Секция «Математика и механика»

Скорость сходимости оценки параметров многомерной логистической регрессии

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

Логистическая регрессия часто используется в машинном обучении для решения проблем классификации. Предполагается, что бинарная случайная величина Y определенным образом зависит от факторов $X_1, \ldots, X_p$, принимающих действительные значения. Нас будет интересовать обобщение модели, предложенной в [1]. А именно, пусть для каждого $n \in \mathbb{N}$ величина $Y = Y(n)$ отображает пространство элементарных исходов Ω в множество $\{0, 1, \ldots, k\}$, а $X = X(n)$ — случайный вектор со значениями в $\mathbb{R}^{p_n}$. Пусть для $j \in \{1, \ldots, k\}$ и $x \in \mathbb{R}^{p_n}$ справедливы соотношения:

$$\mathbb{P}\left(Y(n) = j \,|X(n) = x\,\right) := \pi_j(\alpha^{0,n}, x) = \frac{\exp\{-(\alpha_j^{0,n})^T x\}}{1 + \sum_{t=1}^k \exp\{-(\alpha_t^{0,n})^T x\}},$$

$$\mathbb{P}\left(Y(n) = 0 \,|X(n) = x\,\right) = 1 - \sum_{j=1}^k \pi_j(\alpha^{0,n}, x),$$

где T обозначает транспонирование, неслучайные векторы $\alpha_j^{0,n} \in \mathbb{R}^{p_n}$. По независимым наблюдениям векторов $(X_q(n)^T, Y_q(n))^T$, $q = 1, \ldots, n$, имеющих такое же распределение, как $(X(n)^T, Y(n))^T$, должным образом строится оценка $\widehat{\alpha}_n$ вектора $\alpha^{0,n} = \{(\alpha_1^{0,n})^T, \ldots, (\alpha_k^{0,n})^T\}^T$. Положим $R_n(\alpha) = \left\{(R_n(\alpha)^1)^T, \ldots, (R_n(\alpha)^k)^T\right\}^T$, где $R_n(\alpha)^r = (\{\sum_{q=1}^n \left[\mathbb{I}\left\{Y_q(n) = r\right\} - \pi_r(\alpha, X_q(n))\right] X_q^l(n)\}_{l=1}^{p_n})^T, \alpha \in \mathbb{R}^{kp_n}$. Приведем наш основной результат, обобщающий работу [2], в которой рассматривался случай $k = 1$.

Теорема. *Пусть $p_n^3/n \to 0$ при $n \to \infty$. Предположим, что существуют константы $L, c > 0$ такие, что для любого $n \in \mathbb{N}$ норма $X(n)$ ограничена константой L с вероятностью единица, и все собственные значения ковариационной матрицы вектора $R_n(\alpha^{0,n})/\sqrt{n}$ лежат в полуинтервале $[c, \infty)$. Тогда для любой последовательности $\delta_n > 0$, удовлетворяющей условиям $\delta_n p_n/\sqrt{n} \to 0$, $\sqrt{p_n}/\delta_n \to 0$ при $n \to \infty$, выполняется предельное соотношение*

$$\mathbb{P}\left(\|\alpha^{0,n} - \widehat{\alpha}_n\| \geq \delta_n/\sqrt{n}\right) \to 0,\, n \to \infty.$$

Литература

1. A. Albert, J. Anderson. On the existence of maximum likelihood estimates in logistic regression models // Biometrika. 1984. vol. 71. No. 1. c.1-10.

2. Hua Liang, Pang Du. Maximum likelihood estimation in logistic regression models with a diverging number of covariates // Electronic J. of Statist. 2012. vol. 6. c. 1838-1846.