r^* позволяет избавиться от члена порядка $\frac{1}{\sqrt{n}}$.

Заметим, что условие $\mathsf{D}\ell(f, X, Y) \leq B\mathbf{E}\ell(f, X, Y)$ является абсолютно неизбежным в теореме 2.14. Однако, в частных случаях его легко проверить: например, если $\ell \in [0, 1]$. Тогда удобно выбрать $T(f) = \mathbf{E}\ell(f, X, Y)$.

Применим полученные результаты к задаче классификации. Пусть ℓ – есть индиактор ошибки. Тогда верно следующее следствие теоремы 2.14.

Теорема 2.19. *Если $\mathcal{F}$ имеет конечную размерность Вапника–Червоненкиса, равную V, то для всех $K > 1$, $\delta > 0$ одновременно для всех $f \in \mathcal{F}$ с вероятностью не меньшей $1 - \delta$:*

$$L(f) \leq \frac{K}{K-1} L_n(f) + CK\left(\frac{V\log(\frac{n}{V}) + \log(\frac{1}{\delta})}{n}\right)$$

где $C > 0$ — абсолютная константа.

Игнорируя множитель $\frac{K}{K-1}$, получаем при фиксированной размерности для $L(f) - L_n(f)$ порядок $O(\frac{\log(n)}{n})$ вместо ранее получавшегося $O(\frac{1}{\sqrt{n}})$.

Доказательство следствия основано на прямом построении подкоренной функции. Действительно, леммы 2.15, 2.16 указывают на способ её построения. Оказывается, что для Радемахеровской сложности звездного замыкания класса (в случае классификации) можно построить оценку подкоренной функции порядка $O\left(\sqrt{\frac{rV\log(\frac{n}{V})}{n}}\right)$, что очень напоминает результат теоремы 2.11, полученный с использованием метрической энтропии. Отсюда, для стационарной точки r^*, то есть решения $\psi(r) = r$ выполнено

$$r^* \leq \frac{CV\log(\frac{n}{V})}{n}.$$

Подстановка данной оценки в теорему 2.14 даёт необходимое утверждение.

§2.6 Оценки избыточного риска

Вернемся к оценкам избыточного риска

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

Здесь мы считаем, что в первом слагаемом усреднение берется только по паре (X, Y). Таким обазом, избыточный риск является случайной величиной, так как случайным является выбор $\hat{f}$, основанный на обучающей выборке. Нас интересует *минимизатор*

эмпирического риска, то есть $\hat{f} \in \arg\min_{f\in\mathcal{F}} L_n(f)$. Следующая простая лемма дает возможность ограничить избыточный риск с помощью хорошо изученной оценки на равномерные отклонения.

Лемма 2.20. *Для минимизатора эмпирического риска с вероятностью единица*

$$L(\hat{f}) - \inf_{f\in F} L(f) \le 2\sup_{f\in\mathcal{F}} |L(f) - L_n(f)|.$$

Доказательство.

Так как $\hat{f}$ минимизатор эмпирического риска имеем $L_n(\hat{f}) - L_n(f^*) \le 0$. Пусть также для некоторого f^* выполнено $\inf_{f\in F} L(f) = L(f^*)$. Тогда

$$\begin{aligned}
&L(\hat{f}) = L(\hat{f}) + L(f^*) - L(f^*) \le \\
&L(\hat{f}) + L(f^*) - L(f^*) - L_n(\hat{f}) + L_n(f^*) \le \\
&L(f^*) + 2\sup_{f\in\mathcal{F}} |L(f) - L_n(f)|.
\end{aligned}$$

■

Стоит отметить, что использование леммы §2.6 не всегда дает хорошую оценку для избыточного риска. Поэтому часто избыточный риск оценивают напрямую. Важно понимать, что любые оценки избыточного риска также должны включать в той или иной форме информацию о классе решающих правил или информацию о распределении. Этот принцип заложен в следующей теореме (котор:

Теорема 2.21 (No free lunch theorems [12]). *Какой бы мы не выбрали способ обучения (то есть способ выбора $\hat{f}$ из $\mathcal{F}$ по обучающей выборке) для любого n и $\varepsilon > 0$ найдется распределение P на парах X, Y, такое что $\inf_f L(f) = 0$, а*

$$\mathbf{E}_{(X_i,Y_i)_{i=1}^n} L(\hat{f}) \ge \frac{1}{2} - \varepsilon.$$

Какой бы мы не выбрали способ обучения, для любой сходящейся к нулю последовательности a_n найдется распределение P на парах X, Y такое что $\inf_f L(f) = 0$, а для всех n

$$\mathbf{E}_{(X_i,Y_i)_{i=1}^n} L(\hat{f}) \ge a_n,$$

где $\hat{f} = \hat{f}((X_i, Y_i)_{i=1}^n)$.

Естественно считать, что данная теорема выполнена в самом общем случае, в частности, без условий ограниченности функций потерь. Выигрыш как раз и заключается в выборе ограничивающего семейства $\mathcal{F}$.

§2.7 Локализованная Радемахеровская сложность (продолжение)

Вернемся теперь к тому, как идея локализованной Радемахеровской сложности может быть использована для оценок избыточного риска. Сформулируем сначала основные преимущества оценок, по сравнению, например, с теоремой 2.14.

- Избавление от множителей типа $\frac{K}{K-1}$.
- Переход от невычислимой по данным функции ψ к зависящей от данных $\hat{\psi}$.
- Оценка работает непосредственно со свойствами решающих правил f,а не с функциями ошибок $\ell(f, X, Y)$

Перед формулировкой основого результата нужно задать некоторые ограничения на функцию потерь:

1. Существует $f^* \in \mathcal{F}$ такой что $L(f^*) = \inf\limits_{f \in \mathcal{F}} L(f)$
2. Функция потерь Липшицева по первому аргументу с некоторой константой L, то есть
$$|\ell(y_1, y) - \ell(y_2, y)| \leq L|y_1 - y_2|.$$
3. Существует константа $B > 0$, такая что для всех $f \in \mathcal{F}$
$$\mathbf{E}(f - f^*)^2 \leq B^2(L(f) - L(f^*)).$$

Таким условиям удовлетворяют, например, квадратичная функция потерь, если класс $\mathcal{F}$ выпуклый и равномерно ограниченный. Отметим, что в данном разделе мы не связываем на прямую решающие правила с функциями ошибок, поэтому все алгебраические операции над функциями из $\mathcal{F}$ являются стандартными так же как и определение выпуклости.

Теперь можно сформулировать основную теорему.

Теорема 2.22. *Пусть $\mathcal{F}$ является выпуклым классом функций, принимающих значения из отрезка $[-1, 1]$. Пусть функци потерь удовлетворяет приведенным выше условиям 1–3. Пусть $\hat{f} \in \arg\inf\limits_{f \in \mathcal{F}} L_n(f)$. Определим*

$$\hat{\psi}(r) = c_1 \mathcal{R}_n \left(f \in \mathcal{F} : \sum_{i=1}^{n} (f(X_i) - \hat{f}(X_i))^2 \leq c_3 r \right) + \frac{c_2 \log(\frac{1}{\delta})}{n},$$

где $c_1 = 2L \max(B, 10L)$, $c_2 = 11L^2 + c_1$, $c_3 = 2484 + 4B(11L + 27B)/c^2$. Тогда с вероятностью не меньшей чем $1 - 4\delta$

$$L(\hat{f}) - L(f^*) \leq \frac{705}{B} r^* + \frac{(11L + 27B) \log(\frac{1}{\delta})}{n},$$

где r^* — решение уравнения $\hat{\psi}(r) = r$.

Список литературы

[1] *Bartlett P., Bousquet O., Mendelson S.* Localized Rademacher complexities // The annals of statistics. — 2005.

[2] *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.

[3] *Boucheron S., Bousquet O., Lugosi G.* Introduction to Statistical Learning Theory // 2004. — Pp. 169-207.

[4] *Bousquet O.* A Bennet concentration inequality and its applications to suprema of empirical process // 2002.

[5] *Boucheron S., Lugosi G., Massart P.* Concentration Inequalities: A Nonasymptotic Theory of Independence // 2013. —

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

[7] *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.

[8] *Hsu D., Kakade S. M, Zhang T.* An Analysis of Random Design Linear Regression // . — 2011 http://arxiv.org/pdf/1106.2363v1.pdf

[9] *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.

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

[11] *Massart P.* The tight constant in Dvoretzky-Kiefer-Wolfowitz inequality // Annals of Probability, 1990.

[12] *Rakhlin A.* Statistical Learning Theory and Sequential Prediction // Lecture notes, 2014, http://www-stat.wharton.upenn.edu/ rakhlin/

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

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

[15] *Tsybakov A.M.* Optimal aggregation of classifiers in statistical learning // Annals of Statistics 2004. —

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

[17] *L. G. Valiant* A theory of the Learnable. — Communications of the ACM, 27, 1984.

[18] *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.