Теория статистического обучения

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

2 марта 2016 г.

Материал находится в стадии разработки, может содержать ошибки и неточности. Автор будет благодарен за любые замечания и предложения, направленные по указанному адресу

1 Доказательство No free Lunch теоремы

Доказательство.

Выберем C — подмножество $\mathcal{X}$ мощности $2n$. Всего существует $T = 2^{2n}$ функций из C в $\{0, 1\}$. Обозначим эти функции как $f_1, \ldots, f_T$. Для каждой такой функции введем распределение P_i на $C \times \{0, 1\}$, приписывающая

$$\mathsf{P}_i(\{(X, Y)\}) = \frac{1}{|C|}, \text{если } y = f_i(x).$$

Очевидно, что риск f_i по отношению мере P_i нулевой. Обозначим обучающую выборку как S, тогда $\hat{f} = A(S)$. Докажем, что для любого обучающего алгоритма A, если $\hat{f}$ — получающийся в результате обучения классификатор, то выполняется

$$\max_{i \in \{1, \ldots, 2^{2n}\}} \mathbb{E}_{S \sim \mathsf{P}_i^n} L_{(X,Y) \sim \mathsf{P}_i}(\hat{f}) \geq \frac{1}{4}$$

и нулевую вероятность остальным парам (X, Y). Если удастся доказать это неравенство, то отсюда следует, что существует функция $f = f_i$ и соответствующая мера P_i такие, что $\mathbb{E}L(\hat{f}) \geq \frac{1}{4}$. Для заданных мер математические ожидания можно посчитать явно. Действительно, обозначим $S_j^1, \ldots, S_j^k$ возможные $k = (2n)^n$ реализации выборок с $y = f_j(x)$. Таким образом,

$$\begin{aligned}
\max_{i \in \{1, \ldots, 2^{2n}\}} \mathbb{E}_{S \sim \mathsf{P}_i^n} L_{(X,Y) \sim \mathsf{P}_i}(\hat{f}) &= \max_{i \in \{1, \ldots, 2^{2n}\}} \frac{1}{k} \sum_{j=1}^{k} L_{(X,Y) \sim \mathsf{P}_i}(A(S_j^i)) \\
&\geq \frac{1}{2^{2n}} \sum_{i=1}^{2^{2n}} \frac{1}{k} \sum_{j=1}^{k} L_{(X,Y) \sim \mathsf{P}_i}(A(S_j^i)) \\
&= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{2^{2n}} \sum_{i=1}^{2^{2n}} L_{(X,Y) \sim \mathsf{P}_i}(A(S_j^i)) \\
&\geq \min_{j \in \{1, \ldots, k\}} \frac{1}{2^{2n}} \sum_{i=1}^{2^{2n}} L_{(X,Y) \sim \mathsf{P}_i}(A(S_j^i)).
\end{aligned}$$

Фиксируем некторое $j \in \{1, \ldots, k\}$. Обозначаем $S_j = (x_1, \ldots, x_n)$ и пусть $v_1, \ldots, v_p$ — объекты из C, не попавшие в обучающую выборку. Очевидно, что $p \geq m$. Таким образом, для каждой функции $h : C \to \{0, 1\}$ имеет место

$$\begin{aligned} L_{(X,Y)\sim \mathsf{P}_i}(h) &= \frac{1}{2n} \sum_{x \in C} \mathbf{I}[h(x) \neq f_i(x)] \\ &\geq \frac{1}{2p} \sum_{r=1}^{p} \mathbf{I}[h(v_r) \neq f_i(v_r)]. \end{aligned}$$

Таким образом,

$$\begin{aligned} \frac{1}{2^{2n}} \sum_{i=1}^{2^{2n}} L_{(X,Y)\sim \mathsf{P}_i}(A(S_j^i)) &\geq \frac{1}{2^{2n}} \sum_{i=1}^{2^{2n}} \frac{1}{2p} \sum_{r=1}^{p} \mathbf{I}[A(S_j^i)(v_r) \neq f_i(v_r)] \\ &\geq \frac{1}{2} \min_{r \in \{1,\ldots,p\}} \frac{1}{2^{2n}} \sum_{i=1}^{2^{2n}} \mathbf{I}[A(S_j^i)(v_r) \neq f_i(v_r)] \end{aligned}$$

Теперь можно функции $f_1, \ldots, f_T$ разбить на $\frac{T}{2}$ непересекающихся пар $(f_i, f_{i'})$ таких, что в одной паре значения отличаются только на некотором фиксированном объекте v_r. Для таких пар $S_j^i = S_j^{i'}$. А значит на половине пар последний индикатор будет ненулевым для любого обучающего алгоритма A и:

$$\frac{1}{2} \min_{r \in \{1,\ldots,p\}} \frac{1}{2^{2n}} \sum_{i=1}^{2^{2n}} \mathbf{I}[A(S_j^i)(v_r) \neq f_i(v_r)] = \frac{1}{4}.$$

Таким образом доказано неравенство для математических ожиданий. ■

Упр. 1.1. Перейти от нижней оценки на математические ожидания к нижней оценке на вероятности уклонений.

Упр. 1.2. Сформулируйте No Free Lunch теорему в терминах минимаксных величин.

2 Неравенства концентрации меры

Одним из важнейших математических инструментов, которые мы в дальнейшем будем использовать, являются неравенства концентрации меры. Общий смысл таких неравенств заключается в явном выражении для вероятности отклонения случайных величин и функций от них от их медиан и математических ожиданий. Базовым неравенством является неравенство Маркова:

Лемма 2.1 (неравенство Маркова). *Пусть X – неотрицательная случайная величина с конечным математическим ожиданием. Тогда для любого $t > 0$*

$$\mathsf{P}\left\{X \geq t\right\} \leq \frac{\mathbb{E}X}{t}.$$

Упр. 2.1. Доказать неравенство Маркова.

Лемма 2.2 (неравенство Чебышева). *Пусть X — случайная величина с конечными математическим ожиданием и дисперсией. Тогда для любого $t > 0$*

$$\mathsf{P}\left\{|X - \mathbb{E}X| \geq t\right\} \leq \frac{\mathsf{D}(X)}{t^2}.$$

Доказательство.

$$\mathsf{P}\left\{|X - \mathbb{E}X| \geq t\right\} = \mathsf{P}\left\{(X - \mathbb{E}X)^2 \geq t^2\right\} \leq \frac{\mathsf{D}(X)}{t^2}.$$

■

Важным инструментом для получения неравенств концентрации являются верхние оценки *производящей функции моментов*. Зафиксируем некоторое $\lambda > 0$ и запишем неравенство Маркова:

$$\begin{aligned}\mathsf{P}\left\{|X - \mathbb{E}X| \geq t\right\} &= \mathsf{P}\left\{\lambda|X - \mathbb{E}X| \geq \lambda t\right\} \\ &= \mathsf{P}\left\{\exp(\lambda|X - \mathbb{E}X|) \geq \exp(\lambda t)\right\} \\ &\leq \frac{\mathbb{E}\exp(\lambda|X - \mathbb{E}X|)}{\exp(\lambda t)}.\end{aligned}$$

В числителе последней дроби возникает производящая функция моментов случайной величины X. *Метод Чернова* заключается в минимизации по λ последнего неравенства или его верхней оценки.

Лемма 2.3 (лемма Хеффдинга). *Пусть X – случайная величина, такая что почти наверное $X \in [a, b]$ и $\mathbb{E}X = 0$. Тогда для всех $\lambda > 0$*

$$\mathbb{E}\exp(\lambda X) \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

Упр. 2.2. Доказать лемму Хеффдинга.

Теорема 2.4 (неравенство Хеффдинга). *Пусть $Z_1, \ldots, Z_n$ - независимые случайные величины, такие что с вероятностью единица $Z_i \in [a_i, b_i]$. Обозначим $S_n = \sum_{i=1}^{n} Z_i$, тогда для любого $t > 0$ имеют место неравенства*

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

и

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

Упр. 2.3. Доказать неравенство Хеффдинга с помощью леммы Хеффдинга и метода Чернова.

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

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

Лемма 2.5 (лемма Хеффдинга). *Пусть V — случайная величина, а Z — случайный вектор, такие что $\mathbb{E}(V|Z) = 0$ и для некоторой неотрицательной функции h и константы $c > 0$ с вероятностью единица имеет место неравенство:*

$$h(Z) \le V \le h(Z) + c$$

Тогда для $\lambda > 0$:

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

Доказательство.

Повторяет доказательство леммы 2.3. ∎

Теорема 2.6. *Если $X_1,\dots,X_n$ - независимые случайные величины, а функция g обладает свойством ограниченных разностей, тогда для $t \ge 0$:*

$$\mathsf{P}\{g(X_1,\dots,X_n) - \mathbb{E}g(X_1,\dots,X_n) \ge t\} \le \exp\left(-\frac{2t^2}{\sum\limits_{i=1}^n c_i^2}\right),$$

$$\mathsf{P}\{\mathbb{E}g(X_1,\dots,X_n) - g(X_1,\dots,X_n) \ge t\} \le \exp\left(-\frac{2t^2}{\sum\limits_{i=1}^n c_i^2}\right).$$

Доказательство.

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

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

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

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

Если X_i распределена согласно F_i, то

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

Определим случайные величины

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

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 2.5 для всех $i \in \{1, \ldots, n\}$ с учетом независимости X_i:

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

И почти наверное

$$Z_i \leq V_i \leq W_i.$$

Далее используем метод Чернова и равенство $\mathbb{E}\{XY\} = \mathbb{E}\{Y \mathbb{E}\{X|Y\}\}$ для $\lambda > 0$:

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

Оптимизируя по λ, получаем условие теоремы. ■

Заметим, что неравенство ограниченных разностей является обобщением неравенства Хеффдинга для практически произвольных функций (с ограниченными разностями), зависящих от назависмых 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 $X_i \in [a_i, b_i]$.

3 Обучаемость конечных классов

Вернемся к задаче классификации с бинарной функцией потерь. 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 $\hat{f} = \arg\min_{f \in \mathcal{F}} L_n(f)$. Алгоритмы обучения, основанные на этой идее, будем называть методами *минимизации эмпирического риска*.

Обозначим некоторый классификатор с минимальным риском внутри класса $\mathcal{F}$ как $f^*_{\mathcal{F}}$, то есть $L(f^*_{\mathcal{F}}) = \inf\limits_{f\in\mathcal{F}} L(f)$. Введем *функционал избыточных потерь*:

$$\mathcal{L}(f, X, Y) = \ell(f, X, Y) - \ell(f^*_{\mathcal{F}}, X, Y).$$

Легко видеть, что любой минимизатор эмпирического риска принадлежит случайному множеству:

$$\left\{ f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(f, X_i, Y_i) < 0 \right\}.$$

Таким образом минимизатор эмпирического риска лежит в атипичном классе. С одной стороны, для него $\frac{1}{n}\sum\limits_{i=1}^{n} \mathcal{L}(f, X_i, Y_i) \le 0$ одновременно $\mathbb{E}\mathcal{L}(f, X, Y) > 0$. Рассмотрим теперь избыточный риск минимизатора эмпирического риска:

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

Оба слагаемых исследуются абсолютно одинаково. Исследуем, например, первое:

Лемма 3.1 (Конечный класс функций). *Пусть $\mathcal{F} = \{f_1, \ldots, f_N\}$ — бинарные классификаторы. Тогда для любого $\delta > 0$ одновременно с вероятностью не меньшей $1 - \delta$ выполнено*

$$\forall f \in \mathcal{F} : L(f) \le L_n(f) + \sqrt{\frac{\log(N) + \log\frac{1}{\delta}}{2n}}.$$

Доказательство.

С помощью неравенств Хеффдинга и Буля имеем

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

Отсюда:

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

Обращая оценку, получаем утверждение теоремы. ∎

Применяем аналогичное рассуждение и для $\sup\limits_{f\in\mathcal{F}}(L_n(f) - L(f))$, снова используем неравенство Буля и доказываем следующее утверждение:

Теорема 3.2. *Любой конечный класс является агностически PAC–обучаемым в задаче классификации с бинарной функцией потерь.*

Упр. 3.1. Оцените выборочную сложность в задачи классификации с конечным классом $\mathcal{F}$.

Список литературы

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

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

[3] *Shalev-Shwartz S., Ben-David S.* Understanding Machine Learning: From Theory to Algorithms // Cambridge University Press, 2014

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