Секция «Математика и механика»

Сходимость оценки регрессионной функции, построенной по методу ближайших соседей в случайной метрике

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

Для построения оценки регрессионной функции одним из часто используемых методов является метод ближайших соседей, основанный на бутстрапе. Этот метод предложил Брау в 1996 году[1].

Предположим, что имеется последовательность независимых пар $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ с распределением (X, Y). Для построения оценки регрессионной функции $r(x) = \mathbb{E}(Y|X = x)$ в точке x создается l_n независимых выборок из $\{X_i\}$ мощности k_n. Для каждой выборки ищется ближайшая к x в евклидовой метрике точка $X_{i_j}, j = 1, \ldots, l_n$. Тогда оценка будет равна:

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

Если добавить условие, что $\mathbb{E}|Y|^p \leq \infty$ для некоторого фиксированного $p \geq 1$, то данная оценка будет универсальной L_p-состоятельной. Можно ослабить условие на метрику, заменить евклидову на некоторую случайную. В результате получим новую теорему.

Теорема 1.

Пусть есть последовательность независимых одинаково распределенных $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, где распределение X_1 имеет ограниченный носитель. Также дана последовательность случайных метрик ρ_n, определяемых соотношением

$$\rho_n(x, y) = \|A_n(x - y)\|.$$

Матрица A_n зависит только от $X_1, \ldots, X_n$. Пусть для этой метрики вероятность события, что какие-либо 2 точки X_i и X_j будут равноудалены от некоторой фиксированной точки x, равна нулю. И пусть существуют 2 последовательности неотрицательных случайных величин $\{m_n\}$ и $\{M_n\}$ таких, что для любых $n \in \mathbb{N}$ и $x, y \in \mathbb{R}^d$ выполнено двойное неравенство

$$m_n \|x - y\| \leq \rho_n(x, y) \leq M_n \|x - y\|.$$

Если выполнено

$$\mathrm{P}\left(\liminf_{n\to\infty} \frac{m_n}{M_n} = 0\right) = 0$$

и $l_n \to \infty$, $k_n \to \infty$ и $\frac{k_n}{n} \to 0$, тогда для любого $p \geq 1$ оценка $r_n^\star$ универсальная L_p-состоятельная.

Литература

1. Breiman L. Bagging predictors // Machine Learning. 1996. No. 24. P. 123-140.