Теория статистического обучения

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

11 апреля 2017 г.

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

1 Онлайн обучение в общем случае

Рассмотрим ситуацию, когда не предполагается то, что для некоторого $f^* \in \mathcal{F}$ выполнено $y_t = f(x_t)$. В данном разделе предполагаем для удобства, что $\mathcal{Y} = \{0, 1\}$. В этом случае по аналогии с задачами, когда дана i.i.d. выборка, мы рассмотрим избыточную ошибку, так называемый *регрет* (Regret):

$$\sup_{(x_1,y_1),\dots,(x_t,y_t)} \left(\sum_{t=1}^{T} |p_t - y_t| - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} |f(x_t) - y_t| \right)$$

Очевидно, что в худшем случае регрет будет линейным по T. Оказывается, что даже для самого простого класса $\mathcal{F} = \{f_0, f_1\}$, где f_0 тождественно равна 0 и f_1 тождественно равна 1 регрет нельзя сделать сублинейным. Действительно, в худшем случае $\sum\limits_{t=1}^{T} |p_t - y_t| = T$, при этом для данного класса $\inf\limits_{f\in\mathcal{F}} \sum\limits_{t=1}^{T} |f(x_t) - y_t| \le \frac{T}{2}$. Таким образом, без дополнительных ограничений невозможно гарантировать сублинейность регрета. Однако если предположить, что мы можем выбирать предсказания рандомизированно или принимает значения в отрезке $[0, 1]$, можно гарантировать сублинейность регрета.

Опр. 1.1. *Онлайн обучаемость определяется как существование такого алгоритма, для которого имеет место стремление к нулю регрета, деленного на T при $T \to \infty$.*

§1.1 Алгоритм взвешенного большинства

Сначала нам понадобится некоторый вспомогательный результат. Пусть даны N различных функций принимающих значения $\{0, 1\}$. На каждом шаге мы предоставляем вектор весов $w^{(t)} = (w_1^{(t)}, \dots, w_N^{(t)})$ — задающего вес каждого эксперта. После этого мы получаем вектор весов $v^{(t)} \in [0, 1]^N$. Потеря на шаге t определяется как $(w^{(t)}, v^{(t)})$.

Теорема 1.1. *Существует онлайн алгоритм, выдающий предсказания в отрезке* $[0,1]$*, для которого*

$$\sup_{(x_1,y_1),\ldots,(x_t,y_t)} \left(\sum_{t=1}^{T} |p_t - y_t| - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} |f(x_t) - y_t| \right) \leq \sqrt{2Ldim(\mathcal{F}) \log\left(\frac{eT}{Ldim(\mathcal{F})}\right) T}.$$

Доказательство.

TODO ∎

2 Неравенство Чернова и его применения

В дальнейшем нам понадобятся некоторые результаты, получаемые с помощью неравенства Чернова.

Лемма 2.1 (Неравенство Чернова). *Для* Z*, имеющей биномиальное распределение имеют место неравенства для* $\delta \in (0,1)$

$$P(Z \geq (1+\delta)\mathbb{E}Z) \leq \exp(-\delta^2 \mathbb{E}Z/3).$$

и

$$P(Z < (1-\delta)\mathbb{E}Z) \leq \exp(-\delta^2 \mathbb{E}Z/2).$$

Упр. 2.1. Используйте неравенство Бернштейна для доказательства неравенства Чернова (Возможно получатся другие константы).

Следующая лемма дает способ выбора классификаторов из конечного набора. Подобные результаты часто называются неточными (non-sharp) оракульными неравенствами, так как константа, стоящая перед слагаемым $\min_{i=1,\ldots,M} P(f_i(X) \neq f^*(X))$ больше 1.

Лемма 2.2 (Неточное оракульное неравенство для классификации). *Пусть есть* M *классификаторов* $f_1,\ldots,f_M$ *(не обязательно чтобы среди них был классификатор, не допускающий ошибки). Тогда существует* $c_0 > 0$*, такая что с вероятностью* $1-\delta$ *по обучающей выборке для минимизатора эмпирического риска на выборке длины* n *имеет место неравенство*

$$P(\hat{f}(X) \neq f^*(X)) \leq 2 \min_{i=1,\ldots,M} P(f_i(X) \neq f^*(X)) + c_0 \left(\frac{\log(M)}{n} + \frac{\log(\frac{1}{\delta})}{n} \right).$$

Лемма 2.3 (Версия 2). *Пусть есть* M *классификаторов* $f_1,\ldots,f_M$ *(не обязательно чтобы среди них был классификатор, не допускающий ошибки), среди которых есть классификатор с вероятностью ошибки не больше* $\frac{\varepsilon}{2}$*. Тогда с вероятностью* $1 - (M+1)\exp(-\frac{\varepsilon n}{32})$ *по обучающей выборке минимизатор эмпирического риска будет иметь риск не более* ε*.*

Доказательство.

Докажем второй вариант Леммы. Будем говорить, что классификатор прошел тест, если он прав хотя бы на $(1-\frac{3\varepsilon}{4})$ объектах. С помощью неравенства Чернова

получаем, что вероятность того, что классификатор с ошибкой не более $\frac{\varepsilon}{2}$ провалит тест, не больше чем $\exp(-\varepsilon n/24)$. Одновременно вероятность того, что классификатор с ошибкой большей чем ε пройдет тест также $\exp(-\varepsilon n/32)$. Вероятность того, что произойдет хотя бы одно из этих событий $(M+1)\exp(-\varepsilon n/32)$. ∎

Лемма 2.4 (От математического ожидания к экспоненциальным хвостам). *Если в бесшумном случае для классификатора $\hat{f}$ известна оценка $\mathbb{E}L(\hat{f}) \leq \frac{M}{n}$, то существует его модификация $\hat{g}$, для которой с вероятностью $1-\delta$ имеет место*

$$L(\hat{g}) = O\left(\frac{M\log(\frac{1}{\delta})}{n}\right).$$

Доказательство.

Зафиксируем $\varepsilon > 0$. Рассмотрим выборку размером $\frac{4d}{\varepsilon}$. Ожидаемая ошибка на такой выборке не превосходит $\frac{\varepsilon}{4}$, а по неравенству Маркова с вероятностью 0.5 не превосходит $\frac{\varepsilon}{2}$. Берем $\log_2(2/\delta)$ таких независимых выборок и для каждой обучаем наш классификатор $\hat{f}$. Из этого следует, что с вероятностью $1-\delta/2$ среди всех этих классификаторов есть один, имеющий вероятность ошибки меньше чем $\frac{\varepsilon}{2}$. Далее если протестировать все на последней выборке так чтобы для ее длины n было выполнено $(M+1)\exp(-\varepsilon n/32) \leq \delta/2$, то получится, что для такого алгоритма достаточно запросить $O(\frac{M\log(1/\delta)}{\varepsilon})$ точек. Это эквивалентно тому, что имеет место верхняя оценка на риск $O(\frac{M\log(1/\delta)}{n})$. ∎

Одним из других применений неравенства Чернова является свойства так называемой оценки медианы средних.

Лемма 2.5 (Медианы средних). *Пусть нам дана $N = nk$ элементная простая выборка $X_1, \ldots, X_{nk}$. Пусть $\mu = \mathbb{E}X_1$, $\sigma^2 = \mathsf{D}(X_1)$ и*

$$\mu_{MM} = median\left(\frac{1}{n}\sum_{i=1}^{n} X_i,\ \ldots,\ \frac{1}{n}\sum_{i=(k-1)n+1}^{kn} X_i\right).$$

Тогда с вероятностью $1-\delta$ для $k = 8\log(\frac{1}{\delta})$ выполнено

$$\|\mu_{MM} - \mu\| \leq 8\sigma\sqrt{\frac{\log(\frac{2}{\delta})}{N}}.$$

Эта оценка имеет практически такие же гарантии, как и оценка среднего в гауссовском случае. Однако, в нашем случае требуется лишь существование дисперсии.

Упр. 2.2. Докажите лемму в два шага:

1. С помощью неравенства Чебышева получаем, что среднее в каждом блоке отличается от μ не более чем на $\sigma\sqrt{\frac{8}{n}}$ с вероятностью $\frac{3}{4}$.
2. С помощью неравенства Чернова разберитесь в поведении медианы.

3 Из онлайн в i.i.d.

Теорема 3.1. *Любой консервативный (не изменяющий своего состояния при правильной классификации) онлайновый алгоритм, делающий не более M ошибок на любой конечной выборке, может быть превращен в алгоритм с вероятностью ошибки $O(\frac{M}{n} + \frac{\log(\frac{1}{\delta})}{n})$.*

Нам понадобятся некоторые дополнительные сведения. Последовательность случайных величин — *мартингал с дискретным временем*, если $\mathbb{E}[X_{n+1}|X_n, \dots, X_1] = X_n$ почти наверное. Если неравенство $\geq$, то имеем дело с *субмартингалом*, а иначе с *супермартингалом*.

Полезным свойством мартингальных последовательностей является то, что для них выполнены почти те же неравенства концентрации, что и в случае простых выборок.

Лемма 3.2 (Неравенство Азумы–Хеффдинга.). *Пусть $\{X_i\}_{i=1}^{\infty}$ – супермартингальная последовательность, такая что с вероятностью 1 выполнено $|X_i - X_{i-1}| < c_i$ для всех $i = 2, \dots$. Тогда для всех $t > 0$*

$$P(X_n - X_1 \geq t) \leq \exp\left(-t^2/2 \sum_{i=2}^{n} c_i^2\right).$$

Если в тех же условия последовательность субмартингальная, тогда

$$P(X_1 - X_n \geq t) \leq \exp\left(-t^2/2 \sum_{i=2}^{n} c_i^2\right).$$

Если последовательность мартингальная, тогда

$$P(|X_n - X_1| \geq t) \leq 2\exp\left(-t^2/2 \sum_{i=2}^{n} c_i^2\right).$$

Одним из ярких приложений является доказательство неравенства ограниченных разностей. Пусть $X_1, \dots, X_n$ — независимые случайные величины (в общем случае независимость не требуется). Для функции g обозначим $V_i = \mathbb{E}[g(X_1, \dots, X_n)|X_1, \dots, X_i]$. Очевидно, что $g(X_1, \dots, X_n) = V_n$, а $\mathbb{E}g(X_1, \dots, X_n) = V_0$. Последовательность $V_0, \dots, V_n$ называется *мартингалом Дуба*.

Упр. 3.1. Докажите мартингальные свойства для $V_0, \dots, V_n$, а затем получите неравенство ограниченных разностей.

Пусть $S_i = \sum_{j=1}^{i} X_j$, где $\{X_j\}_{j=1}^{\infty}$ – некоторые случайные величины. Тогда легко убедиться, что тот факт, что $\{S_i\}_{i=1}^{\infty}$ — супермартингал эквивалентен тому, что $\mathbb{E}(S_i - S_{i-1}|X_1, \dots, X_{i-1}) = \mathbb{E}(X_i|X_1, \dots, X_{i-1}) \leq 0$.

Лемма 3.3 (Неравенство Чернова для супермартингалов). *Пусть $X_1, \dots, X_n$ последовательность случайных величин, таких что $X_i \in [0, 1]$. Пусть $c_1, \dots, c_n$ некоторые константы из отрезка $[0, 1]$. Определим $\mu = \frac{1}{n}\sum_{i=1}^{n} c_i$ и последовательность $S_i = \sum_{j=1}^{i}(X_j - c_j)$. Тогда если последовательность S_i — супермартингал, то для любого*

$\alpha \in [0, 1-\mu]$ *выполнено*

$$P(S_n > \alpha n) \leq \exp\left(-\alpha^2 n/2(1-\mu)\right).$$

Лемма 3.4. *У консервативного алгоритма, допускающего M ошибок, обученного на выборке длины* $\max\left(\frac{16\log(2\delta)}{\varepsilon}, \frac{4(m-1)}{\varepsilon}\right)$ *хотя бы один классификатор имеет ошибку, вероятность которой меньше чем $\varepsilon/2$.*

Доказательство.

Рассмотрим случайные величины r_i, равные единице, если на i-ом объекте классификатор не ошибся и полученная на предыдущем шаге функция имеет ошибку с вероятностью большей чем $\varepsilon/2$:

$$r_i = \mathbf{I}[h_{i-1}(x_i) = f^*(x_i)] \wedge \mathbf{I}[P(h_{i-1}(x) \neq f^*(x)) \geq \varepsilon/2].$$

Рассчитаем $\mathbb{E}r_1$. Если h_0 имеет вероятность ошибки меньше чем $\varepsilon/2$, то $\mathbb{E}r_1 = 0$, иначе $\mathbb{E}r_1 \leq 1-\varepsilon/2$. Теперь найдем $\mathbb{E}[r_i|r_{i-1}, \ldots, r_1]$. Докажем, что последовательность $S_j = \sum_{i=1}^{j}(r_i - (1-\varepsilon/2))$ образует супермартингал с дискретным временем. Легко доказать, что $\mathbb{E}[r_i|r_{i-1}, \ldots, r_1] \leq 1-\varepsilon/2$ с вероятностью единица. Из сказанного выше это влечет субмартингальность последовательности S_j. Применяя к ней неравенство Чернова для мартингалов, получаем:

$$P(\sum_{i=1}^{n} r_i - n(1-\varepsilon/2) > \varepsilon n/4) \leq \exp\left(-\frac{\varepsilon^2 n}{2}/(1-\mu)\right) \leq \delta/2.$$

Таким образом, с вероятностью $1-\delta$ мы получаем, что $\sum_{i=1}^{n} r_i \leq n(1-\varepsilon/4)$. Теперь, если $\sum_{i=1}^{n} r_i \leq n - M$, то хотя бы один классификатор имеет вероятность ошибки $\leq \varepsilon/2$. Действительно, иначе получается, что произошло более чем M ошибок. Таким образом, с учетом нашего выбора параметра n получаем, что с вероятностью $1-\delta/2$ хотя бы один классификатор имеет малый риск. ■

После этой леммы для доказательства Теоремы 3.1 останется лишь воспользоваться неточным оракульным неравенством.

Список литературы

[1] *Littlestone N.* From On-line to Batch Learning. Colt 1989.

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