Выбор λ производился с использованием 3-кратной перекрестной проверки. Преимущество этой техники заключается в том, что она уменьшает переобучение без ограничения подмножества набора данных.

Так как присутствует дисбаланс классов обучающей выборки, то в модели использовались сбалансированные веса для классов, то есть вес каждого класса обратно пропорционален количеству экземпляров класса в выборке.

Следует отметить, что применение точного критерия Фишера и метода LASSO привело к различным наборам информативных признаков.

Литература

1. Афифи А., Эйзен С. *Статистический анализ. Подход с использованием ЭВМ.* М: Мир. 1982. Т. 488.

2. Ranstam J., Cook J.A. *LASSO regression* // Journal of British Surgery. 2018. V. 105. №. 10. P. 1348-1348.

ОЦЕНКА СВЕРХУ ДЛЯ БИНОМИАЛЬНЫХ КОЭФФИЦИЕНТОВ В ФОРМЕ МУАВРА — ЛАПЛАСА

С.В. Агиевич

Белгосуниверситет, НИИ прикладных проблем математики и информатики
Независимости 4, 220030 Минск, Беларусь agievich@bsu.by

Теорема Муавра — Лапласа применительно к симметричному биномиальному распределению может быть записана в виде следующей оценки биномиальных коэффициентов:

$$\binom{n}{k} = \frac{2^n}{\sqrt{\pi n/2}} \exp\left(-\frac{2(k-n/2)^2}{n}\right)(1+O(1/\sqrt{n})).$$

Оценка справедлива при $n \to \infty$ и $|k-n/2| = O(\sqrt{n})$, т. е. в так называемой центральной области изменения параметров.

То, что оценка носит асимптотический характер и справедлива только в центральной области, затрудняет ее применение в ряде случаев. Известны неасимптотические оценки, которые справедливы в более широких областях, например,

$$\left(\frac{n}{k}\right)^k \le \binom{n}{k} \le \left(\frac{en}{k}\right)^k, \quad 1 \le k \le n,$$

или, обозначив $H_2(x) = -x\log_2 x - (1-x)\log_2(1-x)$,

$$\frac{2^{nH_2(k/n)}}{\sqrt{8k(1-k/n)}} \le \binom{n}{k} \le \frac{2^{nH_2(k/n)}}{\sqrt{2\pi k(1-k/n)}}, \quad 1 \le k \le n-1$$

(см. соответственно [2] и [1; глава 10, лемма 7]). Однако эти оценки либо недостаточно точны, либо их форма оказывается недостаточно удобной.

Мы нашли оценку сверху для биномиальных коэффициентов, в которой сохраняется форма Муавра — Лапласа и которая справедлива во всей области изменения параметров.

Теорема. Для натурального n и $k \in \{0,1,\dots,n\}$ справедлива оценка

$$\binom{n}{k} \le \frac{2^n}{\sqrt{\pi n/2}} \exp\left(-\frac{2(k-n/2)^2}{n} + \frac{23}{18n}\right).$$

При построении оценки мы следовали подходу работы [3], в свою очередь основанному на ряде предшествующих работ.

Литература

1. MacWilliams F. J., Sloane N. J. A. *The Theory of Error-Correcting Codes.* 2nd Edition. North-holland Publishing Company, 1978.

2. Odlyzko A. M. *Asymptotic Enumeration Methods.* In: Handbook of Combinatorics. R. L. Graham, M. Groetschel and L. Lovasz eds.. Vol. 2. Elsevier. 1995. P. 1063–1229.

3. Szabados T. *A Simple Wide Range Approximation of Symmetric Binomial Distributions.* 2016. arXiv: 1612.01112 [math.PR].

ОБ АСИМПТОТИЧЕСКОЙ МОЩНОСТИ НЕКОТОРЫХ ТЕСТОВ ЧИСТОЙ СЛУЧАЙНОСТИ ДВОИЧНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

Волошко В.А., Трубей А.И.

Белгосуниверситет, НИИ прикладных проблем математики и информатики
Независимости 4, 220030 Минск, Беларусь
valeravoloshko@yandex.ru, trubeia@mail.ru

Пусть наблюдается двоичная случайная последовательность $x_1, x_2, \ldots, x_n \in \{0,1\}$ длины $n \in \mathbb{N}$. Нулевая гипотеза H_0 состоит в предположении чистой случайности наблюдаемой последовательности, то есть в равномерном распределении ее вероятностей:

$$P\{(x_1, \ldots, x_n) = q\} = 2^{-n}, \ \forall q \in \{0,1\}^n.$$

Альтернатива H_1 предполагает, что наблюдаемая последовательность есть стационарная цепь Маркова некоторого произвольного фиксированного порядка $s \in \mathbb{N}$, для которой переходные вероятности имеют вид

$$P\{x_t = 0 | (x_{t-1}, \ldots, x_{t-s}) = q\} = 1/2 + \delta(q), \ q \in \{0,1\}^s.$$

Будем говорить, что альтернатива H_1 контигуально сближается с нулевой гипотезой H_0, если при $n \to \infty$ имеет место следующая асимптотика:

$$\textstyle\sum_{q \in \{0,1\}^s} \delta^2(q) = \mathcal{O}(n^{-1}).$$

Тесты многомерной дискретной равномерности основаны на статистиках хи-квадрат для L-грам, $L \in \mathbb{N}$, встреченных в наблюдаемой последовательности $\{x_i\}_{i=1}^n$. Базовая статистика хи-квадрат имеет следующий общий вид:

$$S_L^\tau = k^{-1} 2^L \textstyle\sum_{q \in \{0,1\}^L} (f_q - k \cdot 2^{-L})^2,$$

где τ – тип теста ($\tau = \pi$ ($\tau = \nu$) для теста по пересекающимся (по непересекающимся) L-грамам), $k \in \mathbb{N}$ – общее число встреченных L-грам, $f_q : \{0,1\}^L \to \mathbb{Z}$ – число встреченных L-грам, равных $q \in \{0,1\}^L$. Для теста МДРН(L) многомерной дискретной равномерности по непересекающимся L-грамам требуется, чтобы n делилось на L, а встреченные L-грамы – это $k = n/L$ непересекающихся L-блоков, на которые разбивается наблюдаемая последовательность. Для теста МДРП(L) многомерной дискретной равномерности по пересекающимся L-грамам $k = n$, а встреченные L-грамы – это все n подслов $q = (x_{i+1}, \ldots, x_{i+L})$, $i = 1, \ldots, n$, длины L в "закольцованной" наблюдаемой последовательности $x_1, \ldots, x_n, x_1, \ldots, x_n, x_1, \ldots$. Статистика теста МДРН($L$) имеет вид $S = S_L^\nu$. Статистика теста МДРП(L) имеет вид $S = S_L^\pi - S_{L-1}^\pi$. Каждый из тестов принимает гипотезу H_0 при $S \le S_\alpha$, и принимает альтернативу H_1 в противном случае. Здесь $S_\alpha \in \mathbb{R}$ – порог теста, отвечающий вероятности ошибки первого рода $\alpha \in (0,1)$. В докладе для асимптотики контигуального сближения Марковской альтернативы H_1 с нулевой гипотезой H_0 найдены пределы вероятностей ошибки второго рода β для тестов МДРН(L) и МДРП(L): они в явном виде выражены через функцию $\delta(q)$ отклонения переходных вероятностей Марковской альтернативы.