Теория статистического обучения

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

9 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 Размерность Вапника–Червоненкиса

В данном разделе речь будет идти только о задачи классификации с бинарной функцией потерь. Для фиксированной обучающей выборки $(X_i, Y_i)_{i=1}^n$ можно определить проекцию $\mathcal{F}$ на эту выборку, как множество различных булевых векторов:

$$\mathcal{F}_{(X_i,Y_i)_{i=1}^n} = \{(\mathbf{I}\{f(X_1) \neq Y_1\}, \dots, \mathbf{I}\{f(X_n) \neq Y_n\}) : f \in \mathcal{F}\}.$$

Функцией роста назовем верхнюю грань по всевозможным выборкам мощности построенной проекции:

$$S_{\mathcal{F}}(n) = \sup_{(X_i,Y_i)_{i=1}^n} |\mathcal{F}_{(X_i,Y_i)_{i=1}^n}|.$$

Очевидно, что если $|\mathcal{F}| = N$, то $S_{\mathcal{F}}(n) \leq N$. Легко показать, что имеют место следующие соотношения:

- $S_{\mathcal{F}}(n) \leq 2^n$.
- $S_{\mathcal{F}}(n+m) \leq S_{\mathcal{F}}(n) S_{\mathcal{F}}(m)$.
- если $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$, то $S_{\mathcal{F}}(n) \leq S_{\mathcal{F}_1}(n) + S_{\mathcal{F}_2}(n)$.

Размерностью Вапника-Червоненкиса семейства $\mathcal{F}$ назовем наибольшее натуральное число V, при котором

$$S_{\mathcal{F}}(V) = 2^V.$$

В случае, если для данного семейства классификаторов такого числа не существует, то считаем, что $V = \infty$.

Пример 1.1. Одномерное семейство пороговых решающих правил

$$\mathcal{F} = \{f_\theta(x) = \mathbf{I}\{x \leq \theta\} : \theta \in [0, 1]\}$$

имеет размерность Вапника-Червоненкиса, равную единице.

Пример 1.2. Семейство классификаторов, представляющее собой семейство разделюящих d-мерных гиперплоскостей имеет размерность Вапника–Червоненкиса, равную $d+1$. Данное утверждение связано с теоремой Радона.

Утв. 1.1 (теорема Радона). *Произвольное подмножество из $d+2$ или более точек d-мерного евклидова пространства может быть разделено на два непересекающихся подмножества, чьи выпуклые оболочки имеют непустое пересечение.*

Пример 1.3. Семейство классификаторов

$$\{\text{sgn}(\sin(tx)) : t \in \mathbb{R}\}$$

имеет размерность равную ∞, даже несмотря на то, что параметризуется лишь одним параметром.

Семейство классификаторов, обладающее конечной ёмкостью обладает замечательным свойством:

Лемма 1.2 (Зауэр, Вапник-Червоненкис). *Для любого семейства классификаторов с размерностью Вапника–Червоненкиса V для $n \geq V$:*

$$S_{\mathcal{F}}(n) \leq \sum_{i=0}^{V} C_n^i$$

Доказательство.

Зафиксируем некоторую выборку $(X_i, Y_i)_{i=1}^n$, на которой достигается супремум в определении функции роста. Пусть $\mathcal{F}_0 = \mathcal{F}_{(X_i,Y_i)_{i=1}^n}$ – соответсвующая проекция. Будем говорить, что множество булевых векторов $\mathcal{F}_i$ *разбивает* множество индексов $S = \{s_1, \ldots, s_m\}$, если ограничение $\mathcal{F}_i$ на эти индексы реализует полный m–мерный булев куб.

Пронумеруем векторы в $\mathcal{F}_0$. Зафиксируем множество первых компонент этих векторов. Последовательно для каждой 1-чной компоненты заменим 1 на 0 в том случае, если данная процедура не создаст повторных векторов в $\mathcal{F}_0$. С нулевыми компонентами не сделаем никаких изменений. После осуществления всех возможных таких замен для первого столбца получаем некоторое множество векторов $\mathcal{F}_1$. Оно совпадает по мощности со множеством $\mathcal{F}_0$ и обладает следующим замечательным свойством: каждое множество S, разбиваемое $\mathcal{F}_1$, разбивается и $\mathcal{F}_0$. Затем по аналогии для второго столбца строим из $\mathcal{F}_1$ множество $\mathcal{F}_2$. И так далее по всем столбцам до множества $\mathcal{F}_n$.

Множество $\mathcal{F}_n$ имеет ту же мощность, что и $\mathcal{F}_0$ и не разбивает ни одного множества мощностью больше чем V. Более того, если $\mathbf{b} \in \mathcal{F}_n$, то для любого $\mathbf{b}' \in \{0,1\}^n$ такого, что $\mathbf{b}'_i \leq \mathbf{b}_i$ имеет место включение $\mathbf{b}'_i \in \mathcal{F}_n$. Таким образом, в $\mathcal{F}_n$ могут быть только векторы, которые содержат не более n единичных компонент, так как иначе $\mathcal{F}_n$ разбило бы некоторое множество, состоящее более чем из V индексов. Максимальная мощность множества булевых векторов с не более чем V единицами равна $\sum_{i=0}^{V} C_n^i$, что и доказывает утверждение леммы. ■

С помощью леммы Зауера можно получить верхнюю полиномиальную верхнюю оценку на функцию роста:

$$S_{\mathcal{F}}(n) \leq (n+1)^V$$

Пусть мы имеем дело с задачей классификации с бинарной функцией потерь. Тогда 4-ое свойство можно переписать в виде

$$\mathcal{R}_n(\ell \circ \mathcal{F}) \leq \sqrt{\frac{2\log(2S_{\mathcal{F}}(n))}{n}}.$$

где неравенство выполнены почти наверное. Что в случае конечной размерности Вапника–Червоненкиса даёт порядок $O\left(\sqrt{\frac{V\log(n)}{n}}\right)$.

Особенность Радемахеровского процесса заключается, что его можно анализировать с помощью горадо более мощных средств теории эмпирических процессов. Действительно, можно рассматривать процесс $\left|\sum_{i=1}^{n}\sigma_i a_i\right|$ как верхнюю оценку эмпирического процесса $\sup_{f\in\mathcal{F}}|L_n(f)-L(f)|$ со множеством состояний A, где A — проекция класса потерь на конечную выборку. В этом случае Радемахеровское среднее есть ни что иное, как ожидаемый супремум этого процесса. Теория эмпирических процессов показывает, что во многих случаях поведение процесса зависит от 'геометрии' пространства состояний. В нашем случае – это метрические свойства множества A.

Условно по обучающей выборке множество $A = A\left((X_i, Y_i)_{i=1}^n\right)$ можно представить себе как набор не более чем $S_{\mathcal{F}}(n)$ различных булевых векторов. Введем на на паре векторов метрику ρ:

$$\rho(a,b) = \sqrt{\frac{1}{n}d_H(a,b)},$$

где d_H – метрика Хэмминга.

Будем говорить, что множества $B \subset \{0,1\}^n$ является ε-покрытием множества A, если объединение замкнутых ε-шаров (по введенной метрике) с центрами в точках B содержат A.

Обозначим $N(\varepsilon, A)$ – *число покрытия*, равное мощности минимального ε-покрытия множества A.

Теорема 1.3. *Для задачи классификации*

$$\mathcal{R}_n(\ell \circ \mathcal{F}) \leq \frac{12}{\sqrt{n}} \sup_{(X_i,Y_i)_{i=1}^n} \int_0^1 \sqrt{\log(2N(\varepsilon, A))}d\varepsilon,$$

где $A = A\left((X_i, Y_i)_{i=1}^n\right)$.

Доказательство.

Зафиксируем конечное множество различных n мерных булевых векторов A. Зафиксируем $B^{(0)} = \{(0,\ldots,0)\}$ – множество состоящее из нулевого вектора, а $B_1, \ldots, B_M$ подмножества $\{0,1\}^n$, являющиеся минимальными 2^{-k}-покрытиями множества A, а $M = \lfloor \log_2(\sqrt{n})\rfloor + 1$.

Пусть для конкретной реализации σ_i вектор $b^* \in A$ доставляет максимум выражения $\left|\sum_{i=1}^{n} \sigma_i b_i\right|$, среди всех векторов A. Обозначим $b^{(k)}$–ближайший к нему вектор в B_k. Из неравенства треугольника так как $\rho(b^{(k)}, b^*) \leq 2^{-k}$ мы имеем

$$\rho(b^{(k)}, b^{(k-1)}) \leq 2^{-k} + 2^{-k+1} = 3 \times 2^{-k}.$$

Тогда

$$\sum_{i=1}^{n} \sigma_i b_i^* = \sum_{k=1}^{M} \sum_{i=1}^{n} \sigma_i (b_i^{(k)} - b_i^{(k-1)}).$$

Тогда

$$\mathbb{E}\left\{\max_{b\in A}\left|\sum_{i=1}^{n}\sigma_i b_i\right|\right\} =$$
$$\mathbb{E}\left\{\left|\sum_{k=1}^{M}\sum_{i=1}^{n}\sigma_i(b_i^{(k)} - b_i^{(k-1)})\right|\right\} \leq$$
$$\sum_{k=1}^{M}\mathbb{E}\left\{\left|\sum_{i=1}^{n}\sigma_i(b_i^{(k)} - b_i^{(k-1)})\right|\right\} \leq$$
$$\sum_{k=1}^{M}\mathbb{E}\left\{\max_{b\in B_k, c\in B_{k-1}, \rho(b,c)\leq\frac{3}{2^k}}\left|\sum_{i=1}^{n}\sigma_i(b_i^{(k)} - b_i^{(k-1)})\right|\right\}.$$

Математическое ожидание под суммой можно представить как математическое ожидание максимума модулей $|B_k||B_{k-1}| \leq N(2^{-k}, A)^2$ экземпляров субгауссовских случайных величин с параметром $\sigma^2 = n(3/2^k)^2$. Условия на параметр σ^2 получаются из независимости σ_i и леммы Хеффдина. Применяя теперь лемму о математическом ожидании максимума субгауссовских величин получаем:

$$\mathbb{E}\left\{\max_{b\in B_k, c\in B_{k-1}, \rho(b,c)\leq\frac{3}{2^k}}\left|\sum_{i=1}^{n}\sigma_i(b_i^{(k)} - b_i^{(k-1)})\right|\right\} \leq 3\sqrt{n}2^{-k}\sqrt{2\log(2N(2^{-k}, A)^2)}.$$

А значит

$$\mathbb{E}\left\{\max_{b\in A}\left|\sum_{i=1}^{n}\sigma_i b_i\right|\right\} =$$
$$3\sqrt{n}\sum_{k=1}^{M}2^{-k}\sqrt{2\log(2N(2^{-k}, A)^2)} \leq$$
$$12\sqrt{n}\sum_{k=1}^{\infty}2^{-k-1}\sqrt{\log(2N(2^{-k}, A))} \leq$$
$$12\sqrt{n}\int_0^1\sqrt{\log(2N(\varepsilon, A))}d\varepsilon.$$

■

Полученная теорема говорит, что Радемахеровское среднее контролируется не логарифмом мощности множества A, а некоторой величиной, которая существенно учитывает структуру A. Будем называть величину $\int\limits_0^1 \sqrt{\log(2N(\varepsilon, A))}d\varepsilon$ *метрической энтропией* множества A.

Важность полученного результата связна с использованием следующей теоремы

Теорема 1.4 (Haussler [2]). *Если множество булевых векторов A состоит из различных векторов ошибок семейства классификаторов с размерностью Вапника-Червоненкиса равной V, то для $0 \leq \varepsilon \leq 1$:*

$$N(\varepsilon, A) \leq e(V+1)\left(\frac{2e}{\varepsilon^2}\right)^V.$$

Применяя данную теорему можно получить, что для некоторой абсолютной константы C для задачи классификации с бинарной функцией потерь

$$\mathcal{R}_n(\ell \circ \mathcal{F}) \leq C\sqrt{\frac{V}{n}}.$$

Пример 1.4 (Теорема Дворецкого-Кифера-Вольфовитца). С помощью данного результата можно получить усиление теоремы Гливенко-Кантелли о равномерной сходимости эмпирической функции распределения к настоящей функции распределения. Пусть $F(x)$ — функция распределения, а $F_n(x)$ — эмпирическая функция распределения. Можно считать, что $x \in \mathbb{R}$ индексирует некоторые классификаторы, которые ошибаются на всех объектах (X, Y) тогда и только тогда, когда $X \leq x$, то есть $\ell(f, X, Y) = \mathbf{I}\{X \leq x\}$. Такие классификаторы обладают единичной размерностью Вапника–Червоненкиса. Таким образом, для некоторой $C > 0$

$$\mathbb{E}\left\{\sup_x |F_n(x) - F(x)|\right\} \leq \frac{C}{\sqrt{n}}$$

Более общий вариант теоремы даёт неулучшаемую явную константу $C = 1$ [4], а также задает хвосты распределения $\sup\limits_x |F_n(x) - F(x)|$.

Упр. 1.1. С помощью неравенства ограниченных разностей оцените хвосты $\sup\limits_x |F_n(x) - F(x)|$.

Список литературы

[1] *Devroye L., Lugosi G.* Combinatorial Methods in Density Estimation // Springer Series in Statistics. Springer-Verlag, 2001.

[2] *Haussler D.* Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension // Journal of Combinatorial Theory. — 1995. — Pp. 217–232.

[3] *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.

[4] *Massart P.* The tight constant in Dvoretzky-Kiefer-Wolfowitz inequality // Annals of Probability, 1990.

[5] *Rakhlin A.* Statistical Learning Theory and Sequential Prediction // Lecture notes, 2014, http://www-stat.wharton.upenn.edu/ rakhlin/

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

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