Кожевин А. А. (Москва, Россия). **Информационный подход к отбору значимых признаков**

Доклад посвящён оценкам условной энтропии [1] и совместной информации в смешанной модели [2], а также новому результату о процедуре отбора значимых признаков (факторов), основанной на информационном подходе.

Смешанная модель описывается в работе [1]. Пусть $\zeta_n = \{(X^i, Y^i)\}_{i=1}^n$ — выборка из н.о.р.с.в., $(X^1, Y^1) \sim (X, Y)$, где $X = (X_1, \dots, X_d)$ — абсолютно непрерывный случайный вектор со значениями в $\mathbb{R}^d$, Y — дискретная случайная величина со значениями в конечном множестве M. Набор индексов $S = \{s_1, \dots s_m\} \subset \{1, \dots, d\}$ ($s_i \neq s_j$ для $i \neq j$) и набор факторов X_S, где $u_L = (u_{l_1}, \dots, u_{l_m})$ для $u = (u_1, \dots, u_d)$ и $L = \{l_1, \dots, l_m\}$, будем называть значимыми, если для каждого $y \in M$ выполняется $f_{Y|X}(y|X) = f_{Y|X_S}(y|X_S)$ п.н. Пусть $Q_m = \{\{l_1, \dots, l_m\} \subset \{1, \dots, d\} \colon l_i \neq l_j, i \neq j\}$. Для каждого $L \in Q_m$ составим выборку $\zeta_{n,L} = \{(X_L^i, Y^i)\}_{i=1}^n$. Оценим совместную информацию $I(X_L; Y)$ для каждой выборки $\zeta_{n,L}$ с помощью метода, предложенного в работе [2]. Соответствующие значения оценок обозначим $\hat{I}_{n,k,L}$, где $k \in \{1, \dots, n-1\}$ — параметр метода.

Положим $\hat{S}_{n,k} = argmax_{L \in Q_m} \hat{I}_{n,k,L}$. Если максимум $\hat{I}_{n,k,L}$ достигается на нескольких множествах из Q_m, то в качестве $\hat{S}_{n,k}$ возьмем первое из них в смысле лексикографического порядка. Для такой процедуры отбора значимых признаков верен следующий новый результат.

Теорема. Пусть m известно, набор значимых факторов длины m единственен. Пусть плотность $f_X(\cdot)$ строго положительна, для каждого набора $L \subset \{1, \dots, d\}$ плотность $f_{X_L,Y}(\cdot, y)$ для каждого $y \in M$ является C_0-стягиваемой ($C_0 > 0$), а также для некоторого $\varepsilon > 0$ выполняется $\mathsf{E}|\log f_{X_L}(X_L)|^{2+\varepsilon} < \infty$. Тогда $\mathsf{P}(\hat{S}_{n,k} = S) \to 1$ при $n \to \infty$ для любого $\alpha \in (0, 1)$ и $k \propto n^\alpha$.

СПИСОК ЛИТЕРАТУРЫ

1. *Bulinski, A., Kozhevin, A.* Statistical Estimation of Conditional Shannon Entropy. ESAIM: Probability and Statistics (Accepted for publication November 28, 2018), p. 1-37, DOI: https://doi.org/10.1051/ps/2018026.
2. *Bulinski, A., Kozhevin, A.* Statistical Estimation of Mutual Information for Mixed Model. (to appear)

Работа выполнена при поддержке гранта МГУ им.М.В.Ломоносова «Современные проблемы фундаментальной математики и механики»