Оценки риска в теории статистического обучения

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

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 для любого нового объекта X. Тем не менее какое-то предположение о природе данных должно существовать.

В данной теории предполагается:

- На $\mathcal{X} \times \mathcal{Y}$ существует некоторая неизвестная вероятностная мера P.
- Все пары (X, Y) из обучающей выборки получены независимо согласно этой мере (вероятностному распределению).
- 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 независимые 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 ошибок*. Функция ошибок — это некоторая функция $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, которая характеризует потери при отнесении 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 $\ell(\hat{f}, X, Y) = |\hat{f}(X) - Y|$.

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. 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 $\mathcal{Y} = \{1, 0\}$ 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 $g(x) = \mathbf{I}\{\eta(x) \geq \frac{1}{2}\}$, где $\eta(x) = \mathbf{E}\,(Y|X = x)$. Отметим, что байесовское решающее правило зависит не от обучающей выборки, а от неизвестной меры P, поэтому одним из способов приближенного построения байесовского решающего правил являются так называемые plug-in rules, основанные на построении по наблюдаемой выборке эмпирического аналога $g(x)$. В самом общем случае, если мера нам неизвестна мы можем только пытаться приблизиться к байесовскому решающему правилу.

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

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 (аффинные подпространства) в случае линейных классификаторов.

Теперь для некоторого построенного решающего правила $\hat{f}$ разумной мерой качества будет так называемый ожидаемый избыточный риск (expected excess risk, regret) [4]

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

Эта величина 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 [2] 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.

2 Анализ минимаксных значений

Пусть $\mathcal{P}$ – некоторое фиксированное семейство распределений на $\mathcal{X} \times \mathcal{Y}$. В качестве $\mathcal{P}^*$ обозначим для удобства семейство состоящее из всех возможных распределений на указанном произведении. Конечно, здесь имеются проблемы с одновременной измеримостью функций относительно семейства всех распределений, но мы специально не будем вдаваться в эти подробности, которые в действительности технически разрешимы. Также считаем, что зафиксировано семейство решающих правил $\mathcal{F}$ и длина обучающей выборки n.

Минимаксным значением назовем следующее число

$$V(\mathcal{F}, \mathcal{P}, n) = \inf_{\tilde{y}} \sup_{\mathsf{P} \in \mathcal{P}} \left\{ \mathbf{E}\ell(\tilde{y}, X, Y) - \inf_{f \in \mathcal{F}} \mathbf{E}\ell(f, X, Y) \right\}.$$

Попробуем подробнее разобраться с введенным определением. Во-первых, стоит отметить, что минимаксное значение не является случайной величиной. $\tilde{y}$ есть решающее правило, обученное по n-элементной выборке. Минимаксное значение можно трактовать следующим образом:

1. Противник выбирает самую плохую вероятностную меру, то есть ту, на которой наш метод обучения дает большую разницу между собственным риском и минимальным в классе (является ли минимальный в классе риск байесовским зависит, в частности, от того, находится ли для данного распределения байесовское решающее правило в классе $\mathcal{F}$).

2. В ответ мы выбираем такой способ построения $\tilde{y}$ по обучающей выборке (метод обучения), который минимизирует эту разницу.

Малое значение минимаксного значения дает основания полагать, что для фиксированного класса $\mathcal{F}$ и выбранной длины обучающей выборки n возможно ввести некоторый метод обучения, с помощью которого можно получать решающие правила с риском, близким к байесовскому. Более того, если для некоторого метода обучения удается показать, что ожидаемый избыточный риск равен минимаксному значения, то может быть доказана в некотором смысле оптимальность данного метода обучения. Мы приведем примеры, когда конкретный метод обучения имеет те же порядки ожидаемого избыточного риска.

Отметим, что в определении минимаксного значения мы считаем, что метод обучения имеет доступ ко всему $\mathcal{X}^{\mathcal{Y}}$, а не только к $\mathcal{F}$. Таким образом, $\tilde{y}$ построено с помощью метода, отображающего $(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{X}^{\mathcal{Y}}$, а не просто $(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$.

Альтернативное определение предполагает, что построенное решающее правило $\tilde{y} \in \mathcal{F}$, то есть используемый метод выбирает решающее правило из множества $\mathcal{F}$. Так определенное минимаксное значение всегда не меньше, чем предыдущее. Тем не менее в дальнейшем все результаты буду верны для обоих определений.

Разумно также на прямую рассматривать минимаксные значения как некоторую глобальную меру сложности класса $\mathcal{F}$.

§2.1 No Free Lunch Theorem

Легко понять, что если класс $\mathcal{F}$ очень большой, а семейство распределений $\mathcal{P}$ совпадает с $\mathcal{P}^*$, то нельзя рассчитывать на то, что некоторый метод обучения одновременно для всех возможных распределений данных сможет хорошо приблизиться к риску лучшего в классе функционала.

Данное утверждение можно сформулировать в виде так называемой No free lunch теоремы [4].

Теорема 2.1. *Если* $\ell(f, X, Y) = (f(X) - Y)^2$, $|\mathcal{X}| > 2n$, *а* $\mathcal{F} = \mathcal{X}^{\mathcal{Y}}$, *то*

$$V(\mathcal{F}, \mathcal{P}^*, n) \geq \frac{1}{8}.$$

Стоит отметить, что эта теорема основана на очень непрактичном случае. Таким образом, для получения малых минимаксных значений мы должно ограничивать класс $\mathcal{F}$ и/или семейство $\mathcal{P}$.

Для дальнейшего анализа минимаксных значений перейдем к анализу некоторых конкретных методов обучения.

§2.2 Размерность 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

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

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

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

имеет размерность равную ∞, даже несмотря на то, что параметризуется лишь одним параметром.

Семейство классификаторов, обладающее конечной ёмкостью обладает замечательным свойством:

Лемма 2.2 (Зауэр, Вапник-Червоненкис). *Для любого семейства классификаторов с размерностью Вапника–Червоненкиса V для $n \ge V$:*

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

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 оценку на функцию роста:

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

§2.3 Минимизация эмпирического риска и Радемахеровский процесс

Введем несколько удобных обозначений. Пусть

$$L(f) = \mathbf{E}\ell(f, X, Y),$$

причем если f зависит от обучающей выборки, то считаем, что математическое ожидание взято и по обучающей выборке. Если нам важно зафиксировать обучающую выборку в этом математическом ожидании, то введем символ $L(f)|_{(X_i,Y_i)}$. Эмпирическим риском назовем средний риск на обучающей выборке

$$L(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f, X_i, Y_i).$$

Минимизатор эмпирического риска — это любое правило $\hat{y}$, выбранное из $\mathcal{F}$ таким, что эмпирический риск $\hat{y}$ не больше чем эмпирический риск любого другого классификатора из $\mathcal{F}$.

Пусть $f^* = \arg\inf f \in \mathcal{F}$. Рассмотрим избыточный риск для минимизатора эмпирического риска:

$$L(\hat{y})|_{(X_i,Y_i)} - L(f^*) = L(\hat{y})|_{(X_i,Y_i)} - L_n(\hat{y}) + L_n(\hat{y}) - L_n(f^*) + L_n(f^*) - L(f^*) \leq$$
$$L(\hat{y})|_{(X_i,Y_i)} - L_n(\hat{y}) + L_n(f^*) - L(f^*) \leq \sup_{f\in\mathcal{F}}$$

Здесь учтено, что $L_n(\hat{y}) \leq L_n(f^*)$. Беря математическое ожидание от обеих частей неравенства, с учетом $\mathbf{E}(L_n(f^*) - L(f^*)) = 0$, получаем

$$L(\hat{y}) - L(f^*) \leq \mathbf{E}\sup_{f\in\mathcal{F}}(L(f) - L_n(f)) \leq \mathbf{E}\sup_{f\in\mathcal{F}}|L(f) - L_n(f)|.$$

Рассмотрим теперь полученный функционал равномерного по классу решающих правил отклонения среднего риска от эмпирического:

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

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

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

Введем *Радемахеровские случайные величины*, то есть независимые в совокупности (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 математическое ожидание по σ_i как $\mathbf{E}_\sigma$, получаем:

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

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

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

Радемахеровскую сложность можно рассматривать как величину, описывающую сложность класса решающих правил. Чем больше Радемахеровская сложность, тем лучше ошибки $\mathcal{F}$ могут коррелировать 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 [2]) 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.
- $\mathcal{R}_n(A) = \mathcal{R}_n(conv(A))$.

В случае бинарной функции потерь 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 $\left|\sum\limits_{i=1}^n \sigma_i a_i\right|$ 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. В частности можно получить, что для некоторой абсолютной константы C для задачи классификации

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

Таким образом, логарифмический фактор под корнем является излишним. Полезным упражнением является осознание того, что предложенная последовательность неравенств дала нам верхнюю оценку и на минимаксные значения, а именно:

$$V(\mathcal{F}, \mathcal{P}^*, n) \leq \sup_{\mathsf{P} \in \mathcal{P}^*} (L(\hat{y}) - L(f^*)) \leq 2 \sup_{\mathsf{P} \in \mathcal{P}^*} \mathcal{R}(\mathcal{F}) \leq C\sqrt{\frac{V}{n}}.$$

Напомним, что данные неравенства выполнены именно для задачи классификации, в которой $\mathcal{F}$ имеет конечную ёмкость V.

§2.4 Massart's low noise condition

В этом разделе мы кратко обсудим некоторые мощные теоретические результаты, с помощью которых можно значительно обобщить и улучшить результаты, описанные в последнем разделе.

Опять же будем работать с задачей классификации и случаем конечной размерности Вапника–Червоненкиса. Для выделенного семейства решающих правил $\mathcal{F}$ выделим семейство распределений $\mathcal{P}(\mathcal{F})$, такое что в нем содержатся только те распределения, байсовское решающее правило которых принадлежит $\mathcal{F}$. Опять же напомним, что байесовское решающее правило в задаче классификации строиться в зависимости от распределения.

В предыдущем разделе мы кратко объяснили почему минимаксные значения ограничены сверху величиной порядка $O(\frac{1}{\sqrt{n}})$. При некоторых достаточно слабых предположения можно показать, что существуют две константы $C_1, C_2 > 0$, такие что

$$C_1\sqrt{\frac{V}{n}} \leq V(\mathcal{F}, \mathcal{P}(\mathcal{F}), n) \leq C_2\sqrt{\frac{V}{n}}.$$

Таким образом, в очень общей постановке в минимаксном смысле не существует метода, который был бы лучше чем простая минимизация эмпирического риска. Конечно, на практике такая ситуация не наблюдается и метод минимизации эмпирического риска не только часто сложен в имплементации, но и не всегда приводит к поиску хорошего решающего правила. Разумно предположить, что основная необоснованность минимаксного подхода в данном случае заключается в учете слишком большого количества распределений. Таким образом, разумно было бы вычленить только те распределения данных, которые соответствуют в некотором смысле реальной ситуации.

Рассмотрим некоторое условие, а именно пусть для фиксированного распределения для всех $X \in \mathcal{X}$ для определенного выше байесовского решающего правила

$$|2\eta(x) - 1| \geq h,$$

где h некоторая константа. Подобные условия называются условиями малого шума (low noise condition). Идея в них следующая, если мы точно знаем что h близко к единице значит байесовское решающее правило хорошо классифицирует объекты. Основные неточности минимаксных оценок заключаются как раз в том, что для некоторых распределений с точки зрения их байесовских решающих правил отнесение объекта к одному из классов практически равновероятно.

Поэтому для фиксированного $h > 0$ введем семейство распределений $\mathcal{P}(h, \mathcal{F})$ состоящее из всех распределений из $\mathcal{P}(\mathcal{F})$, для которых выполнено условие малого шума с параметром h.

Тогда имеет место более продвинутая оценка[3]:

$$C_1 \min\left\{\sqrt{\frac{V}{n}}, \frac{V}{nh}\right\} \leq V(\mathcal{F}, \mathcal{P}(h, \mathcal{F}), n) \leq \begin{cases} C_2\sqrt{\frac{V}{n}}, & \text{if } h \leq \sqrt{\frac{V}{n}} \\ C_3\frac{V}{nh}\left(1 + \log(\frac{nh^2}{V})\right), & \text{if } h > \sqrt{\frac{V}{n}}. \end{cases}$$

Таким образом, если $h > \sqrt{\frac{V}{n}}$ мы не только улучшаем порядки, но и теряем условие на оптимальность в минимаксном смысле минимизатора эмпирического

риска. Оказывается, что присутствие этого логарифмического фактора существенно. Он может быть опущен только за счет дополнительных ограничений на семейство $\mathcal{F}$.

Список литературы

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

[2] *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.

[3] *Massart P., Nedelec E.* Risk bounds for statistical learning // 2006. — Volume 34, Pp. 2326-2366.

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