Семинар 5.
ММП, весна 2013
19 марта

Илья Толстихин
iliya.tolstikhin@gmail.com

Темы семинара:

- Оценки обобщающей способности;
- Неравенства Буля, Маркова, Чебышева, метод Чернова;
- Лемма Хевдинга, неравенство Хевдинга.

1 Оценки обобщающей способности

Мы уже знаем немало способов построения отдельных алгоритмов классификации и регрессии. Мы также узнали что-то про различные способы построения композиций из набора алгоритмов. Чего мы до сих пор практически не затрагивали — так это строгого теоретического анализа качества работы настроенной модели на будущих данных. Мы рассмотрим классический вероятностный подход к изучению обобщающей способности алгоритмов, который принято называть Теорией Статистического Обучения (Statistical Learning Theory или SLT). Основы этого подхода были заложены в конце 70-х годов в работах советских ученых В. Н. Вапника и А. Я. Червоненкиса, которые получили широкое признание (и не только в теории машинного обучения) определением т. н. размерности Вапника–Червоненкиса (VC-dimension).

За 30 с лишним лет существования SLT существенно развивалась. Именно этому подходу принадлежат первые теоретические обоснования успешной работы таких алгоритмов, как SVM и бустинг. Новый толчок SLT пришёлся на конец 90-х годов, что связано с полученным французским математиком Мишелем Талаграном в 95-ом году неравенством (Talagrand's concentration inequality). После этого SLT была существенно развита рядом ученых. Среди прочих открытий стоит отметить введение в теорию машинного обучения В. Колчинским и Д. Панченко понятия Радемахеровской сложности (Rademacher complexity).

Мы рассмотрим самые базовые результаты теории статистического обучения. Мы также познакомимся с некоторыми результаты из теории вероятностей, на которых основано большинство из них: неравенство Буля, неравенства Хевдинга, Бернштейна и МакДиармида, неравенство симметризации, размерность Вапника–Червоненкиса и Радемахеровская сложность. В конце мы увидим, почему бустинг устойчив к переобучению.

1.1 Теория вероятностей

Начнем со следующего неравенства Маркова:

Теорема 1.1 (Неравенство Маркова) *Для любой неотрицательной случайной величины ξ и произвольного $t > 0$:*

$$\mathbb{P}\{\xi \geqslant t\} \leqslant \frac{\mathbb{E}[\xi]}{t}.$$

Задача на дом. *Докажите, что для любой неотрицательной случайной величины ξ выполнено*

$$\mathbb{E}[\xi] = \int_0^\infty \mathbb{P}\{\xi \geqslant t\} dt.$$

Пользуясь этим результатом докажите теорему 1.1.

Задача. *Докажите следующую теорему — неравенство Чебышева:*

Теорема 1.2 (Неравенство Чебышева) *Для любой случайной величины ξ и любого $t > 0$ выполнено:*

$$\mathbb{P}\{|\xi - \mathbb{E}[\xi]| \geqslant t\} \leqslant \frac{\mathrm{Var}[\xi]}{t^2}.$$

Доказательство.

$$\mathbb{P}\{|\xi - \mathbb{E}[\xi]| \geqslant t\} = \mathbb{P}\{(\xi - \mathbb{E}[\xi])^2 \geqslant t^2\} \leqslant \frac{\mathbb{E}\left[(\xi - \mathbb{E}[\xi])^2\right]}{t^2}.$$

■

Мы с вами получили первый пример неравенства, контролирующего отклонение случайной величины от ее мат. ожидания с большой вероятностью. Такие неравенства принято называть *неравенствами концентрации меры* [1] (concentration inequalities). Неравенство Чебышева — вероятно, один из самых простых и слабых результатов. Подобные неравенства лежат в основе теории статистического обучения, и вскоре мы увидим более сильные результаты.

Сейчас мы опишем процедуру, известную как *метод Чернова*, на которой основан вывод (за редким исключением) всех неравенств концентрации (включая упомянутое выше неравенство Талаграна), с которыми вы можете столкнуться в SLT.

Задача. *Докажите, что для любой случайной величины ξ и любого $t > 0$ справедливо*

$$\mathbb{P}\{\xi \geqslant t\} \leqslant \min_{s>0} \mathbb{E}\left[e^{s(\xi - t)}\right] = \min_{s>0} \frac{\mathbb{E}\left[e^{s\xi}\right]}{e^{st}}.$$

Решение: для произвольной $s > 0$:

$$\mathbb{P}\{\xi \geqslant t\} = \mathbb{P}\{e^{s\xi} \geqslant e^{st}\} \leqslant \frac{\mathbb{E}\left[e^{s\xi}\right]}{e^{st}}.$$

Мы видим, что если мы научимся ограничивать *производящую функцию* (или *преобразование Лапласа*) случайной величины ξ: $\mathbb{E}\left[e^{s\xi}\right]$, где $s > 0$:

$$\mathbb{E}\left[e^{s\xi}\right] \leqslant F(s), \tag{1}$$

то мы автоматически получаем оценку для ее *больших отклонений*:

$$\mathbb{P}\{\xi \geqslant t\} \leqslant \frac{F(s)}{e^{st}},$$

которую затем мы можем минимизировать по $s > 0$, получив наиболее точную оценку:

$$\mathbb{P}\{\xi \geqslant t\} \leqslant \min_{s>0} \left(\frac{F(s)}{e^{st}} \right).$$

Как правило, наиболее сложной задачей является получение верхних оценок для преобразования Лапласа (1). Сейчас мы приведем очень простую и изящную лемму Хевдинга, которая ограничивает преобразование Лапласа для ограниченной случайной величины.

Лемма 1.1 (Хевдинг) *Для любой случайной величины ξ, такой что $\mathbb{E}[\xi] = 0$ и $a \leqslant \xi \leqslant b$ с вероятностью 1, для любого $s > 0$ справедливо*

$$\mathbb{E}\left[e^{s\xi}\right] \leqslant e^{\frac{s^2(b-a)^2}{8}}.$$

Задача на дом, сложная. *Докажите лемму Хевдинга. Для этого надо воспользоваться выпуклостью функции e^x, а также разложением в ряд Тейлора чего-то еще.*

Описанные выше результаты позволят нам сейчас получить неравенство концентрации для одного очень важного для теории статистического обучения (и теории вероятностей в целом) объекта — а именно, суммы независимых случайных величин.

Сумма независимых случайных величин. Пусть $\xi_1, \ldots, \xi_n$ — последовательность независимых случайных величин. Введем обозначение $S_n = \sum_{i=1}^n \xi_i$. Мы хотим изучать отклонение случайной величины S_n от ее мат. ожидания $\mathbb{E}[S_n]$. То есть получить неравенство концентрации для $S_n - \mathbb{E}[S_n]$.

Задача. *Покажите, что если случайные величины $\xi_1, \ldots, \xi_n$ помимо всего прочего одинаково распределены и $\mathrm{Var}[\xi_i] = \sigma^2$, то для любого $t > 0$*

$$\mathbb{P}\{|S_n - \mathbb{E}[S_n]| \geqslant t\} \leqslant \frac{n\sigma^2}{t^2}.$$

Задача. *При тех же условиях докажите, что для любого $\delta > 0$ с вероятностью $1 - \delta$ выполнено*

$$|S_n - \mathbb{E}[S_n]| < \sqrt{\frac{n\sigma^2}{\delta}}. \tag{2}$$

Только что проделанная нами операция называется *обращением* вероятности.

Докажем, наконец, более сильный результат, известный как неравенство Хевдинга, пользуясь методом Чернова.

Задача. *Пользуясь методом Чернова докажите следующую теорему,*

Теорема 1.3 (неравенство Хевдинга) *Пусть $\xi_1, \dots, \xi_n$ — последовательность ограниченных и независимых случайных величин, таких что $\xi_i \in [a_i, b_i]$ с вероятностью 1. Тогда для любого $t > 0$ справедливо:*

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

Доказательство. Для любых $s > 0$ справедливо

$$\mathbb{P}\{S_n - \mathbb{E}[S_n] \geqslant t\} = \mathbb{P}\{e^{s(S_n - \mathbb{E}[S_n])} \geqslant e^{st}\} \leqslant \frac{\mathbb{E}\left[e^{s(S_n - \mathbb{E}[S_n])}\right]}{e^{st}} = \frac{\mathbb{E}\left[e^{s\sum_{i=1}^n (\xi_i - \mathbb{E}[\xi_i])}\right]}{e^{st}} =$$
$$= \frac{\mathbb{E}\left[\prod_{i=1}^n e^{s(\xi_i - \mathbb{E}[\xi_i])}\right]}{e^{st}} = \frac{\prod_{i=1}^n \mathbb{E}\left[e^{s(\xi_i - \mathbb{E}[\xi_i])}\right]}{e^{st}} \leqslant \frac{\prod_{i=1}^n e^{s^2(b_i - a_i)^2/8}}{e^{st}} = \frac{e^{s^2 \sum_{i=1}^n (b_i - a_i)^2/8}}{e^{st}} \leqslant$$
$$\leqslant \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Последнее неравенство обусловлено выбором $s = 4t/\sum_{i=1}^n (b_i - a_i)^2$. Лемму Хевдинга мы можем применять, поскольку, несмотря на то, что $\xi_i - \mathbb{E}[\xi_i] \in \Big[a_i - \mathbb{E}[\xi_i], b_i - \mathbb{E}[\xi_i]\Big]$, длина интервала все же составляет $b_i - a_i$. ∎

Задача на дом. *Покажите, что также справедливо следующее неравенство:*

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

а значит и

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

Задача. *Обратите результат из прошлой теоремы и сравните полученное выражение с (2). Убедитесь, что мы существенно улучшили прошлый резльтат!*

В конце этого раздела вспомним простое неравенство Буля (*union bound*):

Теорема 1.4 *Для любых двух событий A и B*

$$\mathbb{P}\{A \cup B\} \leqslant \mathbb{P}(A) + \mathbb{P}(B).$$

Задача на дом. *В каких случаях неравенство Буля завышено? А в каких оно обращается в равенство?*

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