

Теория статистического обучения

Н. К. Животовский

nikita.zhivotovskiy@phystech.edu

31 марта 2015 г.

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

1 Быстрые порядки

На прошлой лекции мы выяснили, что для классов, обладающих конечной размерностью Вапника–Червоненкиса можно получить оценку избыточного риска вида

$$L(\hat{f}) - L(f_{\mathcal{F}}^*) \leq C \sqrt{\frac{V}{n}} + \sqrt{\frac{2 \log(\frac{1}{\delta})}{n}},$$

где неравенство выполняется с вероятностью не меньшей $1 - \delta$. Ставится вопрос, а не является ли порядок $O(\frac{1}{\sqrt{n}})$ слишком пессимистичным. Рассмотрим пример: Пусть некоторая функция f принимает значения 0 и 1 с вероятностями p и $1 - p$ соответственно. С помощью неравенства Хеффдинга имеем

$$P f - P_n f \leq \sqrt{\frac{2 \log(\frac{1}{\delta})}{n}}$$

Оказывается, что неравенство точно отражает скорость сходимости частот к вероятностям только если $P f = \frac{1}{2}$.

Утв. 1.1 (Неравенство Бернштейна). Пусть X_1, \dots, X_n независимые центрированные случайные величины, такие что $|X_i| \leq c$. Пусть $\sigma^2 = \frac{1}{n} \sum_{i=1}^n D(X_i)$. Тогда для любого $\varepsilon \geq 0$:

$$P \left(\frac{1}{n} \sum_{i=1}^n x_i \geq \varepsilon \right) \leq \exp \left(-\frac{n\varepsilon^2}{2\sigma^2 + 2c\varepsilon/3} \right)$$

Доказательство.

Пусть $F_i = \sum_{r=2}^{\infty} \frac{\lambda^{r-2} \mathbb{E}(X_i^r)}{r! \sigma_i^2}$, где $\sigma_i^2 = \mathbb{E}X_i^2$. Очевидно, что

$$\mathbb{E} \exp(\lambda X_i) = 1 + \sum_{r=2}^{\infty} \frac{\mathbb{E}X_i^r}{r!} = 1 + F_i \lambda^2 \sigma_i^2 \leq \exp(F_i \lambda^2 \sigma_i^2).$$

С помощью неравенства Коши–Буняковского имеем

$$\mathbb{E}X_i^r \leq \mathbb{E}X_i^{r-1}X_i \leq \sigma_i \left(\mathbb{E}|X_i^{r-1}|^2 \right)^{\frac{1}{2}}$$

Продолжая так m раз и устремляя m к бесконечности получим, что $\mathbb{E}X_i^r \leq \sigma_i^2 c^{r-2}$. Подставляя это выражение в формулу для F_i получаем, что

$$F_i \leq \frac{1}{\lambda^2 c^2} \sum_{r=2}^{\infty} \frac{\lambda^r c^r}{r!} = \frac{1}{\lambda^2 c^2} (\exp(\lambda c) - 1 - \lambda c).$$

В результате с помощью метода Чернова:

$$\mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n X_i \geq \varepsilon \right) \leq \exp(-\lambda \varepsilon) \exp \left(\lambda^2 n \sigma^2 \frac{\exp(\lambda c) - 1 - \lambda c}{\lambda^2 c^2} \right)$$

Фиксируем $\lambda = \frac{1}{c} \log \left(\frac{\varepsilon c}{n \sigma^2} + 1 \right)$ и получаем, что

$$\mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n X_i \geq \varepsilon \right) \leq \exp \left(\frac{n \sigma^2}{c^2} \left(\frac{\varepsilon c}{n \sigma^2} - \log \left(\frac{\varepsilon c}{n \sigma^2} + 1 \right) \right) - \frac{\varepsilon}{c} \log \left(\frac{\varepsilon c}{n \sigma^2} + 1 \right) \right).$$

Пусть $H(x) = (1+x) \log(1+x) - x$. Тогда

$$\mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n X_i \geq \varepsilon \right) \leq \exp \left(-\frac{n \sigma^2}{c^2} H \left(\frac{\varepsilon c}{n \sigma^2} \right) \right).$$

Чтобы получить неравенство Бернштейна осталось вместо $H(x)$ подставить ее нижнюю оценку $G(x) = \frac{3}{2} \frac{x^2}{x+3}$. ■

Обращая неравенство Бернштейна для предыдущего примера и учитывая $\mathbb{P} f \leq \mathbb{P} f(1 - \mathbb{P} f)$ получаем, что

$$\mathbb{P} f - \mathbb{P}_n f \leq \sqrt{\frac{2 \mathbb{P} f \log(\frac{1}{\delta})}{n}} + \frac{2 \log(\frac{1}{\delta})}{3n}.$$

В областях с малым математическим ожиданием скорость сходимости частот к вероятностям значительно лучше.

Рассмотрим теперь задачу бинарной классификации с классами $\{-1, 1\}$ и индикаторной функцией потерь. Тогда байесовское решающее правило имеет вид

$$f^*(x) = \text{sign}(\eta(x)).$$

где $\eta(x) = \mathbb{E}[Y|X = x]$.

Упр. 1.1. Докажите оптимальность Байесовского решающего правила.

Опр. 1.1 (Условия малого шума Массара). Пусть $f^* \in \mathcal{F}$. Говорят, что для распределения \mathbb{P} на (X, Y) и семейства \mathcal{F} выполнено условие малого шума, если с вероятностью единица

$$|\eta(x)| \geq \frac{1}{c}$$

для некоторой константы c .

Упр. 1.2. Докажите, что для всех $f \in \mathcal{F}$ имеет место соотношение:

$$L(f) - L(f^*) = \mathbb{E}(|\eta(X)|\mathbf{I}[f(X)f^*(X) < 0]).$$

Утв. 1.2. Условие малого шума эквивалентно тому, что для любой $g \in (\ell \circ \mathcal{F})^*$:

$$\mathbb{P} g^2 \leq c \mathbb{P} g.$$

Доказательство.

Пусть $g \in (\ell \circ \mathcal{F})^*$ соответствует $f \in \mathcal{F}$. Тогда

$$\begin{aligned} \mathbb{P} g &= L(f) - L(f^*) = \\ &= \mathbb{E}(|\eta(X)|\mathbf{I}[f(X)f^*(X) < 0]) \geq \\ &= \frac{1}{c} \mathbb{E}(\mathbf{I}[f(X)f^*(X) < 0]). \end{aligned}$$

С другой стороны

$$\begin{aligned} \mathbb{P} g^2 &= \\ &= \mathbb{E}(\mathbf{I}[f(X) \neq Y] - \mathbf{I}[f^*(X) \neq Y])^2 = \\ &= \mathbb{E}(\mathbf{I}[f(X) \neq f^*(X)]) = \\ &= \mathbb{E}(\mathbf{I}[f(X)f^*(X) < 0]). \end{aligned}$$

■

Утв. 1.3. Для неотрицательных чисел A, B, C имеет место, что если $A \leq B + C\sqrt{A}$, то $A \leq B + C^2 + \sqrt{BC}$

Теорема 1.4. В задаче классификации с бинарной функцией потерь в случае, если $|\mathcal{F}| = N$ и выполнено условие малого шума Массара с константой c , то для минимизатора эмпирического риска \hat{f} :

$$L(\hat{f}) - L(f^*) \leq \frac{2(1 + 3\sqrt{c} + c) \log(\frac{N}{\delta})}{3n}$$

Доказательство.

Пусть $h \in (\ell \circ \mathcal{F})^*$. Пусть Dh — дисперсия h . Тогда с помощью неравенства Бернштейна с вероятностью не меньшей $1 - \delta$:

$$\mathbb{P} h \leq \mathbb{P}_n h + \sqrt{\frac{2Dh \log(\frac{1}{\delta})}{n}} + \frac{2 \log(\frac{1}{\delta})}{3n}.$$

Объединяя с помощью неравенства Буля предыдущее неравенство по всему классу, имеем с вероятностью не меньшей $1 - \delta$ одновременно для всех $g \in (\ell \circ \mathcal{F})^*$:

$$\mathbb{P} g \leq \mathbb{P}_n g + \sqrt{\frac{2Dg \log(\frac{N}{\delta})}{n}} + \frac{2 \log(\frac{N}{\delta})}{3n}.$$

Из условия малого шума следует, что $Dg \leq P g^2 \leq c P g$. Тогда в тех же условиях

$$P g \leq P_n g + \sqrt{\frac{2c P g \log(\frac{N}{\delta})}{n}} + \frac{2 \log(\frac{N}{\delta})}{3n}.$$

Пусть $\hat{g} = \mathbf{I}[\hat{f}(X) \neq Y] - \mathbf{I}[f^*(X) \neq Y]$. Ясно, что $P \hat{g} = L(\hat{f}) - L(f^*)$ и $P_n \hat{g} \leq 0$. Теперь, используя что из $A \leq B + C\sqrt{A}$ следует $A \leq B + C^2 + \sqrt{BC}$, получаем, что с вероятностью не меньшей $1 - \delta$:

$$L(\hat{f}) - L(f^*) \leq \frac{2 \log(\frac{N}{\delta})}{3n} + \frac{2c \log(\frac{N}{\delta})}{n} + \sqrt{\frac{2 \log(\frac{N}{\delta})}{3n}} \sqrt{\frac{2c \log(\frac{N}{\delta})}{n}} \leq \frac{2(1 + 3\sqrt{c} + c) \log(\frac{N}{\delta})}{3n}$$

■

Таким образом в условиях малого шума избыточный риск для конечных \mathcal{F} имеет порядок $\frac{1}{n}$.

Список литературы

- [1] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A Survey of Some Recent Advances // ESAIM: Probability and Statistics, 2005.
- [2] *Koltchinskii V.* Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems // Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics. Springer-Verlag, 2011.
- [3] *Rakhlin A.* Statistical Learning Theory and Sequential Prediction // Lecture notes, 2014, <http://www-stat.wharton.upenn.edu/~rakhlin/>
- [4] *Shalev-Shwartz S., Ben-David S.* Understanding Machine Learning: From Theory to Algorithms // Cambridge University Press, 2014
- [5] *Vapnik V.* Statistical Learning Theory. — John Wiley and Sons, New York, 1998.