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 Выбор модели, регуляризация и устойчивость

Методы, изложенные в предыдущем разделе, тесно связаны с так называемыми регуляризованными методами. Однако ранее они использовались как методы выбора модели. Мы уже знаем, что в случае когда класс потерь является равномерным классом Гливенко–Кантелли, можно обучаться с помощью минимизации эмпирического риска. В этой лекции мы считаем, что наша задача не обязательно задача классификации, а функция потерь не обязательно индикатор ошибки. Более того мы рассмотрим так называемые *методы минимизации регуляризованного риска* с помощью которых будет доказана агностическая PAC—обучаемость для задачи, класс потерь которой не обязательно является равномерным классом Гливенко–Кантелли.

Если класс решающих правил $\mathcal{F}$ линейный, то отождествим каждую функцию в нем с соответствующим вектором весов $\mathbf{w}$.

Пример 1.1. Рассмотрим хорошо известную задачу *гребневой регрессии* (Ridge Regression). В этой задаче выбирается вектор весов по правилу

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left(L_n(\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2\right),$$

для некоторого $\lambda > 0$. В случае квадратичной функции потерь в классе линейных решающих правил задача имеет аналитическое решение. В наших обозначениях привычнее записывать задачу в нематричном виде

$$\hat{\mathbf{w}} = \arg\min \left(\lambda\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}((x_i, \mathbf{w}) - y_i)^2)\right).$$

Если $\mathbf{A} = \left(\sum\limits_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T\right)^{-1}$, $\mathbf{b} = \sum\limits_{i=1}^{n}\mathbf{x}_i y_i$, то

$$\hat{\mathbf{w}} = (\mathbf{A} + \lambda n I)^{-1}\mathbf{b}.$$

Данный пример является одним из стандартных примеров регуляризованной минимизации риска: к эмпирическому риску прибавляется член $\lambda\|\mathbf{w}\|_2^2$, который не позволяет весам разрастаться слишком сильно. Стоит отметить, что подобная квадратичная регуляризация является частным случаем *регуляризации по Тихонову*.

§1.1 Устойчивость алгоритмов обучения

Неформально *устойчивость* алгоритмов заключается в том, что небольшое изменение обучающей выборке не изменяет значительно результата обучения. Обозначим $S, S^{(i)}$ — две обучающие выборки, отличающиеся только на i-ом объекте.

Лемма 1.1. *Пусть U – равномерное распределение на индексах $\{1, \ldots, n\}$. Тогда для любого метода обучения*

$$\mathbb{E}(L(\hat{f}_S) - L_n(\hat{f}_S)) = \mathbb{E}_{S, i \sim U}(\ell(\hat{f}_{S^{(i)}}, X_i, Y_i) - \ell(\hat{f}_S, X_i, Y_i)).$$

Доказательство.

Воспользуемся аддитивностью математических ожиданий и сравним первые слагаемые:

$$\mathbb{E}(L(\hat{f}_S)) = \mathbb{E}_{S, X', Y'}[\ell(\hat{f}_S, X', Y')] = \mathbb{E}_{S, i \sim U}[\ell(\hat{f}_{S^{(i)}}, X_i, Y_i)].$$

Аналогично доказываем, что

$$\mathbb{E}L_n(\hat{f}_S) = \mathbb{E}_{S, i \sim U}\ell(\hat{f}_S, X_i, Y_i)).$$

■

Говорят, что алгоритм обучения *переобучается*, если частота ошибок выбранного им решающего правила на обучении сильно меньше чем реальная ошибка.

Опр. 1.1. *Алгоритм обучения называется устойчивым в среднем, если для некоторой убывающей функции $\varepsilon : \mathbb{N} \to \mathbb{R}$, такой что $\varepsilon(n) \to 0$ при $n \to \infty$:*

$$\mathbb{E}_{S, i \sim U}(\ell(\hat{f}_{S^{(i)}}, X_i, Y_i) - \ell(\hat{f}_S, X_i, Y_i)) \leq \varepsilon(n).$$

Последняя лемма показывает, что алгоритм не переобучается тогда и только тогда, когда он устойчив в среднем. Однако отсутствие переобучения не есть гарант хорошей обобщающей способности. Пусть, например, алгоритм обучения выбирает всегда некоторое фиксированное решающее правило вне зависимости от обучающей выборки. Тогда он является устойчивым в среднем, более того $\varepsilon(n) \equiv 0$, однако его эмпирическая ошибка, а значит и реальный риск может быть велик.

Опр. 1.2 (Выпуклая задача статистического обучения). *Задача статистического обучения называется выпуклой, если класс $\mathcal{F}$ является выпуклым множеством и для всех (X, Y) функция $\ell(\mathbf{w}, X, Y)$ является выпуклой по первому аргументу.*

Пример 1.2. В задаче линейной регрессии классу $\mathcal{F}$ мы сопоставляем пространство $\mathbb{R}^d$, далее легко видеть, что функция потерь как функция одного аргумента является выпуклой. Таким образом данная задача — выпуклая.

Опр. 1.3 (Сильно выпуклая функция). *Функция называется сильно выпуклой с параметром λ, если для всех $\mathbf{u}, \mathbf{w}, \alpha \in (0, 1)$:*

$$f(\alpha\mathbf{w} + (1-\alpha)\mathbf{u}) \leq \alpha f(\mathbf{w}) + (1-\alpha)f(\mathbf{u}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{w} - \mathbf{u}\|^2.$$

Лемма 1.2. *Имеют место следующие свойства:*

- $f(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2 - 2\lambda$ *сильно выпуклая функция.*
- *Сумма сильно выпуклой и выпуклой функции — сильно выпуклая с тем же параметром.*
- *Если f — сильно выпуклая, гладкая и $\mathbf{u}$ минимизирует f, то*

$$f(\mathbf{w}) - f(\mathbf{u}) \geq \frac{\lambda}{2}\|\mathbf{w} - \mathbf{u}\|_2^2.$$

Упр. 1.1. Доказать лемму.

В выпуклой задаче обозначим $\mathbf{w}_S$ как вектор, получаемый с помощью регуляризованной минимизации эмпирического риска по обучающей выборке S. То есть:

$$\mathbf{w}_S = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left(L_n(\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2\right),$$

Лемма 1.3. *В выпуклой задаче для регуляризованной минимизации эмпирического риска имеет место неравенство:*

$$\lambda\|\mathbf{w}_S - \mathbf{w}_{S^i}\|_2^2 \leq \frac{1}{n}(\ell(\mathbf{w}_{S^i}, X_i, Y_i) - \ell(\mathbf{w}_S, X_i, Y_i)) + \frac{1}{n}(\ell(\mathbf{w}_S, X', Y') - \ell(\mathbf{w}_{S^i}, X', Y'))$$

Доказательство.

Пусть $g_S(\mathbf{w}) = L_{n,S}(\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2$. Данная функция является сильно выпуклой как сумма сильно выпуклой и выпуклой функции. Мы знаем, что

$$g_S(\mathbf{w}) - g_S(\mathbf{w}_S) \geq \frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}_S\|_2^2.$$

С другой стороны:

$$\begin{aligned} g_S(\mathbf{u}) - g_S(\mathbf{w}) &= L_{n,S}(\mathbf{u}) + \lambda\|\mathbf{u}\|_2^2 - L_{n,S}(\mathbf{w}) - \lambda\|\mathbf{w}\|_2^2 \\ &= L_{n,S^{(i)}}(\mathbf{u}) + \lambda\|\mathbf{u}\|_2^2 - L_{n,S^{(i)}}(\mathbf{w}) - \lambda\|\mathbf{w}\|_2^2 \\ &+ \frac{1}{n}(\ell(\mathbf{w}_{S^i}, X_i, Y_i) - \ell(\mathbf{w}_S, X_i, Y_i)) + \frac{1}{n}(\ell(\mathbf{w}_S, X', Y') - \ell(\mathbf{w}_{S^i}, X', Y')). \end{aligned}$$

Заменяем $\mathbf{u}$ на $\mathbf{w}_{S^{(i)}}$ и $\mathbf{w}$ на $\mathbf{w}_S$ получаем, что

$$\lambda\|\mathbf{w}_S - \mathbf{w}_{S^i}\|_2^2 \leq \frac{1}{n}(\ell(\mathbf{w}_{S^i}, X_i, Y_i) - \ell(\mathbf{w}_S, X_i, Y_i)) + \frac{1}{n}(\ell(\mathbf{w}_S, X', Y') - \ell(\mathbf{w}_{S^i}, X', Y'))$$

■

Опр. 1.4 (Липшицева функция потерь). *Функция потерь ℓ называется липшицевой с параметром ρ, если для всех допустимых X, Y функция, задаваемая $\ell(\mathbf{w}, X, Y)$, является липшицевой по первому аргументу с параметром ρ:*

$$\|\ell(\mathbf{w}, X, Y) - \ell(\mathbf{u}, X, Y)\| \leq \rho\|\mathbf{w} - \mathbf{u}\|.$$

Лемма 1.4. *Если функция потерь ℓ является липшицевой с параметром ρ, то в выпуклой задаче регуляризованный метод минимизации эмпирического риска является устойчивым в среднем с $\varepsilon(n) = \frac{2\rho^2}{\lambda n}$.*

Доказательство.

Воспользуемся предыдущей леммой. Из определения липшицевости

$$\ell(\mathbf{w}_{S^i}, X_i, Y_i) - \ell(\mathbf{w}_S, X_i, Y_i) \le \rho\|\mathbf{w}_{S^i} - \mathbf{w}_S\|.$$

Таким образом,

$$\lambda\|\mathbf{w}_S - \mathbf{w}_{S^i}\|^2 \le \frac{2\rho\|\mathbf{w}_{S^i} - \mathbf{w}_S\|}{n}$$

Следовательно,

$$\|\mathbf{w}_S - \mathbf{w}_{S^i}\| \le \frac{2\rho}{\lambda n}.$$

Еще раз применяя Липшицевость, получаем, что

$$\ell(\mathbf{w}_{S^i}, X_i, Y_i) - \ell(\mathbf{w}_S, X_i, Y_i) \le \frac{2\rho^2}{\lambda n}.$$

Беря математические ожидания, получаем утверждение леммы. ■

2 Баланс эмпирического риска и устойчивости

Для любого правила $\hat{f}$ можно сделать очевидное разложение:

$$\mathbb{E}L(\hat{f}) = \mathbb{E}L_n(\hat{f}) + \mathbb{E}\left(L(\hat{f}) - L_n(\hat{f})\right).$$

Первый член отвечает за ожидаемую ошибку на обучающей выборке, а второй — за устойчивость. Для того чтобы ожидаемый риск был мал необходимо чтобы оба слагаемых были в балансе. Рассмотрим, например случай выпуклой и липшицевой функции потерь в задачи регуляризованных методов минимизации эмпирического риска. Очевидно, что для любого вектора $\mathbf{w}^*$ имеет место соотношение:

$$L_n(\mathbf{w}_S) \le L_n(\mathbf{w}_S) + \lambda\|\mathbf{w}_S\|^2 \le L_n(\mathbf{w}^*) + \lambda\|\mathbf{w}^*\|^2.$$

Теорема 2.1 (Оракульное неравенство). *Если функция потерь ℓ является липшицевой с параметром ρ, то в выпуклой задаче для регуляризованного метода минимизации эмпирического риска для любого вектора $\mathbf{w}^*$ выполнено:*

$$\mathbb{E}L(\mathbf{w}_S) \le L(\mathbf{w}^*) + \lambda\|\mathbf{w}^*\|^2 + \frac{2\rho^2}{\lambda n}.$$

Доказательство.

Доказательство получается с помощью неравенства $L_n(\mathbf{w}_S) \le L_n(\mathbf{w}^*) + \lambda\|\mathbf{w}^*\|^2$ и леммы 1.4. ■

Термин оракульное неравенство здесь уместен. Каждый вектор весов можно рассматривать как отдельную модель, а регуляризованный метод минимизации эмпирического риска как метод выбора модели. Тогда данное неравенство сравнивает ожидаемый риск выбранной модели, в частности, с риском оракульной модели.

Теорема 2.2 (Оракульное неравенство, ограниченный случай). *Если функция потерь ℓ является липшицевой с параметром ρ, то в выпуклой задаче для регуляризованного метода минимизации эмпирического риска параметр λ может быть выбран таким образом, что будет выполнено неравенство:*

$$\mathbb{E}L(\mathbf{w}_S) \le \inf_{\mathbf{w}} L(\mathbf{w}) + \rho B\sqrt{\frac{8}{n}}.$$

Упр. 2.1. Доказать теорему.

Упр. 2.2. Доказать агностическую PAC–обучаемость выпуклой ограниченной задачи с липшицевой функцией потерь.

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] *Massart P.* Concentration Inequalities and Model Selection // Ecole d'Etŕe de Probabilitŕes de Saint-Flour XXXIII-2003. Springer-Verlag, 2007.

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