Теория статистического обучения

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

22 мая 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 Фундаментальная теорема PAC-обучения

Теорема 1.1. *В задаче классификации с бинарной функцией потерь следующие утверждения эквивалентны*

1. *Класс потерь $\ell \circ \mathcal{F}$ является равномерным классом Гливенко–Кантелли.*
2. *$\mathcal{F}$ является агностически PAC обучаемым.*
3. *$\mathcal{F}$ является PAC обучаемым.*
4. *Класс $\mathcal{F}$ имеет конечную размерность Вапника–Червоненкиса.*

При этом обучаемость производится с помощью минимизации эмпирического риска.

Доказательство.

Переход $1 \to 2$ следует из того, что для минимизатора эмпирического риска $L(\hat{f}) - L(f^*_{\mathcal{F}}) \leq 2 \sup\limits_{f \in F} |L_n(f) - L(f)|$. Переход $2 \to 3$ очевиден. Переход $3 \to 4$ выполняется с помощью No free lunch теоремы. Переход $4 \to 1$ осуществляется за счет того, что в семействе с конечной размерностью V величина $\sup\limits_{f \in F} |L_n(f) - L(f)|$ с большой вероятностью имеет порядок $O\left(\sqrt{\frac{V}{n}}\right)$. ∎

Оказывается порядок $\frac{1}{\sqrt{n}}$ является в общем случае неулучшаемым. Мы покажем, что он не улучшаем для минимизации эмпирического риска даже в случае, когда $|\mathcal{F}| = 2$. Пусть в задаче бинарной классификации $Y \equiv 0$, а $\mathsf{P}(X = 1) = \frac{1}{2} - \frac{1}{\sqrt{n}}$ и $\mathsf{P}(X = -1) = \frac{1}{2} + \frac{1}{\sqrt{n}}$. Рассмотрим $\mathcal{F} = \{f_1, f_2\}$, где $f_1 = \mathbf{I}[0, 1]$, а $f_2 = \mathbf{I}[-1, 0]$. Имеют место равенства $L(f_1) = \frac{1}{2} - \frac{1}{\sqrt{n}}$, $L(f_2) = \frac{1}{2} + \frac{1}{\sqrt{n}}$. Как следствие $f^*_{\mathcal{F}} = f_1$. Осталось показать, что с константной вероятностью минимизатор эмпирического риска выбирает f_2. Тогда избыточный риск будет иметь величину $\frac{2}{\sqrt{n}}$ с константной вероятностью.

Лемма 1.2 (Неравенство Берри–Эссеена). *Пусть дана последовательность независимых одинаково распределенных случайных величин X_n, таких что $\mathbb{E}(X) = 0$, $\mathbb{E}|X|^3 < \infty$. Тогда для функции распределения $F^{(n)}$ суммы $\frac{\sum_{i=1}^{n} X_i}{\sigma\sqrt{n}}$ для всех t:*

$$|F^{(n)}(t) - \Phi(t)| \leq \frac{c_0 \mathbb{E}|X|^3}{\sigma^3\sqrt{n}},$$

где Φ — функция распределения стандартного нормального распределения, а c_0 — константа Берри–Эссеена.

Тогда, если g имеет стандартное нормальное распределение, то

$$\left|\mathsf{P}\left(L_n(f_2) - L_n(f_1) \leq L(f_2) - L(f_1) + \frac{\sigma t}{\sqrt{n}}\right) - \mathsf{P}(g \leq t)\right| \leq \frac{c_0 \mathbb{E}|X|^3}{\sqrt{n}},$$

где $\sigma^2 = \mathbb{E}(X - \mathbb{E}X)^2 = 1 - \frac{1}{4n}$. Заметим, что $L(f_2) - L(f_1) = \frac{2}{\sqrt{n}}$. Для $t = -3$ и достаточно большого n имеет место $\mathsf{P}(g \leq t) - \frac{c_0}{\sqrt{n}} > 0$. И с этой вероятностью

$$L_n(f_2) - L_n(f_1) < 0,$$

то есть минимизатор эмпирического риска выберет f_2.

2 Структурная минимизация эмпирического риска

В предыдущих лекциях мы считали, что семейство $\mathcal{F}$ задано априорно. Изучая агностическую PAC-обучаемость, мы не интересовались ошибкой аппроксимации, то есть величиной $L(f^*_{\mathcal{F}}) - L(f^*)$. Залогом успеха в практических задачах является именно выбор оптимального семейства $\mathcal{F}$. Одним из наиболее хорошо изученных методов является метод *структурной минимизации эмпирического риска*. Предположим, что теперь у нас имеется возможно бесконечный набор семейств $\{\mathcal{F}_i\}_{i=1}^{m}$, $m \leq \infty$. Например, это могут быть семейства с возрастающей размерностью Вапника–Червоненкиса или классы полиномов с увеличивающимися максимальными степенями.

Теперь учет ошибки аппрокисмации будет заключаться в следующем: пусть $\hat{f}_k$ — некоторая оценка (например минимизатор эмпирического риска), построенная с помощью семейства $\mathcal{F}_k$. Но не известно, какая из *моделей* $\mathcal{F}_i$ дает наименьшую ошибку аппроксимации. Поэтому первостепенной задачей было бы найти модель $\tilde{k}$, такую что

$$\mathbb{E}(L(\hat{f}_{\tilde{k}}) - L(f^*)) = \inf_k \mathbb{E}(L(\hat{f}_k) - L(f^*)).$$

На данном шаге единственным требованием от оценки будет стремление внутри модели к нулю избыточного риска. Если бы $\tilde{k}$ было известно, то задача фактически решена, так известна конкретная модель, в которой ошибка оценки минимальна.

Очевидно, что доступа к *оракульному* значению $\tilde{k}$ у нас нет, поэтому мы должны построить некоторую его оценку $\hat{k}$. Финальной целью будет получение так называемого *оракульного неравенства*:

$$\mathbb{E}(L(\hat{f}_{\hat{k}}) - L(f^*)) \leq C \inf_k \left(\mathbb{E}(L(\hat{f}_k) - L(f^*)) + C' \frac{\gamma(k, n)}{n}\right),$$

где $C \geq 1$, $C' \geq 0$, а γ — некоторая медленно растущая функция. В случае, если $C = 1$ оракульное неравенство называется точным (sharp). Обозначим $L(f^*_{\mathcal{F}_k})$ как $L(f^*_k)$. Тогда оракульное неравенство удобно представить в виде

$$\mathbb{E}(L(\hat{f}_{\hat{k}}) - L(f^*)) \leq C \inf_k \left(L(f^*_k) - L(f^*) + \mathbb{E}(L(\hat{f}_k) - L(f^*_k)) + C' \frac{\gamma(k,n)}{n} \right).$$

Если окажется, что член $C' \frac{\gamma(k,n)}{n}$ мал в сравнении с ошибкой оценивания внутри модели $\mathbb{E}(L(\hat{f}_k) - L(f^*_k))$, а константа C близка к единице, то наша оценка вместе с методом выбора модели хорошо приближают лучшую аппроксимацию $L(f^*)$ с помощью оракульной модели.

Возникает вопрос: каким образом строить оценку $\hat{k}$. Задача сводится к умению попарно сравнивать пару произвольных моделей $\mathcal{F}_i$ и $\mathcal{F}_j$. Если алгоритм обучения внутри модели выбран, то разумно из двух моделей выбирать ту, которая дает меньший риск оценки, а именно, если $\mathbb{E}L(\hat{f}_i) \leq \mathbb{E}L(\hat{f}_j)$, то $\mathcal{F}_i$ предпочтительнее $\mathcal{F}_j$. Но как и ранее нам неизвестно распределение, поэтому истинный риск приходится заменять на имперический и производить сравнения $L_n(\hat{f}_i)$ и $L_n(\hat{f}_j)$.

Наиболее популярные методы основаны на введении некоторого порога τ такого, что $\mathcal{F}_i$ предпочтительнее $\mathcal{F}_j$, если

$$L_n(\hat{f}_i) - L_n(\hat{f}_j) < \tau(i,j,n).$$

Например, введя порог $\tau(i,j,n) = pen(j,n) - pen(i,n)$, получим, что наиболее предпочтительное семейство связано со следующим критерием:

$$\hat{k} = \arg\min_k (L_n(\hat{f}_k) + pen(k,n)).$$

Такие методы называют *методами выбора модели с пенализацией*. Следующая лемма усиливает свойства концентрации для условной радемахеровской сложности.

Лемма 2.1. *Пусть класс функций $\mathcal{G}$ принимает значения в отрезке $[-1,1]$. Тогда*

$$\mathsf{D}\mathcal{R}_n(\mathcal{G}) \leq \frac{1}{n}\mathcal{R}_n(\mathcal{G}),$$
$$\mathsf{P}\left(\mathcal{R}_n(\mathcal{G}) \geq \mathbb{E}\mathcal{R}_n(\mathcal{G}) + \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{\mathbb{E}} 2(\mathbb{E}\mathcal{R}_n(\mathcal{G}) + \varepsilon/3) \right),$$
$$\mathsf{P}\left(\mathcal{R}_n(\mathcal{G}) \leq \mathbb{E}\mathcal{R}_n(\mathcal{G}) - \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\mathbb{E}\mathcal{R}_n(\mathcal{G})} \right).$$

Теорема 2.2. *Введем для задачи классификации с бинарной функцией потерь следующую пенализацию:*

$$pen(n,k) = 3\mathcal{R}_n(\ell \circ \mathcal{F}_k) + 2\sqrt{\frac{\log(k)}{n}} + \frac{18\log(k)}{n}.$$

Пусть $\hat{k} = \arg\min_k (L_n(\hat{f}_k) + pen(k,n))$, а $\hat{f}_k$ минимизирует эмпирический риск в k-ой модели. Тогда выполнено следующее оракульное неравенство:

$$\mathbb{E}\left(L(\hat{f}_{\hat{k}}) - L(f^*) \right) \leq \inf_k \left(L(f^*_k) - L(f^*) + 5\mathbb{E}\mathcal{R}_n(\ell \circ \mathcal{F}_k) + 2\sqrt{\frac{\log(k)}{n}} + \frac{18\log(k)}{n} \right) + \sqrt{\frac{2\pi}{n}} + \frac{18}{n}.$$

Доказательство.

Из способа выбора $\hat{k}$ имеем для всех k:

$$L(\hat{f}_{\hat{k}}) - L(f^*)$$
$$\leq L(\hat{f}_k) - L(f^*) - \Big(L(\hat{f}_k) - L_n(\hat{f}_k) - pen(n,k))\Big) + \Big(L(\hat{f}_{\hat{k}}) - L_n(\hat{f}_{\hat{k}}) - pen(n,\hat{k}))\Big)$$

Взяв математические ожидания получаем:

$$\mathbb{E}\left(L(\hat{f}_{\hat{k}}) - L(f^*)\right)$$
$$\leq \mathbb{E}\left(L(\hat{f}_k) - L(f^*)\right) - \mathbb{E}\left(L(\hat{f}_k) - L_n(\hat{f}_k) - pen(n,k))\right) + \mathbb{E}\left(L(\hat{f}_{\hat{k}}) - L_n(\hat{f}_{\hat{k}}) - pen(n,\hat{k}))\right)$$
$$\leq \mathbb{E}\left(L(\hat{f}_k) - L(f^*) + pen(n,k)\right) + \mathbb{E}\left(\sup_k \left(L(\hat{f}_k) - L_n(\hat{f}_k) - pen(n,k)\right)\right)$$
$$\leq \mathbb{E}\left(L(\hat{f}_k) - L(f^*) + pen(n,k)\right) + \mathbb{E}\left(\sup_k \left(\sup_{f\in\mathcal{F}_k}(L(f) - L_n(f)) - pen(n,k)\right)\right)$$
$$\leq \mathbb{E}\left(L(\hat{f}_k) - L(f^*) + pen(n,k)\right) + \sum_k \mathbb{E}\left(\sup_{f\in\mathcal{F}_k}(L(f) - L_n(f)) - pen(n,k)\right)_+.$$

Осталось разобраться с последней суммой. Для любого $\delta > 0$, используя неравенство $\mathbb{E}\sup_{f\in\mathcal{F}_k}(L(f) - L_n(f)) \leq 2\mathbb{E}\mathcal{R}_n(\ell\circ\mathcal{F}_k)$:

$$\mathsf{P}\left(\sup_{f\in\mathcal{F}_k}(L(f) - L_n(f)) \geq pen(n,k) + 2\delta\right)$$
$$= \mathsf{P}\left(\sup_{f\in\mathcal{F}_k}(L(f) - L_n(f)) \geq 3\mathcal{R}_n(\ell\circ\mathcal{F}_k) + 2\sqrt{\frac{\log(k)}{n}} + \frac{18\log(k)}{n} + 2\delta\right)$$
$$\leq \mathsf{P}\left(\sup_{f\in\mathcal{F}_k}(L(f) - L_n(f)) \geq \mathbb{E}\sup_{f\in\mathcal{F}_k}(L(f) - L_n(f)) + 2\sqrt{\frac{\log(k)}{n}} + \delta\right)$$
$$+ \mathsf{P}\left(\mathcal{R}_n(\ell\circ\mathcal{F}_k) \leq \frac{2}{3}\mathbb{E}\mathcal{R}_n(\ell\circ\mathcal{F}_k) - \frac{18\log(k)}{3n} - \frac{\delta}{3}\right)$$
$$\leq \frac{1}{k^2}\exp(-2n\delta^2) + \frac{1}{k^2}\exp\left(-\frac{n\delta}{9}\right),$$

где последние неравенства получены с помощью неравенства ограниченных разностей и неравенство типа Бернштейна 2.1 для условных Радемахеровских сложностей. Осталось вспомнить, что для случайной величины $Y \geq 0$ имеет место неравенство $\mathbb{E}Y \leq \int_0^\infty \mathsf{P}(Y \geq \delta)d\delta$. На последнем шаге нужно воспользоваться тем, что $\hat{f}_k$ минимизирует эмпирический риск и снова воспользоваться неравенством

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

■

Полученный результат достаточно общий и дает явный метод выбора модели. Вернемся теперь к частному случаю полученной оценки, а именно к структурной минимизации эмпирического риска. Пусть теперь $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_h \subset \ldots$, причем $\mathcal{F}_i$

имеет размерность Вапника–Червоненкиса V_i. Заменив в пенализации условные радемахеровские сложности на верхние оценки их математических ожиданий, получим вариант принципа структурной минимизации эмпирического риска:

Теорема 2.3. *Введем для задачи классификации с бинарной функцией потерь следующую пенализацию:*

$$pen(n,k) = C\sqrt{\frac{V_k}{n}} + 2\sqrt{\frac{\log(k)}{n}}.$$

Пусть $\hat{k} = \arg\min_k(L_n(\hat{f}_k) + pen(k,n))$, *а* $\hat{f}_k$ *минимизирует эмпирический риск в* k*-ой модели. Тогда выполнено следующее оракульное неравенство:*

$$\mathbb{E}\left(L(\hat{f}_{\hat{k}}) - L(f^*))\right) \leq \inf_k \left(L(f_k^*) - L(f^*) + 2C\sqrt{\frac{V_k}{n}} + 2\sqrt{\frac{\log(k)}{n}}\right) + \sqrt{\frac{2\pi}{n}},$$

где C *— наилучшая из констант в верхней оценке Радемахеровской сложности.*

Список литературы

[1] *Boucheron S., Bousquet O., Lugosi G.* Theory of classificatiom: 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'Etŕe de Probabilitŕes de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics. Springer-Verlag, 2011.

[3] *Lugosi G., Wegkamp M.* Complexity Regularization via Localized Random Penalties // The Annals of Statistics, 2004.

[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.