Домашнее задание по оценкам обобщающей способности 2. ММП, весна 2013 13 апреля

В списке литературы приведен хороший текст о радемахеровской сложности.

1. Рассмотрим классификацию с двумя классами $\mathbb{Y} = \{-1, +1\}$ и бинарной функцией потерь $\ell(y, y') = [y \neq y]$. Зафиксируем множество классификаторов $\mathcal{G} = \{g\colon \mathbb{X} \to \mathbb{Y}\}$.

   (a) Докажите для этого случая следующее свойство радемахеровской сложности:

   $$\mathcal{R}(\ell \circ \mathcal{G}) = \mathbb{E}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell\bigl(g(X_i), Y_i\bigr)\right] = \frac{1}{2}\mathbb{E}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(X_i)\right] = \frac{1}{2}\mathcal{R}(\mathcal{G}).$$

   (b) Покажите также, что то же самое свойство выполняется и для условной радемахеровской сложности:

   $$\mathcal{R}_n(\ell \circ \mathcal{G}) = \mathbb{E}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell\bigl(g(X_i), Y_i\bigr) \Big| X^n\right] = $$
   $$= \frac{1}{2}\mathbb{E}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(X_i) \Big| X^n\right] = \frac{1}{2}\mathcal{R}_n(\mathcal{G}).$$

   Подсказка: При доказательстве обеих утверждений воспользуйтесь следующим полезным наблюдением (которое тоже стоит доказать): *случайные величины σ_i и $\sigma_i Y_i$ имеют одинаковые распределения.*

   Таким образом мы показали, что в случае классификации, теорема 2.2 с последнего семинара запишется в виде:

   $$L(g) \leqslant L_n(g) + \mathcal{R}(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}, \tag{*}$$

   и

   $$L(g) \leqslant L_n(g) + \mathcal{R}_n(\mathcal{G}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}. \tag{**}$$

2. Продолжая полученные результаты прошлой задачи, докажите, что при тех же условиях, справедливо равенство

$$\frac{1}{2}\mathcal{R}_n(\mathcal{G}) = \frac{1}{2} - \mathbb{E}_\sigma \left[\inf_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \big[g(X_i) \neq \sigma_i \big] \right].$$

Интерпретируйте полученный результат: что нужно сделать, чтобы при фиксированных знаках $\sigma_1, \ldots, \sigma_n$ вычислить условную радемахеровскую сложность? Также, пользуясь последней формулой, еще раз подумайте над смыслом радемахеровской сложности. В частности: когда $\mathcal{R}_n(\mathcal{G}) = 1$? О чем это свидетельствует?

3. Пусть $\mathcal{G}$ и $\mathcal{F}$ — два семейства ограниченных функций, отображающих пространство $\mathcal{X}$ в $\mathbb{R}$. Зафиксируем выборку $X^n = \{x_1, \ldots, x_n\} \subset \mathcal{X}$. Докажите следующие свойства условной радемахеровской сложности:

   (a)
   $$\mathcal{R}_n(\mathcal{G} \cup \mathcal{F}) \leqslant \mathcal{R}_n(\mathcal{G}) + \mathcal{R}_n(\mathcal{F}).$$

   (b) Для $c > 0$ и $c \cdot \mathcal{G} = \{cg(x), g \in \mathcal{G}\}$ выполнено
   $$\mathcal{R}_n(c \cdot \mathcal{G}) = c \cdot \mathcal{R}_n(\mathcal{G}).$$

   (c) Пусть класс $\mathcal{G} = \{g_1, \ldots, g_N\}$ имеет конечную мощность N. Рассмотрим выпуклую оболочку этого класса функций
   $$\operatorname{conv}(\mathcal{G}) = \left\{ \sum_{i=1}^{N} \alpha_i g_i, \quad \sum_{i=1}^{N} \alpha_i = 1, \quad \alpha_i \geqslant 0,\ i = 1, \ldots, N \right\}.$$

   Покажите, что
   $$\mathcal{R}_n(\mathcal{G}) = \mathcal{R}_n\big(\operatorname{conv}(\mathcal{G})\big).$$

4. Свойство (c) имеет одно очень интересное следствие.

   Представим, что мы решаем задачу регрессии $\mathbb{Y} = \mathbb{R}$ с квадратичной функцией потерь $\ell(y, y') = (y - y')^2$. При этом будем считать, что ответы объектов $\mathbb{Y}$ и значения используемых нами функций из $\mathcal{G}$ по модулю ограничены константой 1. Тем самым мы удовлетворили требованию, использовавшемуся на протяжении всех последних семинаров, что функция потерь ℓ ограничена единицей (убедитесь в этом). В этом случае, оказывается, справедливо следующее утверждение (без доказательства):
   $$\mathcal{R}_n(\ell \circ \mathcal{G}) \leqslant 2\mathcal{R}_n(\mathcal{G}).$$

   Это свойство очень похоже на то, что мы вывели в задаче 1: в обоих случаях мы получаем возможность «избавиться» от функции потерь и перейти к самому классу алгоритмов $\mathcal{G}$.

   Предположим, что у нас есть конечный «базовый» класс $\mathcal{G} = \{g_1, \ldots, g_N\}$ функций. Для решения задачи регрессии мы будем использовать определенный ранее класс $\operatorname{conv}(\mathcal{G})$. Докажите следующую теорему

Теорема 0.1 *Для любого $\delta > 0$ с вероятностью не меньше $1-\delta$ одновременно для всех функций $g \in \mathrm{conv}(\mathcal{G})$ справедливо*

$$L(g) \leqslant L_n(g) + 4\mathcal{R}_n(\mathcal{G}) + 3\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Подсказка: достаточно воспользоваться теоремой 2.2 из последнего семинара (убедитесь, что ей воспользоваться мы действительно можем), а также воспользоваться некоторыми из уже доказанных свойств условной радемахеровской сложности.

То, что мы сейчас доказали — сильный результат, поскольку в правую часть неравенства входит лишь радемахеровская сложность базового класса. Получается, если у нас есть некоторый конечный набор функций (возможно, предварительно полученных нами разными методами машинного обучения), то усреднение этих функций не ведет к увеличению сложностного слагаемого оценки теоремы 2.2. Казалось бы, «мы усложняем используемое семейство», но сложность этого семейства не увеличивается. Очень похожим способом обосновывается устойчивость к переобучению бустинга — ведь бустинг выбирает в качестве классификатора взвешенную композицию базовых классификаторов.

Список литературы

[1] *Maria-Florina Balcan* Lecture 25. Rademacher complexity. — Machine Learning Theory, Georgia Tech Lectures, 2011.
http://www.cc.gatech.edu/~ninamf/ML11/lect1115.pdf