Теория статистического обучения

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

27 февраля 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 Принцип равномерной сходимости

В прошлой лекции мы доказали агностическую PAC-обучаемость конечных классов для задачи бинарной классификации. В качестве алгоритма мы привели метод минимизации эмпирического риска. Введем понятие *класса потерь*:

$$\ell \circ \mathcal{F} = \{(x, y) \to \ell(f, x, y) : f \in \mathcal{F}\}.$$

Введем одновременно понятие *класса избыточных потерь*:

$$(\ell \circ \mathcal{F})^* = \{(x, y) \to \ell(f, x, y) - \ell(f^*, x, y) : f \in \mathcal{F}\}.$$

Скоро мы выясним, что свойства обучаемости существенно зависят от геометрических свойств класса потерь (класса избыточных потерь). Рассмотрим некоторую функцию $g : X \to \mathbb{R}_+$. Введем два стандартных обозначения: $\mathsf{P}\, g = \mathbb{E} g$ и $\mathsf{P}_n\, g = \frac{1}{n}\sum\limits_{i=1}^{n} g(x_i)$, где математическое ожидание берется по распределению на X, которое также обозначается P, а суммирование ведется по реализации независимых x_i, распределенных согласно распределению P.

Говорят, что для класса функций $\mathcal{G}$ выполняется *независищий от распределения усиленный закон больших чисел* (*distribution free uniform strong law of large numbers*), если для любого $\varepsilon > 0$

$$\lim_{n\to\infty} \sup_{\mathsf{P}} \mathsf{P}\left\{ \sup_{m \geq n} \sup_{\mathcal{G}} |\, \mathsf{P}_n\, g - \mathsf{P}\, g| > \varepsilon \right\} = 0;$$

Классы, для которых выполнено данное свойство называются *равномерными классами Гливенко–Кантелли.*

Утв. 1.1 (Принцип равномерной сходимости). *Для того чтобы класс $\mathcal{F}$ был агностически PAC-обучаемым достаточно, чтобы соответсвующий класс потерь $\ell \circ \mathcal{F}$ был равномерным классом Гливенко–Кантелли.*

Доказательство.

Докажем, что обучаемость достигается с помощью метода минимизации эмпирического риска. Действительно, для минимизатора эмпирического риска $\hat{f}$:

$$\begin{aligned}
&L(\hat{f}) - L(f^*_{\mathcal{F}}) = \\
&L(\hat{f}) - L_n(\hat{f}) + L_n(f^*_{\mathcal{F}}) - L(f^*_{\mathcal{F}}) + L_n(\hat{f}) - L_n(f^*_{\mathcal{F}}) \le \\
&L(\hat{f}) - L_n(\hat{f}) + L_n(f^*_{\mathcal{F}}) - L(f^*_{\mathcal{F}}) \le \\
&2 \sup_{f \in \mathcal{F}} |L(f) - L_n(f)| = \\
&2 \sup_{g \in \ell \circ \mathcal{F}} |\,\mathsf{P}\, g - \mathsf{P}_n\, g|.
\end{aligned}$$

∎

Упр. 1.1. Воспользуйтесь определением равномерного класса Гливенко–Кантелли для доказательства агностической PAC-обучаемости.

Условие теоремы вовсе не необходимое для обучаемости. Фактически оно является лишь достаточным для того, чтобы агностически обучиться с помощью метода минимизации эмпирического риска. Далее мы покажем, что для задач классификации с бинарной функцией потерь варианты равномерных законов больших чисел являются также и необходимыми для обучаемости с помощью метода минимизации эмпирического риска. Тем не менее нужно согласовывать это со следующим элементарным результатом.

Пример 1.1 (обучение без равномерной сходимости). Возьмем произвольный класс $\mathcal{F}$ и добавим к нему такой функционал $f^{(1)}$, что $\ell(f^{(1)}, x, y) < \inf\limits_{f \in \mathcal{F}} \ell(f, x, y)$ с вероятностью единица. Тогда очевидно, что минимизатор эмпирического риска с вероятностью единица будет выбирать $f^{(1)}$, который доставляет и минимальный риск в новом классе. Заметим, что для класса $\mathcal{F}$ в этом случае может не выполняться никаких вариаций равномерного закона больших чисел.

Упр. 1.2. Возможна ли ситуация из предыдущего примера в задаче классификации с бинарной функцией ошибок?

2 Радемахеровский процесс

Первым шагом на пути выяснения, а какие же классы потерь являются равномерными является так называемая *симметризация*. Рассмотрим на первом шаге математическое ожидание

$$\mathbb{E} \sup_{f \in \mathcal{F}} |L(f) - L_n(f)|.$$

Обозначим $L'_n(f)$ – эмипирическое среднее по независимой копии обучающей выборки. Соответсвующее ей математическое ожидание будем обозначать $\mathbb{E}'$. С помощью неравенства Йенсена имеем

$$\mathbb{E}\sup_{f\in\mathcal{F}}|L(f)-L_n(f)| =$$
$$\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{E}'L_n'(f)-L_n(f)| \le$$
$$\mathbb{E}\mathbb{E}'\sup_{f\in\mathcal{F}}|L_n'(f)-L_n(f)| =$$
$$\frac{1}{n}\ \mathbb{E}\mathbb{E}'\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}(\ell(f,X_i',Y_i')-\ell(f,X_i,Y_i))\right|.$$

Введем *Радемахеровские случайные величины*, то есть независимые в совокупности (и от X_i, Y_i) случайные величины σ_i, принимающие равновероятно значения 1 и -1. Легко видеть, что для всех i распределения случайных величин $(\ell(f,X_i',Y_i')-$ $-\ \ell(f,X_i,Y_i))$ и $\sigma_i(\ell(f,X_i',Y_i')-\ell(f,X_i,Y_i))$ одинаковы. Поэтому, обозначая математическое ожидание по всем σ_i как $\mathbb{E}_\sigma$, получаем:

$$\frac{1}{n}\ \mathbb{E}\mathbb{E}'\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}(\ell(f,X_i',Y_i')-\ell(f,X_i,Y_i))\right| =$$
$$\frac{1}{n}\ \mathbb{E}\mathbb{E}'\mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i(\ell(f,X_i',Y_i')-\ell(f,X_i,Y_i))\right| \le$$
$$\frac{2}{n}\ \mathbb{E}\mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i\ell(f,X_i,Y_i)\right|$$

Введем для фиксированной выборки $(X_i,Y_i)_{i=1}^n$ *условную Радемахеровскую сложность*:

$$\mathcal{R}_n(\ell\circ\mathcal{F}) = \frac{1}{n}\ \mathbb{E}_\sigma\sup_{g\in\ell\circ\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i g(X_i,Y_i)\right| = \frac{1}{n}\ \mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i\ell(f,X_i,Y_i)\right|$$

и просто *Радемахеровскую сложность*

$$\mathcal{R}(\ell\circ\mathcal{F}) = \mathbb{E}\mathcal{R}_n(\ell\circ\mathcal{F}) = \frac{1}{n}\ \mathbb{E}\sup_{g\in\ell\circ\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i g(X_i,Y_i)\right| = \frac{1}{n}\ \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i\ell(f,X_i,Y_i)\right|.$$

Таким образом, мы получили, что

$$\mathbb{E}\sup_{f\in\mathcal{F}}|L(f)-L_n(f)| \le 2\mathcal{R}(\ell\circ\mathcal{F}).$$

Радемахеровскую сложность можно рассматривать как величину, описывающую сложность класса решающих правил. Чем больше Радемахеровская сложность, тем лучше ошибки $\mathcal{F}$ могут кореллировать со случайным шумом σ_i. Как только мы зафиксировали выборку $(X_i,Y_i)_{i=1}^n$ условную Радемахеровскую сложность можно рассматривать как *Радемахеровское среднее*, связанное со множеством $A\subset\mathbf{R}^n$:

$$\mathcal{R}_n(A) = \frac{1}{n}\ \mathbb{E}_\sigma\sup_{a\in A}\left|\sum_{i=1}^{n}\sigma_i a_i\right|,$$

где множество A является множеством векторов ошибок $\mathcal{F}$ на $(X_i, Y_i)_{i=1}^n$.

Рассмотрим простые свойства Радемахеровских средних. Если A, B – ораниченные множества в $\mathbb{R}^n$, $c \in \mathbb{R}$.

1. $\mathcal{R}_n(A \cup B) \leq \mathcal{R}_n(A) + \mathcal{R}_n(B)$.

2. $\mathcal{R}_n(cA) = |c|\mathcal{R}_n(A)$.

3. $\mathcal{R}_n(A \oplus B) \leq \mathcal{R}_n(A) + \mathcal{R}_n(B)$.

4. Если $A = \{a^{(1)}, \dots, a^{(N)}\}$, то $\mathcal{R}_n(A) \leq \max\limits_j ||a^{(j)}||_2 \sqrt{\frac{2\log(2N)}{n}}$.

5. (Contraction inequality [2]) Если $\varphi : \mathbb{R} \to \mathbb{R}$ Липшицева с константой L, причем $\varphi(0) = 0$, то $\mathcal{R}_n(\varphi(A)) \leq L\mathcal{R}_n(A)$, где φ действует на векторы A покомпонентно.

6. $\mathcal{R}_n(A) = \mathcal{R}_n(conv(A))$.

Доказательства первых трёх пунктов являются простыми упражнениями. Разберёмся подробно с 4-ым пунктом.

Случайная величина Y называется сабгауссовской с параметром σ^2 ($Y \in SG(\sigma^2)$), если для $\lambda > 0$:

$$\mathbb{E}\exp(\lambda Y) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

Можно легко показать, что, если $(Y_i)_{i=1}^n$ независимые случайные величины, такие, что $Y_i \in SG(\sigma_i^2)$, то $\sum\limits_{i=1}^n Y_i \in SG\left(\sum\limits_{i=1}^n \sigma_i^2\right)$.

Более того, все так определенные случайные величины имеют нулевое математическое ожидание.

Лемма 2.1. *Пусть $Y_i \in SG(\sigma_i^2)$, $i = 1, \dots, N$. Тогда*

$$\mathbb{E}\max_i |Y_i| \leq \max_i \sigma_i \sqrt{2\ln(2n)}.$$

Доказательство.

$$\exp\left(\lambda\mathbb{E}\left\{\max_{i=1,\dots,N} Y_i\right\}\right) \leq$$
$$\mathbb{E}\left\{\exp(\lambda \max_{i=1,\dots,N} Y_i)\right\} =$$
$$\mathbb{E}\left\{\max_{i=1,\dots,N} \exp(\lambda Y_i)\right\} \leq$$
$$N\exp\left(\frac{\lambda^2 \max\limits_{i=1,\dots,N} \sigma_i^2}{2}\right).$$

Логарифимруя обе части и оптимизируя по λ, получаем, что

$$\mathbb{E}\max_i Y_i \leq \max \sigma_i \sqrt{2\ln(n)}.$$

Для получения утверждения леммы нужно применить полученное оценку к набору $Y_1, \dots, Y_n, -Y_1, \dots, -Y_n$. ■

Упр. 2.1. С помощью доказанной леммы завершите доказательство 4-го свойства Радемахеровских средних.

Рассмотрим $\sup\limits_{f \in \mathcal{F}} |L(f) - L_n(f)|$ как функцию от наблюдаемой выборки $(X_i, Y_i)_{i=1}^n$. Непосредственно убеждаемся, что, если функция потерь равномерно ограничена единицей, то введенная функция является функцией с ограниченными приращениями с $c_i = \frac{2}{n}$. Аналогично, рассматривая условную Радемахеровску сложность как функцию от наблюдаемой выборки, доказываем, что она также является функцией с ограниченными разностями с $c_i = \frac{2}{n}$.

Теорема 2.2. *С вероятностью не меньшей* $1-\delta$ *для функций потерь, принимающих значения в отрезке* $[0, 1]$*:*

$$\sup_{f \in \mathcal{F}} |L(f) - L_n(f)| \le 2\mathcal{R}(\ell \circ \mathcal{F}) + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}}.$$

Также

$$\sup_{f \in \mathcal{F}} |L(f) - L_n(f)| \le 2\mathcal{R}_n(\ell \circ \mathcal{F}) + 3\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}$$

Доказательство.

Доказательство первого неравенства заключается в применении неравенства ограниченных разностей для $\sup\limits_{f \in \mathcal{F}} |L(f) - L_n(f)|$ и ограничении сверху $\mathbb{E} \sup\limits_{f \in \mathcal{F}} |L(f) - L_n(f)|$ с помощью $2\mathcal{R}(\ell \circ \mathcal{F})$.

Доказательство второго неравенства заключается в использовании неравенства ограниченных разностей для $\mathcal{R}_n(\ell \circ \mathcal{F})$ и учёте того, что $\mathbb{E}\mathcal{R}_n(\ell \circ \mathcal{F}) = \mathcal{R}(\ell \circ \mathcal{F})$. ■

Список литературы

[1] *Alon N., Ben-David S., Cesa-Bianci N., Haussler D.* Scale-Sensitive Dimensions, Uniform Convergence, and Learnability // 1997,

[2] *Koltchinskii V.* Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems // Ecole d'Etґe de Probabilitґes 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., Shamir O., Srebro N., Sridharan K.* Learnability, Stability and Uniform Convergence // Journal of Machine Learning Research 11, 2010

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

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