Семинары 6–7.
ММП, весна 2013
26 марта

Илья Толстихин
iliya.tolstikhin@gmail.com

Темы семинара:

- Оценки обобщающей способности;
- Оценка для одной функции;
- Оценка для конечного класса функций;
- Оценка для счетного класса функций: Бритва Оккама;
- Оценка для общего случая: функция роста, VC-размерность, VC-оценка;

В начале этого семинара приведем неравенство Хевдинга, полученное нами на прошлом семинаре.

Теорема 0.1 (неравенство Хевдинга) *Пусть $\xi_1, \ldots, \xi_n$ — последовательность ограниченных и независимых случайных величин, таких что $\xi_i \in [a_i, b_i]$ с вероятностью 1. Тогда для любого $t > 0$ справедливо:*

$$\mathbb{P}\{S_n - \mathbb{E}[S_n] \geqslant t\} \leqslant \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right);$$

$$\mathbb{P}\{S_n - \mathbb{E}[S_n] \leqslant -t\} \leqslant \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right);$$

$$\mathbb{P}\{\left|S_n - \mathbb{E}[S_n]\right| \geqslant t\} \leqslant 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

1 Оценки обобщающей способности

Мы освоили часть необходимого нам материала из теории вероятностей и готовы приступить к получению оценок обобщающей способности. Этот семинар мы начнем с рассмотрения общей **вероятностной** постановкой задачи обучения по прецедентам, принятой в теории статистического обучения (Statistical Learning Theory, SLT). В этом разделе мы еще раз вспомним ряд понятий и определений, с которыми мы сталкивались на протяжении предыдущей части курса.

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

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

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

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

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

всюду далее нижний индекс мат. ожидания (где это неочевидно) будет указывать, по какой мере идет интегрирование.

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

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

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

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

Введем удобное обозначение

$$L(g) = \mathsf{E}_{(X,Y)\sim\mathsf{P}}\,\ell\big(g(X),Y\big)$$

— средний риск алгоритма g. Из технических соображений, мы будем предполагать, что потери $\ell\big(g(X),Y\big)$ для функций $g \in \mathcal{G}$ **равномерно ограничены** — иными словами, существует такое неотрицательное число U, что

$$\sup_{g\in\mathcal{G}}\sup_{(X,Y)\in\mathbb{X}\times\mathbb{Y}}\big|\ell\big(g(X),Y\big)\big| \leqslant U.$$

Для удобства будем просто полагать, что $\ell\colon \mathbb{Y}^2 \to [\,0,1]$, то есть $U = 1$.

Наша задача сводится к

$$L(g) \to \min_{g\in\mathcal{G}}. \tag{1}$$

Проблема, конечно, в том, что нам неизвестно распределение P. Зато мы знаем поведение потерь $\ell\big(g(X),Y\big)$, $g \in \mathcal{G}$ на обучающей выборке $\{(X_i,Y_i)\}_{i=1}^n$. Используя **эмпирическое распределение**

$$\mathsf{P}_n(X,Y) = \frac{1}{n}\sum_{i=1}^{n} I\big((X,Y)=(X_i,Y_i)\big)$$

(дискретное распределение с вероятностями $1/n$ в точках обучающей выборки), введем понятие **эмпирического риска**:

$$L_n(g) = \mathsf{E}_{(X,Y)\sim\mathsf{P}_n}\ell\big(g(X),Y\big) = \frac{1}{n}\sum_{i=1}^{n}\ell\big(g(X_i),Y_i\big).$$

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

$$L_n(g) \to \min_{g\in\mathcal{G}}. \tag{2}$$

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

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

- Если не ограничивать $\mathcal{G}$ вообще, то возможна ситуация очевидного переобучения: $L_n(g) = 0$, $L(g) = 1$. Алгоритм не ошибается на обучающей выборке и ошибается на всех прочих объектах с вероятностью 1.

- Если класс $\mathcal{G}$ очень мал, то $\min_{g\in\mathcal{G}} L(g)$ будет очень далек от байесовского риска.
- Если $\mathcal{G}$ сильно увеличить, то, во-первых, снова велик шанс переобучиться, и, во-вторых, становится сложным решить задачу (2).

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

1. Оценка **избыточного риска** функции g (excess risk bound):
$$\mathcal{E}(g) = \mathcal{E}_P(g) = L(g) - \inf_{g\in\mathcal{G}} L(g) \leqslant B_1(n,\mathcal{G}). \tag{3}$$
2. Оценка **обобщающей способности** функции g (error bound):
$$L(g) - L_n(g) \leqslant B_2(n,\mathcal{G}). \tag{4}$$

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

Для дальнейших рассуждений крайне важно отметить, что $L(\hat{g}_n)$ — **случайная величина**, поскольку функция $\hat{g}_n$ выбирается на основе **случайной** выборки $\{(X_i,Y_i)\}_{i=1}^n$. Она может быть формально записана в виде условного матожидания:
$$L(\hat{g}_n) = \mathsf{E}_{(X,Y)\sim\mathsf{P}}\big\{\ell\big(\hat{g}_n(X),Y\big)|\{(X_i,Y_i)\}_{i=1}^n\big\}$$
— тогда становится очевидна ее зависимость от обучающей выборки. Поэтому неравенства вида (3) и (4) для функции $\hat{g}_n$ будут всегда иметь вероятностный характер, например:
$$\mathbb{P}\{\mathcal{E}(\hat{g}_n) \geqslant t\} \leqslant \varepsilon(t,n,\mathcal{G}) \quad \text{или} \quad \mathbb{P}\{L(\hat{g}_n) - L_n(\hat{g}_n) \geqslant t\} \leqslant \varepsilon(t,n,\mathcal{G}), \tag{5}$$
где распределение $\mathbb{P}$ — декартово произведение $\mathsf{P}^{\otimes n}$ исходного распределения P — распределение на случайных простых (i.i.d) выборках длиной n из распределения P (обучающих выборках).

Бóльшая часть теории статистического обучения посвящена построению как можно более точных оценок вида (5), учитывающих «структуру» класса функций $\mathcal{G}$. Причем построению как ненаблюдаемых и зависящих от распределения P оценок (distribution dependant bounds), так и оценок, вычислимых по обучающей выборке (data dependant bounds). Очевидно, второй класс оценок представляется нам наиболее полезным. Для начала мы займемся получением простейших оценок обобщающей способности (4).

1.1 Оценка для одной фиксированной функции

Пристально присмотримся к величине $L(g) - L_n(g)$, забыв, для начала, что нас интересует конкретный случай $g = \hat{g}_n$. То есть, зафиксируем какую-то функцию $g \in \mathcal{G}$.

Перед нами разность матожидания случайной величины $\ell\big(g(X), Y\big)$ и ее среднего выборочного значения:

$$L(g) - L_n(g) = \mathsf{E}_{(X,Y)\sim \mathsf{P}} \ell\big(g(X), Y\big) - \frac{1}{n}\sum_{i=1}^{n} \ell\big(g(X_i), Y_i\big).$$

Но мы с вами на прошлом семинаре вывели неравенство Хевдинга 0.1, которое с большой вероятностью ограничивает эту разность! Итак, мы без труда получаем следующую теорему:

Теорема 1.1 (Оценка для одной функции) *Для любой фиксированной функции $g \in \mathcal{G}$ и любого $\delta > 0$ с вероятностью не меньше $1 - \delta$ (относительно реализации обучающей выборки) выполнено:*

$$L(g) \leqslant L_n(g) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \tag{a}$$

$$|L(g) - L_n(g)| \leqslant \sqrt{\frac{\log \frac{2}{\delta}}{2n}}. \tag{b}$$

Задача: *докажите теорему 1.1 (a).*

Задача на дом: *докажите теорему 1.1 (b).*

Здесь очень важно понимать следующее: результат справедлив **для фиксированной функции** и вероятность рассматривается относительно повторного выбора случайной обучающей выборки $\{(X_i, Y_i)\}_{i=1}^n$. Если функция g зависит от обучающей выборки (как в нашем случае $\hat{g}_n$), то этот результат **неприменим**.

Ограниченность этого результата состоит в следующем: утверждается, что для каждой функции g в классе $\mathcal{G}$ существует событие S_g, реализуемое с большой вероятностью, на котором значение $L(f) - L_n(g)$ мало. Однако, ничего не утверждается о связи событий S_g для разных функций класса $\mathcal{G}$ — они могут оказаться совершенно разными для разных функций. Значит, при конкретной реализации обучающей выборки $\{(X_i, Y_i)\}_{i=1}^n$ неравенство 1.1 будет справедливо только для заранее неизвестного **подмножества** класса $\mathcal{G}$. Если посмотреть на рисунок 1, то можно лучше понять картину происходящего (здесь R и R_n обозначают наши L и L_n соответственно). Кривая R обозначает средний риск и она фиксирована. Кривая R_n — эмпирический риск, и она меняется вместе с обучающей выборкой. Неравенство Хевдинга описывает колебание точек $L_n(g)$ фиксированной функции g вокруг значения $L(g)$. Если класс функций достаточно большой, то для конкретной обучающей выборки найдутся функции, для которых $L(g) - L_n(g)$ велико.

Более того, легко построить интуитивно ясный пример, почему неравенства Хевдинга нам недостаточно. Представим, что по обучающей выборке мы выбираем ту функцию, для которой достигается максимум разности $L(g) - L_n(g) \to \min_{g\in\mathcal{G}}$ (функциця, для которой на картинке зазор между двумя кривыми максимален). Очевидно, что для такой функции неравенство Хевдинга не может выполняться.

Как же нам оценить волнующую нас величину $L(\hat{g}_n) - L_n(\hat{g}_n)$? Ниже приведено уже классическое неравенство, в котором берут свое начало многие подходы в теории статистического обучения:

$$L(\hat{g}_n) - L_n(\hat{g}_n) \leqslant \sup_{g\in\mathcal{G}} \big\{L(g) - L_n(g)\big\}. \tag{6}$$

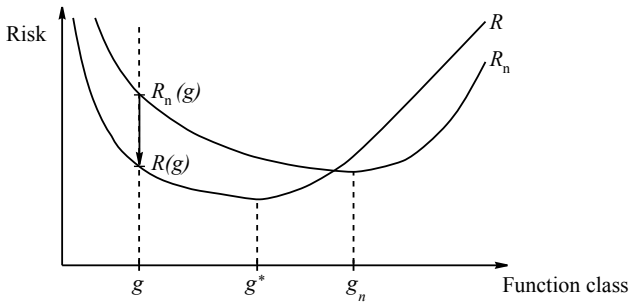

Рис. 1: Эмпирический и средний риск класса функций.

Идея простая: ограничив максимальное по классу отклонение, мы тем самым ограничим его и для минимизатора эмпирического риска $\hat{g}_n$. Это очень простая идея и ей пошли с самого начала. Рассмотрим несколько примеров оценок правой части неравенства в различных ситуациях. Заметим, что в правой части неравенства (6) стоит по-прежнему случайная величина, поэтому мы снова займемся получением вероятностных неравенств, ограничивающих ее.

1.2 Оценка для конечного класса функций

Рассмотрим случай, когда класс функций конечен $\mathcal{G} = \{g_1, \dots, g_N\}$. Покажем, как в этом случае получить равномерный по классу функций $\mathcal{G}$ аналог оценки Хевдинга.

Теорема 1.2 (Оценка для конечного класса) *Для любого конечного класса функций $\mathcal{G} = \{g_1, \dots, g_N\}$ и любой $\delta > 0$ с вероятностью не меньше $1 - \delta$ одновременно для всех $g \in \mathcal{G}$ справедливо:*

$$L(g) \leqslant L_n(g) + \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2n}}. \tag{7}$$

$$|L(g) - L_n(g)| \leqslant \sqrt{\frac{\log N + \log \frac{2}{\delta}}{2n}}. \tag{b}$$

Здача. *Докажите теорему 1.2 (7).*

Здача на дом. *Докажите теорему 1.2 (b).*

Решение. Используя неравенство Буля (union bound) вместе с неравенством Хевдинга, мы получаем:

$$\mathbb{P}\Big\{\sup_{g\in\mathcal{G}}\big\{L(g) - L_n(g)\big\} > t\Big\} = \mathbb{P}\Big\{\exists g \in \{g_1, \dots, g_N\} \colon L(g) - L_n(g) > t\Big\} \leqslant$$

$$\leqslant \sum_{i=1}^{N} \mathbb{P}\{L(g_i) - L_n(g_i) > t\} \leqslant N \exp(-2n\varepsilon^2).$$

Обозначив $\delta = N \exp(-2n\varepsilon^2)$, мы получили желаемый результат.

Обратим внимание, что единственное отличие этого результата от 1.1 — присутствие величины $\log N$ в числителе дроби. Это наша плата за требование равномерного по классу $\mathcal{G}$ контроля уклонения эмпирического риска от равномерного и мы еще ни раз столкнемся с похожими выражениями.

1.3 Оценка для счетного класса функций

Пусть теперь класс функций $\mathcal{G} = \{g_1, g_2 \dots\}$ не более чем счетный.

Теорема 1.3 (Оценка для счетного класса) *Для любого счетного класса функций $\mathcal{G} = \{g_1, g_2, \dots\}$, любой $\delta > 0$ и любой фиксированной весовой функции $p\colon \mathcal{G} \to [0,1]$, такой что $\sum_{g\in\mathcal{G}} p(g) = 1$, с вероятностью не меньше $1-\delta$ одновременно для всех $g \in \mathcal{G}$ справедливо:*

$$L(g) \leqslant L_n(g) + \sqrt{\frac{\log \frac{1}{p(g)} + \log \frac{1}{\delta}}{2n}}. \tag{8}$$

$$|L(g) - L_n(g)| \leqslant \sqrt{\frac{\log \frac{1}{p(g)} + \log \frac{2}{\delta}}{2n}}. \tag{b}$$

Здача. *Докажите теорему 1.3 (8).*

Здача на дом. *Докажите теорему 1.3 (b).*

Решение. В этом случае работает такая же логика, как и в прошлом пункте. Для каждой отдельно взятой функции $g_i \in \mathcal{G}$ мы можем записать

$$\mathbb{P}\left\{ L(g_i) - L_n(g_i) > \sqrt{\frac{\log \frac{1}{\delta(g_i)}}{2n}} \right\} \leqslant \delta(g_i).$$

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

$$\mathbb{P}\left\{ \exists g \in \mathcal{G}\colon L(g) - L_n(g) > \sqrt{\frac{\log \frac{1}{\delta(g)}}{2n}} \right\} \leqslant$$

$$\leqslant \sum_{i=1}^{\infty} \mathbb{P}\left\{ L(g_i) - L_n(g_i) > \sqrt{\frac{\log \frac{1}{\delta(g_i)}}{2n}} \right\} \leqslant \sum_{g_i\in\mathcal{G}} \delta(g_i) = \delta.$$

Обратив вероятность, получаем желаемое утверждения.

Оценка (8) в литературе известна под названием «бритва Оккама». Например, если весовую функцию $p(\cdot)$ мы выберем до наблюдения обучающей выборки таким образом, что бóльший вес получат «более простые» функции $g \in \mathcal{G}$, то посыл последней оценки можно трактовать следующим образом: если у нас есть две функции $g', g'' \in \mathcal{G}$ с одинаковыми значениями эмпирического риска, то истинный средний риск меньше у «более простой» из них. В этих размышлениях понятие «простоты» функции можно трактовать в том же смысле, в каком, например, в SVM извилистые разделяющие поверхности считались более сложными по сравнению с гладкими.

Если $\mathcal{G}$ — множество всех решающих деревьев конечной высоты, то проще те деревья, глубины которых меньше. Заметим, что различных деревьев большой высоты гораздо и гораздо больше, чем деревьев маленькой глубины (вы можете сосчитать аналитически их число и сравнить). То есть сложных функций в классе гораздо больше, чем простых. Это указывает на то, что, вообще говоря, выбрать весовую

функцию p так, чтобы бóльшие веса получали функции более сложные (деревья большой высоты в прошлом примере) по сравнению с более простыми, — задача довольно нетривиальная (вам надо, чтобы функция p была при этом нормирована).

Задача на дом. *Попробуйте доказать теорему 1.2 как следствие теоремы 1.3.*

Еще один важный момент: эта оценка дает нам возможность использовать некие «дополнительные» или *априорные* знания о задаче и семействе $\mathcal{G}$. Если бы нам посчастливилось сосредоточить всю величину $p(\cdot)$ на минимизаторе эмпирического риска $\hat{g}_n$, то мы бы в правой части (8) получили в точности величину, стоящую в правой части оценки для одной функции 1.1 — полный аналог оценки типа Хевдинга в случая единственной функции в семействе: $\mathcal{G} = \{g_0\}$. Однако, дело здесь в том, что мы должны выбрать $p(\cdot)$ **до того, как увидим обучающую выборку**. А значит, мы не можем предугадать, какая из функций будет минимизировать эмпирический риск. Тем не менее хороший выбор величины $p(\cdot)$ позволяет улучшать оценку. Можно считать, что описанная процедура — в некотором роде способ улучшения простго неравенства Буля.

1.4 Бесконечный несчетный класс функций

Пусть множество функций $\mathcal{G}$ на этот раз бесконечно и несчетно. Сейчас мы ограничим рассмотрение случаем $\ell(y, y^*) = I(y \neq y^*)$ — бинарной функцией потерь.

Теперь логика прошлых доказательств уже не работает — например, потому что мы не можем применить неравенство Буля (у нас есть счетная аддитивность, но нет несчетной). Обозначим обучающую выборку, как обычно, с помощью $X^n = \{(X_i, Y_i)\}_{i=1}^n$.

Нам пригодится понятие **функции роста** $S_{\mathcal{G}}(n)$:

Определение 1.1 (Функция роста) *Функцией роста класса* $\mathcal{G}$

$$S_{\mathcal{G}}(n) = \sup_{X^n} |\mathcal{G}_{X^n}| = \sup_{X^n} \left| \left\{ \left(\ell\big(g(X_1), Y_1\big), \ldots, \ell\big(g(X_n), Y_n\big) \right) \right\} : g \in \mathcal{G} \right|.$$

То есть $S_{\mathcal{G}}(n)$ подсчитывает максимальную мощность множества n-мерных векторов потерь функций класса $\mathcal{G}$ на обучающей выборке длины n.

Задача. *Докажите следующее элементарное неравенство* $S_{\mathcal{F}}(n) \leqslant 2^n$.

Оказывается, справедлива следующая оценка

Теорема 1.4 (оценка Вапника–Червоненкиса) *Для любого класса функций* $\mathcal{G}$ *и любой* $\delta > 0$ *с вероятностью не меньше* $1 - \delta$ *относительно случайной реализации обучающей выборки одновременно для всех* $g \in \mathcal{G}$ *справедливо:*

$$L(g) \leqslant L_n(g) + 2\sqrt{2\frac{\log S_{\mathcal{G}}(2n) + \log \frac{1}{\delta}}{n}}. \tag{9}$$

$$|L(g) - L_n(g)| \leqslant 2\sqrt{2\frac{\log S_{\mathcal{G}}(2n) + \log \frac{2}{\delta}}{n}}. \tag{b}$$

Задача на дом. *Докажите оценку Вапника–Червоненкиса (b). Для этого, заглянув в доказательство леммы симметризации [2][Lemma 2], убедитесь, что симметризация выполнена и для* $L_n(g) - L(g)$.

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

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

$$\mathbb{P}\Big\{\sup_{g\in\mathcal{G}}\big\{L(f) - L_n(g)\big\} \geqslant t\Big\} \leqslant 2\mathbb{P}'\Big\{\sup_{g\in\mathcal{G}}\big\{L'_n(g) - L_n(g)\big\} \geqslant t/2\Big\}. \tag{10}$$

Здесь $L'_n(g) = \frac{1}{n}\sum_{i=1}^n \ell\big(g(X'_i), Y'_i\big)$ — эмпирический риск функции g относительно независимой «призрачной» (ghost) выборки $\{(X'_i, Y'_i)\}_{i=1}^n$ — независимой копии обучающей выборки. Соответственно, вероятность $\mathbb{P}'$ в правой части — вероятность относительно случайной реализации объединения обучающей и призрачной выборок (i.i.d. выборки из нашего неизвестного распределения P длиной $2n$). Таким образом симметризация позволяет заменить неизвестный средний риск функции ее средним выборочным значением на еще одной независимой выборке. Подобное введение дополнительной **рандомизации** — частый прием в теории вероятностей. В результате правая часть неравенства (10) зависит лишь от проекции класса $\mathcal{G}$ на двойную выборку $\{(X_1, Y_1), \ldots, (X_n, Y_n), (X'_1, Y'_1), \ldots, (X'_n, Y'_n)\}$ длины $2n$. Поскольку мощность этой проекции **конечна** (напомним: потому что мы используем сейчас бинарную функцию потерь, а множество бинарных векторов длины $2n$ как ни крути конечно), мы можем снова использовать неравенство Буля вместе со следующей слегка модифицированной версией неравенства Хевдинга:

Теорема 1.5 (симметризованное неравенство Хевдинга) *Для любой фиксированной функции* $g \in \mathcal{G}$ *и любого* $\delta > 0$ *с вероятностью не меньше* $1 - \delta$ *справедливо:*

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

Задача на дом. *Докажите теорему 1.5.*

Собрав все кубики вместе, получим:

$$\begin{aligned}\mathbb{P}\Big\{\sup_{g\in\mathcal{G}}\big\{L(g) - L_n(g)\big\} \geqslant t\Big\} &\leqslant 2\mathbb{P}'\Big\{\sup_{g\in\mathcal{G}}\big\{L'_n(g) - L_n(g)\big\} \geqslant t/2\Big\} = \\ &= 2\mathbb{P}'\Big\{\sup\nolimits_{g\in\mathcal{G}_{\{X^n, X'^n\}}}\big\{L'_n(g) - L_n(g)\big\} \geqslant t/2\Big\} \leqslant \\ &\leqslant 2S_{\mathcal{G}}(2n)\mathbb{P}\{L'_n(g) - L_n(g) > t/2\} \leqslant 2S_{\mathcal{G}}(2n)\exp\Big(-\frac{nt^2}{8}\Big).\end{aligned}$$

Обращение вероятности завершает доказательство оценки Вапника–Червоненкиса. □

Определение 1.2 (VC-размерность) *Размерностью Вапника–Червоненкиса (VC dimension) класса функций* $\mathcal{G}$ *называется наибольшее целое число* n, *такое что*

$$S_{\mathcal{G}}(n) = 2^n.$$

Размерность Вапника–Червоненкиса класса $\mathcal{G}$ *мы будем обозначать* $\mathrm{VC}(\mathcal{G})$.

Заметим, что если $S_{\mathcal{G}}(n) = 2^n$, это означает что существует обучающая выборка, состоящая из n различных объектов, которую классификаторы из нашего класса $\mathcal{G}$ могут разбить на два класса всевозможными способами.

Зачем мы ввели понятие VC-размерности? Что оно значит? В свете оценки Вапника–Червоненкиса следовало бы ожидать, что VC-размерность класса функций $\mathcal{G}$ как-то поможет нам ограничить сверху функцию роста этого класса $S_{\mathcal{G}}(n)$. Но как именно? И какой, собственно, результат мы хотим получить с помощью оценки Вапника–Червоненкиса? Все, что пока ясно — что при $k \leqslant \mathrm{VC}(\mathcal{G})$ мы получаем $S_{\mathcal{G}}(k) = 2^k$, а для $k > \mathrm{VC}(\mathcal{G})$ мы получаем $S_{\mathcal{G}}(k) < 2^k$. И, казалось бы, это нам ничего хорошего не дает и вот почему. Все прошлые оценки утверждали, что при росте размера обучающей выборки n эмпирические риски $L_n(g)$ функций класса $\mathcal{G}$ *равномерно* сходились к их истинным средним рискам $L(g)$. Чтобы получить что-то похожее в текущей ситуации, нам необходимо, чтобы при $n \to \infty$ было выполнено $\log S_{\mathcal{G}}(2n) = o(n)$, то есть что $S_{\mathcal{G}}(2n)$ растет медленнее экспоненты по n.

Нашу цепочку рассуждений заканчивает следующий комбинаторный результат.

Лемма 1.2 (Лемма Сауэра) *Пусть $\mathcal{G}$ — класс с конечной VC-размерностью*

$$\mathrm{VC}(\mathcal{G}) < \infty.$$

Тогда для всех натуральных n справедливо:

$$S_{\mathcal{G}}(n) \leqslant \sum_{i=1}^{\mathrm{VC}(\mathcal{G})} C_n^i,$$

и для всех $n \geqslant \mathrm{VC}(\mathcal{G})$:

$$S_{\mathcal{G}}(n) \leqslant \left(\frac{e \cdot n}{\mathrm{VC}(\mathcal{G})}\right)^{\mathrm{VC}(\mathcal{G})}.$$

Мы наблюдаем удивительную картину: если класс функций имеет конечную VC-размерность, его функция роста, начиная с этой VC-размерности, растет полиномиально! (изобразить на доске). Это чрезвычайно важный комбинаторный результат, позволивший Вапнику и Червоненкису заложить основы теории эмпирических процессов.

Задача. *Докажите, что для класса функций $\mathcal{G}$ с конечной VC-размерностью $\mathrm{VC}(\mathcal{G}) = h$, величина $\sup_{g \in \mathcal{G}} \{L(g) - L_n(g)\}$ имеет порядок $\sqrt{h \log n} n$, то есть мы получаем равномерную по классу $\mathcal{G}$ сходимость значений эмпирических рисков к истинным средним рискам функций.*

Подытоживая, стоит отметить, что функция роста и размерность Вапника–Червоненкиса характеризуют *эффективную мощность} класса функций.*

Список литературы

[1] Stephane B., Lugosi G., and Bousquet O. *Concentration inequalities. — Machine Learning Summer School 2003.*
www. econ. upf. edu/ ~lugosi/ mlss_ conc. pdf

[2] Bousquet O., Boucheron S., Lusogi G. *Introduction to statistical learning theory. — Advanced Lectures in Machine Learning, Springer, pp.169–207, 2004.*
http://www.econ.upf.edu/~lugosi/mlss_slt.pdf