Секция «Математика и механика»

Оценка уклонения эмпирически оптимального классификатора, определяемого схемой голосования.

Атрощенко Михаил Юрьевич
Студент
Московский государственный университет имени М.В. Ломоносова,
Механико-математический факультет, Москва, Россия
E-mail: m.atroshenko@gmail.com

Пусть дана выборка $D_n = \{(X_1, Y_1), \dots (X_n, Y_n)\}$, которая состоит из независимых пар $(X_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\}$, распределенных так же, как пара (X, Y). Рассмотрим семейство классификаторов вида

$$\mathcal{G}_p = \{g(x) = \operatorname{sgn}(c_0 + c_1 h_1(x) + \dots c_p h_p(x)); c_i \in \mathbb{R}, h_i \in \mathcal{H}\},$$

где $p \geq 1$ – целое, а $\mathcal{H}$ — есть некоторое множество функций $h : \mathbb{R}^d \to \{-1, 0, 1\}$, причем $|\mathcal{H}| < \infty$.

Выбор классификатора из этого класса осуществляется на основе минимизации эмпирического риска

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(g(X_i) \neq Y_i).$$

Введем обозначение

$$g_n^* = \arg\min_{g \in \mathcal{G}_p} L_n(g).$$

Основным результатом является следующая теорема.

Теорема 1.
Для любого $\varepsilon > 0$, натуральных n и p верно неравенство

$$\mathsf{P}\left(\mathsf{P}(g_n^*(X) \neq Y | D_n) - \inf_{g \in \mathcal{G}_p} \mathsf{P}(g(X) \neq Y) > \varepsilon\right) \leq \frac{2c}{p!} 3^{p^2} \left(\frac{|\mathcal{H}|e}{p}\right)^p e^{-n\varepsilon^2/2},$$

где константа $c \leq 4e^{4\varepsilon + 4\varepsilon^2}$.

Литература

1. Вапник В., Червоненкис А. Теория распознавания образов. Наука. Москва., 1974
2. Devroye L., Gyofri L., Lugosi G. Probobalistic theory of pattern recognition. Springer., 1996