Семинар 5.
ММП, весна 2013
19 марта

Илья Толстихин
iliya.tolstikhin@gmail.com

Темы семинара:

- Оценки обобщающей способности;
- Неравенство МакДиармида;
- Симметризация и рademахеровская сложность;
- Оценка для произвольной ограниченной функции потерь.

1 Неравенство МакДиармида

До настоящего момента мы видели неравенства концентрации только для одного типа функций $Z = f(\xi_1, \ldots, \xi_n)$, зависящих от последовательности независимых случайных величин $\xi_1, \ldots, \xi_n$, а именно — для их суммы $Z = \sum_{i=1}^n \xi_i$. Оказывается, подобные неравенства справедливы и для гораздо более общих функций f. Сегодня нам пригодится одно из них, которое в литературе известно как неравенство МакДиармида или *неравенство ограниченных разностей* (bounded difference inequality).

Рассмотрим последовательность независимых случайных величин $\xi_1, \ldots, \xi_n$, принимающих значения в $\mathcal{X}$, и произвольную функцию $f\colon \mathcal{X}^n \to \mathbb{R}$. Нас как и раньше интересует, насколько случайная величина $Z = f(\xi_1, \ldots, \xi_n)$ сосредоточена вокруг своего мат. ожидания $\mathbb{E}[Z]$.

Введем следующее удобное обозначение

$$\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot|\xi_1, \ldots, \xi_i].$$

Тогда $\mathbb{E}_0[Z] = \mathbb{E}[Z]$ и $\mathbb{E}_n[Z] = Z$. Давайте обозначим $\Delta_i = \mathbb{E}_i[Z] - \mathbb{E}_{i-1}[Z]$, тогда

$$Z - \mathbb{E}[Z] = \mathbb{E}_n[Z] - \mathbb{E}_{n-1}[Z] + \mathbb{E}_{n-1}[Z] - \mathbb{E}_{n-2}[Z] + \cdots + \mathbb{E}_1[Z] - \mathbb{E}_0[Z] = \sum_{i=1}^n \Delta_i.$$

Что нам может дать такая запись случайной величины Z (она известна как Doob martingale representation)? Рассмотрим особый случай, когда для всех i с вероятностью 1 выполнено $|\Delta_i| \leqslant c_i$.

Задача [Неравенство Азумы–Хевдинга]. *Докажите, что в этом случае для всех $\lambda > 0$ справедливо следующее неравенство:*

$$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leqslant e^{\lambda^2\left(\sum_{i=1}^n c_i^2\right)/2}.$$

Решение.

$$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] = \mathbb{E}\left[e^{\lambda(\sum_{i=1}^n \Delta_i)}\right] = \mathbb{E}\left[\mathbb{E}_{n-1}\left[e^{\lambda(\sum_{i=1}^n \Delta_i)}\right]\right] = \mathbb{E}\left[e^{\lambda(\sum_{i=1}^{n-1} \Delta_i)}\mathbb{E}_{n-1}\left[e^{\lambda\Delta_n}\right]\right] \leqslant \tag{1}$$

$$\leqslant \mathbb{E}\left[e^{\lambda(\sum_{i=1}^{n-1} \Delta_i)}e^{\lambda^2 c_n^2/2}\right] = e^{\lambda^2 c_n^2/2}\mathbb{E}\left[e^{\lambda(\sum_{i=1}^{n-1} \Delta_i)}\right] \leqslant \cdots \leqslant e^{\lambda^2\left(\sum_{i=1}^n c_i^2\right)/2},$$

где первый знак неравенства соответствует использованию леммы Хевдинга.

Раз мы получили верхнюю оценку для производящей функции случайной величины, как мы знаем, мы легко можем выписать неравенство концентрации для нее:

$$\mathbb{P}\{Z - \mathbb{E}[Z] \geqslant t\} = \mathbb{P}\{e^{\lambda(Z-\mathbb{E}[Z])} \geqslant e^{\lambda t}\} \leqslant e^{\lambda^2\left(\sum_{i=1}^n c_i^2\right)/2 - \lambda t}.$$

Выбрав $\lambda = \frac{t}{\sum_{i=1}^n c_i^2}$, мы окончательно получим

$$\mathbb{P}\{Z - \mathbb{E}[Z] \geqslant t\} \leqslant \exp\left\{-\frac{t^2}{2\sum_{i=1}^n c_i^2}\right\}.$$

Теорема 1.1 (Неравенство МакДиармида) *Пусть существуют действительные числа $c_1, \ldots, c_n$, такие что функция $f\colon \mathcal{X}^n \to \mathbb{R}$ удовлетворяет следующему условию:*

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{X}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1} \ldots, x_n)| \leqslant c_i, \quad \forall i, \tag{b.d.}$$

и пусть $\xi_1, \ldots, \xi_n$ — последовательность независимых случайных величин. Тогда для случайной величины $Z = f(\xi_1, \ldots, \xi_n)$ справедливо:

$$\mathbb{P}\{Z - \mathbb{E}[Z] \geqslant t\} \leqslant \exp\left\{-2\frac{t^2}{\sum_{i=1}^n c_i^2}\right\};$$

$$\mathbb{P}\{\mathbb{E}[Z] - Z \geqslant t\} \leqslant \exp\left\{-2\frac{t^2}{\sum_{i=1}^n c_i^2}\right\};$$

$$\mathbb{P}\{|Z - \mathbb{E}[Z]| \geqslant t\} \leqslant 2\exp\left\{-2\frac{t^2}{\sum_{i=1}^n c_i^2}\right\}.$$

Задача. *Докажите неравенство МакДиармида.*

Мы его уже фактически доказали. Нам остается заметить, что

$$a_i = \inf_{\xi_i} \mathbb{E}[Z|\xi_1, \ldots, \xi_i] - \mathbb{E}[Z|\xi_1, \ldots, \xi_{i-1}] \leqslant$$
$$\leqslant \Delta_i = \mathbb{E}[Z|\xi_1, \ldots, \xi_i] - \mathbb{E}[Z|\xi_1, \ldots, \xi_{i-1}] \leqslant$$
$$\leqslant \sup_{\xi_i} \mathbb{E}[Z|\xi_1, \ldots, \xi_i] - \mathbb{E}[Z|\xi_1, \ldots, \xi_{i-1}] = b_i,$$

причем из условия теоремы следует, что

$$b_i - a_i \leqslant c_i.$$

Таким образом, в (1) мы получим $e^{\lambda^2 c_i^2/8}$. После применения метода Чернова мы получаем желаемый результат.

Обратим внимание, что полученное неравенство чрезвычайно легко применять. Действительно, нам достаточно убедиться, что выполнено условие (b.d.) — то есть, что наша функция не слишком сильно увеличится при изменении только одной переменной. Функции, удовлетворяющие этому условию, принято называть *функциями с ограниченными разностями*. Итак, мы получили неравенство концентрации, справедливое для произвольной функции с ограниченными разностями. Вскоре мы увидим, что в этот класс входят в том числе некоторые интересующие нас функции.

Задача. *Докажите, что в случае $f(\xi_1, \ldots, \xi_n) = \sum_{i=1}^n \xi_i$ неравенство МакДиармида в точности дает неравенство Хевдинга.*

Замечание *Тот из вас, кто успел расправиться с домашним заданием и знает, что такое неравенства Беннета и Бернштейна, и как они связаны с неравенством Хевдинга, с интересом отметит следующий момент. Полученное нами только что неравенство МакДиармида можно считать обобщением неравенства Хевдинга на класс функций, более общий, чем суммы независимых случайных величин. Действительно, неравенство ограниченных сумм дает такую же экспоненциально убывающую оценку хвоста распределения, учитывающую лишь одно свойство функции — ее ограниченность. То же самое происходит в случае неравенства Хевдинга. Более того, как было отмечено, неравенство МакДиармида, будучи применено к сумме независимых случайных величин, даст в точности неравенство Хевдинга. «А нельзя ли получить обобщение неравенств типа Бернштейна на более общий класс функций?» — спросите вы. Ответ положительный, подобные результаты научились относительно легко получать совсем недавно и именно на них основаны все свежие результаты теории статистического обучения [3, 2, 5, 6]. Эта область известна как Concentration-of-measure inequalities, и в ней сегодня работает довольно много математиков.*

2 Радемахеровская сложность

Давайте еще раз выпишем оценку Вапника–Червоненкиса, которую мы получили на прошлом семинаре.

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

Какие главные недостатки мы можем заметить в оценке Вапника–Червоненкиса (кроме ее очевидно сильной завышенности)? Наверно, напрашивается примерно следующие рассуждения. Во-первых, оценка, почему-то, не зависит от распределения P, которым, вообще говоря, вся картина должна определяться. Неважно, что мы его не знаем — нас устроило бы хотя бы присутствие в оценке каких-то выборочных оценок

параметров распределения P. Понятно, что это существенно улучшило бы оценку. Оценка же Вапника–Червоненкиса справедлива сразу для **всех** распределений, в чем, с одной стороны, ее сила, но с другой — слабая сторона, поскольку будучи справедливой сразу для всех распределений, она справелива и для распределения «самого плохого», для которого разница $L(g) - L_n(g)$ может оказаться очень большой. Во-вторых, странно, что оценка практически не зависит от выбранного алгоритма $g \in \mathcal{G}$. Она учитывает лишь эмпирический риск этого алгоритма, но второе слагаемое (которое принято называть *сложностью семейства*, *complexity term*) от алгоритма не зависит. Это, в некотором смысле, глобальная характеристика класса алгоритмов, хотя и более детальная, чем, например, мощность семейства алгоритмов (в случае конечного класса). В-третьих, оценка не зависит от обучающей выборки — она опять же учитывает эмпирический риск, но не более того. Что снова удивительно, ведь обучающая выборка — все, что у нас есть о задаче. Хотелось бы как-то существеннее использовать эту информацию.

С первым недостатком в общем случае справиться сложно — для этого нам надо накладывать различные условия на рассматриваемый класс распределений, выполнение которых проверить обычно достаточно сложно. Несмотря на это в теории классификации получено ряд сильных результатов, дающих, в том числе, скорость равномерной сходимости эмпирических рисков к средним потерям порядка $1/n$ — так называемые Tsybakov and Massart low noise conditions [5].

Со вторым мы уже пытались бороться — неравенство бритвы Оккама отчасти решало эту проблему. Более сильные результаты дает так называемый PAC-Bayes подход. Считается, что этот подход дает наиболее точные оценки обобщающей способности и в данный момент он активно развивается.

Как справиться с третьим мы узнаем сейчас. Заодно мы получим способ получения оценок для произвольных действительнозначных функций потерь, чего не удавалось при использовании исключительно комбинаторной сложности Вапника–Червоненкиса.

2.1 Симметризация, еще одна...

Вспомним, что мы договорились сконцентрировать свое внимание на получении оценок обобщающей способности для метода минимизации эмпирического риска. Для этого мы, пользуясь следующим элементарным неравенством

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

начали строить верхние оценки для $\sup_{g\in\mathcal{G}}\big(L(g) - L_n(g)\big)$ во всевозможных случаях. Следующая теорема открывает нам еще один путь получения такого сорта оценок.

Для фиксированного класса функций $\mathcal{G}$ введем понятие радемахеровской сложности:

Определение 2.1 (Радемахеровская сложность) *Радемахеровской сложностью класса $\mathcal{G}$ при фиксированной функции потерь ℓ назовем*

$$\mathcal{R}(\ell \circ \mathcal{G}) = \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell\big(g(X_i), Y_i\big)\right],$$

где $\sigma_1, \ldots, \sigma_n$ — последовательность независимых радемахеровских случайных величин, принимающих значения $+1$ и -1 с вероятностями $1/2$. Математическое ожидание в определении берется одновременно по объектам обучающей выборки и радемахеровским случайным величинам.

Условной радемахеровской сложностью класса $\mathcal{G}$ при фиксированной функции потерь ℓ назовем

$$\mathcal{R}_n(\ell \circ \mathcal{G}, X^n) = \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell\big(g(X_i), Y_i\big)\Big| X^n\right],$$

где математическое ожидание берется только по случайным знакам при фиксированной обучающей выборке.

Давайте сразу же заметим, что выполнено следующее соотношение:

$$\mathbb{E}\left[\mathcal{R}_n(\ell \circ \mathcal{G}, X^n)\right] = \mathcal{R}(\ell \circ \mathcal{G}, X^n). \tag{1}$$

Теорема 2.1 (Симметризация для мат. ожиданий) *Для любого класса функций $\mathcal{G}$ выполнено*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\big(L(g) - L_n(g)\big)\right] \leqslant 2\mathcal{R}(\mathcal{G}).$$

Докажем это неравенство.

$$\begin{aligned}
\mathbb{E}\left[\sup_{g\in\mathcal{G}}\big(L(g) - L_n(g)\big)\right] &= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\big(\mathbb{E}'[L'_n(g)] - L_n(g)\big)\right] = \mathbb{E}\left[\sup_{g\in\mathcal{G}}\big(\mathbb{E}'[L'_n(g) - L_n(g)]\big)\right] \leqslant \\
&\leqslant \mathbb{E}\mathbb{E}'\left[\sup_{g\in\mathcal{G}}\big(L'_n(g) - L_n(g)\big)\right] = \\
&= \mathbb{E}\mathbb{E}'\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{n}\sum_{i=1}^{n}\Big(\ell\big(g(X_i), Y_i\big) - \ell\big(g(X'_i), Y'_i\big)\Big)\right)\right] = \\
&= \mathbb{E}\mathbb{E}_\sigma\mathbb{E}'\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\Big(\ell\big(g(X_i), Y_i\big) - \ell\big(g(X'_i), Y'_i\big)\Big)\right)\right] \leqslant \\
\leqslant \mathbb{E}\mathbb{E}_\sigma\mathbb{E}'\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell\big(g(X_i), Y_i\big)\right)\right] &+ \mathbb{E}\mathbb{E}_\sigma\mathbb{E}'\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell\big(g(X'_i), Y'_i\big)\right)\right] = \\
&= 2\mathbb{E}_\sigma\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell\big(g(X_i), Y_i\big)\right)\right] = 2\mathcal{R}(\mathcal{G}).
\end{aligned}$$

Мы сначала ввели призрачную выборку X'^n. Затем воспользовались очевидным свойством супремума. Затем м заметили, что, поскольку X^n и $X'n$ независимы и имеют одинаковое распределение, то под знаком мат. ожидания мы вольны менять местами пары (X_i, Y_i) и (X'_i, Y'_i) местами. Таким образом, ничего не изменится, если мы добавим перед скобками случайные знаки σ_i (замена пар местами), а усредним по этим знакам. Снова воспользовавшись свойствами супремума, мы можем «разорвать» его по знаку вычитания. Наконец, мы получаемый желаемый результат, поскольку подынтегральные выражения в обоих слагаемых одинаково распределены.

2.2 Получаем еще одну оценку

Зачем же нам верхняя оценка мат. ожидания $\mathbb{E}\left[\sup_{g\in\mathcal{G}}\big(L(g)-L_n(g)\big)\right]$, когда на самом деле нас интересует верхняя оценка того, что стоит **под** математическим ожиданием? Здесь надо привыкнуть к стандартному трюку в теории концентрации: хочешь с большой вероятностью оценить случайную величину сверху — получи результат концентрации (тем самым «привязав» случайную величину к ее математическому ожиданию), а затем получи верхнюю оценку мат. ожидания. Так мы сейчас и поступим.

Теорема 2.2 *Для произвольного класса алгоритмов $\mathcal{G}$ и произвольной ограниченной единицей функции потерь ℓ, для любого $\delta>0$ с вероятностью не меньше, чем $1-\delta$, одновременно для всех $g\in\mathcal{G}$ справедливо:*

$$L(g)\leqslant L_n(g)+2\mathcal{R}(\ell\circ\mathcal{G})+\sqrt{\frac{\log\frac{1}{\delta}}{2n}},$$

а также с вероятностью не меньше, чем $1-\delta$, одновременно для всех $g\in\mathcal{G}$ справедливо:

$$L(g)\leqslant L_n(g)+2\mathcal{R}_n(\ell\circ\mathcal{G})+3\sqrt{\frac{\log\frac{2}{\delta}}{n}}.$$

Задача. *Докажите последнюю теорему.*

Воспользуемся планом, описанным выше. Для того, чтобы сварить суп, нам понадобится результат концентрации. Концентрации для какой случайной величины? Поскольку мы хотим в итоге воспользоваться результатом 2.1, то нам нужно неравенство концентрации для $\sup_{g\in\mathcal{G}}\big(L(g)-L_n(g)\big)$. Есть ли у нас такой результат? Конечно же есть! Давайте убедимся, что функция

$$F\Big((X_1,Y_1),\dots,(X_n,Y_n)\Big)=\sup_{g\in\mathcal{G}}\left(L(g)-\frac{1}{n}\sum_{i=1}^{n}\ell\big(g(X_i),Y_i\big)\right)$$

удовлетворяет условию ограниченных разностей.

Задача. *Докажите этот факт и найдите для этой функции константы c_i из определения свойства ограниченных разностей.*

Если заменить одну отдельно взятую пару (X_i,Y_i) на другую (X_i',Y_i'), то легко убедиться, что значение функции изменится не больше, чем на $1/n$. Таким образом, мы можем применить неравенство МакДиармида для нашей функции с $c_i=1/n$ и получить, что

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}}\big(L(g)-L_n(g)\big)-\mathbb{E}\left[\sup_{g\in\mathcal{G}}\big(L(g)-L_n(g)\big)\right]\geqslant t\right\}\leqslant e^{-2nt^2},$$

или, после обращения, для любого $\delta>0$ с вероятностью не менее $1-\delta$

$$\sup_{g\in\mathcal{G}}\big(L(g)-L_n(g)\big)-\mathbb{E}\left[\sup_{g\in\mathcal{G}}\big(L(g)-L_n(g)\big)\right]\leqslant\sqrt{\frac{\log\frac{1}{\delta}}{2n}}. \tag{2}$$

Полученное неравенство вместе с 2.1 завершают доказательство первой части теоремы.

Вторую часть мы докажем повторным применением только что описанного трюка, на этот раз к функции $\mathcal{R}_n(\ell \circ \mathcal{G})$. Действительно,

$$\mathcal{R}_n(\ell \circ \mathcal{G}) = G\Big((X_1, Y_1), \dots, (X_n, Y_n)\Big) = \mathbb{E}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell\big(g(X_i), Y_i\big) \Big| X^n\right].$$

Может быть, эта функция тоже удовлетворяет условию ограниченных разностей? Давайте проверим.

Задача. *Проверьте, удовлетворяет ли приведенная выше функция условию ограниченных разностей. Если да, то найдите константы c_i.*

Итак, совсем несложно убедиться, что эта функция снова удовлетворяет условию ограниченных разностей с константами $c_i = 1/n$. Тогда с учетом (1) неравенство МакДиармида даст нам

$$\mathbb{P}\left\{\mathbb{E}\left[\mathcal{R}_n(\ell \circ \mathcal{G}) - \mathcal{R}_n(\ell \circ \mathcal{G})\right] \geqslant t\right\} \leqslant e^{-2nt^2}$$

или после обращения

$$\mathcal{R}(\ell \circ \mathcal{G}) \leqslant \mathcal{R}_n(\ell \circ \mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \tag{3}$$

Если мы теперь объединим с помощью неравенства Буля (2) и (3) с вероятностями $\delta/2$, то получим, что с вероятностью не менее $1 - \delta$ справедливо

$$\sup_{g \in \mathcal{G}}\big(L(g) - L_n(g)\big) \leqslant 2\mathcal{R}_n(\ell \circ \mathcal{G}) + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}} + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Список литературы

[1] *Boucherone S., Lugosi G., and Bousquet O.* Concentration inequalities. — Machine Learning Summer School 2003.
www.econ.upf.edu/~lugosi/mlss_conc.pdf

[2] *Lugosi G.* Concentration inequalities. — Machine Learning Summer School 2012. Videolectures.net
http://videolectures.net/gabor_lugosi/

[3] *Bousquet O.* Advanced Statistical Learning Theory. — Machine Learning Summer School 2007. Videolectures.net
http://videolectures.net/mlss07_bousquet_slt/

[4] *V. Koltchinskii.* Excess risk bounds in machine learning. — Machine Learning Summer School 2009. Videolectures.net
http://videolectures.net/mlss09us_koltchinskii_berml/

[5] *Koltchinskii V.* Oracle inequalities in empirical risk minimization and sparse recovery problems. — École d'Été de Probabilités de Saint-Flour XXXVIII-2008, 2011. http://www.stat.washington.edu/jaw/COURSES/EPWG/PAPERS-11/sflour_book.pdf

[6] *Seldin Y., Shawe-Taylor J., Laviolette F.* PAC-Bayes analysis and its applications. — ECML 2012. Videolectures.net http://videolectures.net/ecmlpkdd2012_seldin_laviolette_shawe_taylor_pac/