Секция «Математика и механика»

Скорость сходимости оценки регрессионной функции, построенной по модифицированному методу ближайшего соседа

Хапланов Арсений Юрьевич
Аспирант
Московский государственный университет имени М.В. Ломоносова,
Механико-математический факультет, Москва, Россия
E-mail: khaplanova@gmail.com

Для построения оценки регрессионной функции в [1] использован метод ближайшего соседа в сочетании с равномерным усреднением. Полученная оценка оказалась состоятельной. Нас интересует среднеквадратичное отклонение упомянутой оценки от регрессионной функции.

Предположим, что имеется последовательность независимых одинаково распределенных векторов $(X, Y), (X_1, Y_1), \ldots$ со значениями в $\mathbb{R}^d \times \mathbb{R}$. Пусть дана реализация (x_i, y_i), $i = 1, \ldots, n$. Из набора $(x_1, \ldots, x_n)$ выбираем без возвращения k_n элементов и проделываем это независимым образом l_n раз. Для каждой выборки ищется ближайшая к x в евклидовой метрике точка x_{i_j} $(j = 1, \ldots, l_n)$. Оценка регрессионной функции $r(x) = \mathbb{E}(Y|X = x)$, где $x \in \mathbb{R}^d$, задается формулой

$$r_n(x) = \frac{1}{l_n} \sum_{j=1}^{l_n} Y_{i_j}.$$

Будем обозначать условную дисперсию $\sigma^2(x) = \operatorname{var}(Y \mid X = x)$.

Теорема. *Пусть существуют положительные константы* σ, C, M, α, *для которых* $P(\|X\| > t) \le e^{-\alpha t}$ *при всех* $t \ge M$, *а также для любых* $x, x' \in \mathbb{R}^d$

$$\sigma^2(x) \le \sigma^2, \ |r(x) - r(x')| \le C \|x - x'\|,$$

где $\|\cdot\|$ *– евклидова норма. Тогда найдется константа* $R > 0$, *зависящая только от* α *и* M, *такая, что верно неравенство*

$$\mathbb{E}\left([r_n(X) - r(X)]^2\right) \le \sigma^2 \left[\frac{1}{l_n} + \left(1 - \frac{1}{l_n}\right) \frac{k_n}{n} \frac{1}{(1 - k_n/n + 1/n)^2}\right] + 4\, RC^2 (\ln k_n)^3\, f_d(n).$$

Здесь $f_1(n) = 2\, k_n^{-1/2}$, $f_2(n) = \sqrt{2(1 + \ln k_n)}\, k_n^{-1/2}$, $f_d(n) = \sqrt{\frac{2}{1-2/d}}\, k_n^{-1/d}$, *если* $d \ge 3$.

Заметим, что в отличие от [2] не предполагается ограниченность нормы вектора X.

Литература

1. Biau G. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification // J. Multivariate Analysis. 2010. No. 101. C. 2499–2518.

2. Biau G. On the rate of convergence of the bagged nearest neighbor estimate // J. Machine Learning Research. 2010. No. 11. P. 687-712.