

УДК 519. 2

ДИСКРИМИНАНТНЫЙ АНАЛИЗ ДИСКРЕТНЫХ СЛУЧАЙНЫХ ВЕКТОРОВ

Ю. С. Харин¹⁾, М. Н. Колячко²⁾

^{1), 2)} *Белорусский государственный университет, пр. Независимости, 4,
220030, г. Минск, Беларусь, kharin@bsu.by*

Разработана общая математическая модель дискриминантного анализа дискретно-значных случайных N -векторов из L классов. Построены байесовские и подстановочные байесовские решающие правила для общей модели и пяти основных (стандартных) многомерных дискретных вероятностных распределений вероятностей. Приводятся результаты компьютерных экспериментов.

Ключевые слова: дискретные данные; случайный вектор; дискриминантный анализ.

DISCRIMINANT ANALYSIS OF DISCRETE-VALUED RANDOM VECTORS

Yu. S. Kharin¹⁾, M. N. Kolyachko²⁾

^{1), 2)} *Belarusian State University, Independence av., 4,
220030, Minsk, Belarus, kharin@bsu.by*

General mathematical model for discriminant analysis of discrete-valued random N -vectors is developed. Bayesian and plug-in Bayesian decision rules are constructed for general model and five standard multivariate discrete probability distributions. Results of computer experiments are given.

Keywords: discrete data; random vector; discriminant analysis.

1. Введение

Любое цифровизированное общество порождает множество дискретно-значных данных, описывающих различные объекты или явления стохастической природы. Если каждый исследуемый объект характеризуется N числовыми показателями (признаками), то результатом наблюдения за объектом является дискретный N -мерный вектор

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \in A,$$

принимаящий значения из некоторого дискретного множества A .

Теория дискриминантного анализа (являющегося частью статистического анализа) и соответствующее программное обеспечение глубоко разработаны для «непрерывных» данных, когда A – подмножество евклидова пространства R^N с ненулевой мерой Лебега, а для «дискретного» случая с дискретным множеством наблюдений теория и программное обеспечение начинают формироваться [1 – 6].

Данная статья посвящена разработке методов, алгоритмов и программного обеспечения дискриминантного анализа дискретных случайных векторов. Это исследование выполнено при частичной финансовой поддержке в рамках проекта № 20212422 ГПНИ «Информатика-2025».

2. Математическая модель

Пусть на вероятностном пространстве (Ω, F, P) определен случайный вектор с N компонентами (случайными величинами)

$$\Xi = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_N \end{pmatrix} \in A,$$

с некоторым дискретным N -мерным распределением вероятностей

$$\begin{aligned} P\{\xi_1 = i_1, \xi_2 = i_2, \dots, \xi_N = i_N\} &= P\{\Xi = X\} = p(X; \theta), \\ X \in A, \sum_{X \in A} p(X; \theta) &= 1, \end{aligned} \quad (1)$$

где A – некоторое дискретное множество (конечное, если $2 \leq |A| \leq M < +\infty$, или счетное, если $|A| = +\infty$); $\theta = (\theta_i) \in R^m$ – вектор параметров распределения.

Имеется $2 \leq L < +\infty$ различных классов распределений вероятностей (1): $p^{(1)}(X; \theta^{(1)}), p^{(2)}(X; \theta^{(2)}), \dots, p^{(L)}(X; \theta^{(L)})$.

Задача дискриминантного анализа заключается в построении решающего правила для определения номера класса $1 \leq l \leq L$, к которому относится наблюдение (зарегистрированные данные) Ξ , с наименьшей возможной вероятностью ошибочного решения. При этом возможны два уровня априорной информации: истинные значения параметров $\{\theta^{(l)}\}$ известны априорно (получены в результате предыдущих исследований;

$\{\theta^{(l)}\}$ – априорно не известны, но наблюдается классифицированная обучающая выборка из всех L классов:

$$S^{(l)} = \{X^{(l,1)}, X^{(l,2)}, \dots, X^{(l,M_l)}\} \in A^{M_l}, l=1, 2, \dots, L, \quad (2)$$

где $S^{(l)}$ – случайная выборка объема $M_l > m$ из l -го класса.

Будем использовать байесовский подход к решению задачи дискриминантного анализа дискретных многомерных данных [1]. Обозначим: $v \in \{1, 2, \dots, L\}$ – случайный номер класса, $\pi_l = P\{v=l\}$ – априорная вероятность l -го класса, $\Xi^{(l)} = (\xi_i^{(l)}) \in A$ – случайный вектор наблюдений из l -го класса с дискретным распределением $p^{(l)}(X; \theta^{(l)})$.

Будем рассматривать два типа решающих правил [2]. Нерандомизированное решающее правило для сформулированной выше задачи дискриминантного анализа имеет следующий общий вид:

$$d = d(X) = \begin{cases} 1, & X \in B_1, \\ 2, & X \in B_2, \\ \dots & \\ L, & X \in B_L, \end{cases} \quad (3)$$

где $\{B_l\}$ – разбиение множества (пространства) наблюдений A на L подмножеств:

$$A = \bigcup_{l=1}^L B_l, B_k \cap B_l = \emptyset, k \neq l, \quad (4)$$

$d \in D = \{1, 2, \dots, L\}$ – решение (номер класса) для наблюдения X , D – пространство решений. Таким образом, нерандомизированное решающее правило (3) задается разбиением (4).

Рандомизированное решающее правило в отличие от (3), (4) зависит от «случая» $\omega \in \Omega$ и имеет следующий общий вид $d = d(X, \omega)$:

$$P\{d(X, \omega) = l | X\} = \varphi_l(X), l \in D, X \in A, \quad (5)$$

где $\{\varphi_l(X)\}$ – так называемые критические функции:

$$0 \leq \varphi_l(X) \leq 1, \sum_{l=1}^L \varphi_l(X) \equiv 1, \quad (6)$$

полностью определяющие решающее правило (5). Таким образом, при наличии наблюдения X решение d о его принадлежности к одному из L классов определяется по жребию с распределением вероятностей $\{\phi_l(X)\}$.

Вероятность ошибочного решения (риск) для нерандомизированного решающего правила (3), (4) имеет вид:

$$r = r(d) ::= P\{d(\Xi^{(v)}) \neq v\}, \quad (7)$$

а для рандомизированного решающего правила (5), (6):

$$r = r(d) ::= P\{d(\Xi^{(v)}, \omega) \neq v\}.$$

Критерий оптимальности решающего правила – минимум вероятности ошибки (риска):

$$r(d) \rightarrow \min_{d(\cdot)}. \quad (8)$$

Оптимальное решающее правило $d_0(\cdot)$, удовлетворяющее (8), принято называть байесовским решающим правилом (БРП) [1].

3. БРП для общей модели дискретных случайных векторов

Рассмотрим вначале случай нерандомизированного решающего правила (3), (4). Обозначим $\mathbf{1}\{C\}$ – индикатор события C .

Теорема 1. Если верна определенная выше вероятностная модель наблюдения (1), и истинные значения параметров $\{\theta^{0(l)}\}$ априорно известны, то оптимальным решающим правилом, минимизирующим вероятность ошибки принятия решения, является байесовское решающее правило (БРП):

$$d = d_0(X) = \arg \max_{1 \leq l \leq L} (\pi_l p^{(l)}(X; \theta^{0(l)})), \quad X \in A. \quad (9)$$

При этом минимальная вероятность ошибки, достигаемая БРП, равна:

$$r_0 = 1 - \sum_{X \in A} \max_{1 \leq l \leq L} (\pi_l p^{(l)}(X; \theta^{0(l)})). \quad (10)$$

Доказательство. Вычислим вероятность для РП (3), (4), используя математическую модель наблюдений (1), формулу полной вероятности и условие нормировки:

$$r = r(d) := P\{d(\Xi^{(v)}) \neq v\} = \sum_{l=1}^L P\{v = l\} \cdot P\{d(\Xi^{(l)}) \neq l\} =$$

$$\sum_{l=1}^L \pi_l \cdot \left(1 - P\{d(\Xi^{(l)}) = l\}\right) = 1 - \sum_{X \in A} \left(\sum_{l=1}^L \pi_l p^{(l)}(X; \theta^{0(l)}) \mathbf{1}\{X \in B_l\} \right). \quad (11)$$

Решим задачу минимизации вероятности ошибки (7), которая в силу (11) эквивалентна поточечной ($X \in A$) максимизации выражения:

$$\sum_{l=1}^L \pi_l p^{(l)}(X; \theta^{0(l)}) \mathbf{1}\{X \in B_l\} \rightarrow \max_{\{B_l\}}. \quad (12)$$

Решение задачи (12) очевидно: наблюдение X необходимо включить в B_{l^*} , где $l^* = \arg \max_{1 \leq l \leq L} \pi_l p^{(l)}(X; \theta^{0(l)})$, что приводит к оптимальному БРП (9). $\square$

Подставляя (9) в (11), получаем (10).

Замечание. Отметим, что если в точке ($X \in A$) максимум в правой части (9) достигается в $2 \leq K \leq L$ точках $l^* \in \{l_1, l_2, \dots, l_K\} \subseteq D$, то решающее правило (9) эквивалентно рандомизированному решающему правилу (5), (6): в точке X решение может быть отнесено к классам номер $l_1, l_2, \dots, l_K$ с любыми вероятностями $\phi_{l_1}(X), \phi_{l_2}(X), \dots, \phi_{l_K}(X) \in [0, 1]$, сумма которых равна единице.

Теорема 2. Если для модели (1) $\{p^{(l)}(\cdot)\}$ удовлетворяют условиям регулярности, истинные значения параметров $\{\theta^{0(l)}\}$ априорно неизвестны и $\{\hat{\theta}^{(l)}\}$ – некоторые состоятельные (при $\min_l M_l \rightarrow +\infty$) оценки этих параметров по классифицированной обучающей выборке (2), то подстановочное байесовское решающее правило (ПБРП):

$$d = \hat{d}_0(X) := \arg \max_{1 \leq l \leq L} \left(\pi_l p^{(l)}(X; \hat{\theta}^{(l)}) \right), \quad X \in A, \quad (13)$$

имеет риск, сходящийся к r_0 :

$$r(\hat{d}_0) \rightarrow r_0. \quad (14)$$

Доказательство. В условиях регулярности [2] при $\min_l M_l \rightarrow +\infty$

$$\hat{\theta}^{(l)} \xrightarrow{P} \theta^{0(l)}, \quad l \in D.$$

В силу конечности L, X , условий регулярности и (10)

$$r(\hat{d}_0) = 1 - \sum_{X \in A} \pi_l E \left\{ \max p^{(l)}(X; \hat{\theta}^{(l)}) \right\} \rightarrow r_0.$$

□

Построим байесовское решающее правило для основных (стандартных) дискретных многомерных распределений вероятностей [3, 5].

4. БРП для полиномиального распределения

Полиномиальное распределение – совместное распределение случайных величин $\xi_1, \dots, \xi_N$, принимающих целые неотрицательные значения $x_1, \dots, x_N$, удовлетворяющие условию $\sum_{i=1}^N x_i = n$, с вероятностями [3, 5]:

$$p(x_1, x_2, \dots, x_N) = \frac{n!}{x_1! x_2! \dots x_N!} \prod_{i=1}^N p_i^{x_i}, \quad \sum_{i=1}^N p_i = 1, \quad x = (x_1, x_2, \dots, x_N) -$$

N -мерный вектор, $x \in A = \{(x_1, x_2, \dots, x_N) : 0 \leq x_i \leq n, x_1 + x_2 + \dots + x_N = n\}$.

Имеется $2 \leq L < +\infty$ различных классов распределений вероятностей

$$p^{(l)}(x_1, x_2, \dots, x_N) = \frac{n!}{x_1! x_2! \dots x_N!} \prod_{i=1}^N (p_i^{(l)})^{x_i}, \quad (15)$$

$\sum_{i=1}^N p_i^{(l)} = 1; \pi_l = P\{v = l\}$, где $v \in D$ – случайный номер класса, $l \in D$,

$D = \{1, 2, \dots, L\}$.

Построим байесовское решающее правило для классов, описываемых полиномиальным распределением вероятностей, на основе теоремы 1.

Следствие 1. Для полиномиальной модели (15) БРП имеет вид:

$$d_0(x) = \arg \max_{l \in D} \left(\ln \pi_l + \sum_{i=1}^N x_i \ln p_i^{(l)} \right); \quad (16)$$

вероятность ошибки этого решающего правила

$$r_0 = 1 - \sum_{x \in A} \frac{n!}{x_1! x_2! \dots x_N!} \max_{l \in D} \left(\pi_l \prod_{i=1}^N (p_i^{(l)})^{x_i} \right).$$

Рассмотрим отдельно случай двух классов ($L = 2$). Определим функцию

$$B(x) = \ln \frac{\pi_1}{\pi_2} + \sum_{i=1}^N x_i \ln \frac{p_i^{(1)}}{p_i^{(2)}}.$$

Тогда БРП (16) можно записать в следующем виде:

$$d_0(x) = \begin{cases} 1, & B(x) \geq 0, \\ 2, & B(x) < 0. \end{cases}$$

Вероятность ошибки построенного правила

$$r_0 = 1 - \sum_{x \in A} \frac{n!}{x_1! x_2! \dots x_N!} \max \left(\pi_1 \cdot \prod_{i=1}^N (p_i^{(1)})^{x_i}, \pi_2 \cdot \prod_{i=1}^N (p_i^{(2)})^{x_i} \right).$$

Подстановочное БРП (13) имеет вид:

$$\hat{d}_0(x) = \arg \max_{l \in D} \left(\ln \hat{\pi}_l + \sum_{i=1}^N x_i \ln \hat{p}_i^{(l)} \right)$$

и получается подстановкой в (16) следующих оценок максимального правдоподобия (ОМП):

$$\hat{\pi}_l = \frac{M_l}{\sum_{i=1}^L M_i}, \quad \hat{p}_i^{(l)} = \frac{\sum_{j=1}^{M_l} x_i^{(l,j)}}{nM_l}, \quad l \in D, \quad i = 1, 2, \dots, N.$$

Рассмотрим следующий численный пример: $L = 2$; $N = 10$; $\pi_1 = 0.5$, $\pi_2 = 1 - \pi_1 = 0.5$, $M = M_1 + M_2$ – суммарный объем обучающей выборки, $p^{(1)} = (0.06, 0.08, 0.03, 0.14, 0.04, 0.7, 0.14, 0.12, 0.13, 0.19)$, $p^{(2)} = (0.06, 0.03, 0.15, 0.03, 0.09, 0.18, 0.11, 0.09, 0.04, 0.22)$. Число экзаменационных реализаций $M_1^* = M_2^* = 50$.

График зависимости оценки $\hat{r}$ вероятности ошибки классификации от объема M обучающей выборки представлен на рис. 1.

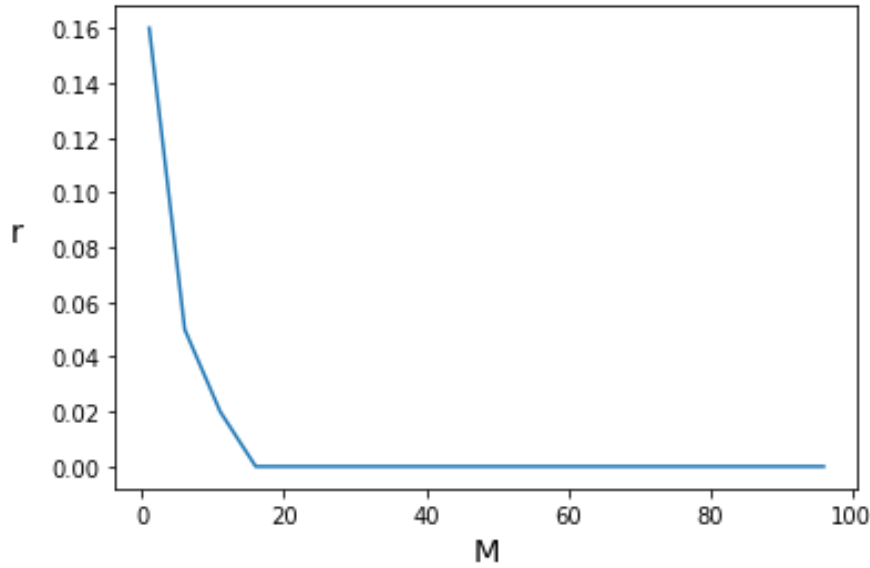


Рис.1. Зависимость оценки $\hat{r}$ вероятности ошибки классификации от объема M обучающей выборки из полиномиального распределения

5. БРП для многомерного биномиального распределения

Многомерное биномиальное распределение вероятностей – совместное распределение случайных величин $\xi_1, \dots, \xi_N$, принимающих значения $x_1, \dots, x_N, x_i \in \{0, 1, \dots, n\}$, с вероятностями [3, 5]:

$$p(x) = p(x_1, x_2, \dots, x_N) = \prod_{i=1}^N \binom{n}{x_i} p^{x_i} q^{(n-x_i)}, x \in A,$$

где $p + q = 1, 0 \leq p \leq 1$; $x = (x_1, x_2, \dots, x_N)$ – N -мерный вектор, $x \in A = \{0, 1, \dots, n\}^N$.

Имеется $2 \leq L < +\infty$ различных классов распределений вероятностей:

$$p^{(l)}(x_1, x_2, \dots, x_N) = \prod_{i=1}^N \binom{n}{x_i} (p^{(l)})^{x_i} (q^{(l)})^{(n-x_i)}, q^{(l)} = 1 - p^{(l)}, 0 \leq p^{(l)} \leq 1. \quad (17)$$

Следствие 2. Байесовское решающее правило для L классов многомерного биномиального распределения вероятностей имеет вид:

$$d_0(x) = \arg \max_{l \in D} \left(\ln \pi_l + v \ln \frac{p^{(l)}}{1 - p^{(l)}} + Nn \ln(1 - p^{(l)}) \right), v = \sum_{i=1}^N x_i, \quad (18)$$

а вероятность ошибки

$$r_0 = 1 - \sum_{x \in A} \max_{l \in D} \left(\pi_l \cdot \prod_{i=1}^N \binom{n}{x_i} (p^{(l)})^{x_i} (q^{(l)})^{(n-x_i)} \right).$$

В случае двух классов ($L = 2$) БРП (18) можно записать в следующем виде:

$$d_0(x) = \begin{cases} 1, & B(x) \geq 0, \\ 2, & B(x) < 0, \end{cases}$$

$$B(x) = \ln \frac{\pi_1}{\pi_2} + v \ln \frac{p^{(1)} \cdot (1 - p^{(2)})}{p^{(2)} \cdot (1 - p^{(1)})}, \quad v = \sum_{i=1}^N x_i.$$

Вероятность ошибки для этого БРП:

$$r_0 = 1 - \sum_{x \in A} \max \left(\pi_1 \cdot \prod_{i=1}^N \binom{n}{x_i} (p^{(1)})^{x_i} (q^{(1)})^{(n-x_i)}, \right. \\ \left. \pi_2 \cdot \prod_{i=1}^N \binom{n}{x_i} (p^{(2)})^{x_i} (q^{(2)})^{(n-x_i)} \right).$$

Подстановочное БРП принимает вид:

$$\hat{d}_0(X) = \arg \max_{l \in D} \left(\ln \hat{\pi}_l + v \ln \frac{\hat{p}^{(l)}}{1 - \hat{p}^{(l)}} + N \ln (1 - \hat{p}^{(l)}) \right), \quad v = \sum_{i=1}^N x_i,$$

где $\hat{\pi}_l = M_l / \sum_{i=1}^N M_i$; $\hat{p}^{(l)}$ – оценка максимального правдоподобия параметра распределения $p^{(l)}$:

$$\hat{p}^{(l)} = \frac{v^{(l)}}{nM_lN}, \quad v^{(l)} = \sum_{j=1}^{M_l} \sum_{i=1}^N x_i^{(l,j)}, \quad l \in D.$$

Оценка $\hat{p}^{(l)}$ является несмещенной и состоятельной.

Рассмотрим следующий численный пример: $L = 2$; $N = 10$; $p^{(1)} = 0.76$, $p^{(2)} = 0.45$; $\pi_1 = 0.5$, $\pi_2 = 1 - \pi_1 = 0.5$. Число экзаменационных реализаций $M_1^* = M_2^* = 50$.

График зависимости оценки $\hat{r}$ вероятности ошибки классификации от объема $M = M_1 + M_2$ обучающей выборки представлен на рис. 2.

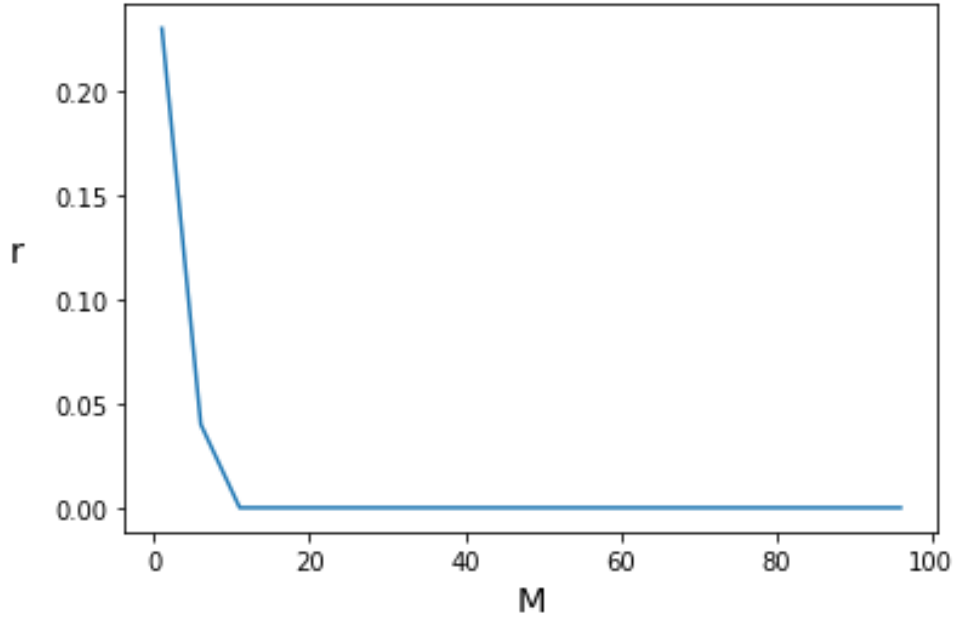


Рис. 2. Зависимость оценки $\hat{r}$ вероятности ошибки классификации от объема M обучающей выборки из многомерного биномиального распределения

6. БРП для многомерного распределения Пуассона

Многомерное распределение Пуассона – совместное распределение случайных величин $\xi_1, \dots, \xi_N$, принимающих значения $x_1, \dots, x_N$, $x_i \in \{0, 1, 2, \dots\}$ с вероятностями [5]:

$$p(x) = p(x_1, x_2, \dots, x_N) = \prod_{i=1}^N e^{-\theta_i} \cdot \frac{\theta_i^{x_i}}{x_i!},$$

где $x = (x_1, x_2, \dots, x_N)$ – случайный N -мерный вектор,

$x \in A = \{(x_1, x_2, \dots, x_N) : x_i \in \{0, 1, 2, \dots\}\}$; $\theta_i \geq 0$ – параметр распределения.

Имеется $2 \leq L < +\infty$ различных классов распределений вероятностей:

$$p_l(x) = p^{(l)}(x_1, x_2, \dots, x_N) = \prod_{i=1}^N e^{-\theta_i^{(l)}} \cdot \frac{(\theta_i^{(l)})^{x_i}}{x_i!}, \quad l \in D. \quad (19)$$

Следствие 3. Байесовское решающее правило для L классов многомерного распределения Пуассона имеет вид:

$$d_0(x) = \arg \max_{l \in D} \left(\ln \pi_l + \sum_{i=1}^N (-\theta_i^{(l)} + x_i \ln \theta_i^{(l)}) \right), \quad (20)$$

а вероятность ошибки решающего правила (20)

$$r_0 = 1 - \sum_{x \in A} \max_{l \in D} \left(\pi_l \cdot \prod_{i=1}^N e^{-\theta_i^{(l)}} \cdot \frac{(\theta_i^{(l)})^{x_i}}{x_i!} \right).$$

В случае двух классов ($L = 2$) БРП можно записать в следующем виде:

$$d_0(x) = \begin{cases} 1, & B(x) \geq 0, \\ 2, & B(x) < 0, \end{cases}$$

где

$$B(x) = \ln \frac{\pi_1}{\pi_2} + \sum_{i=1}^N \left(-(\theta_i^{(1)} - \theta_i^{(2)}) + x_i \ln \frac{\theta_i^{(1)}}{\theta_i^{(2)}} \right),$$

а вероятность ошибки построенного решающего правила

$$r_0 = 1 - \sum_{x \in A} \max \left(\pi_1 \cdot \prod_{i=1}^N e^{-\theta_i^{(1)}} \cdot \frac{(\theta_i^{(1)})^{x_i}}{x_i!}, \pi_2 \cdot \prod_{i=1}^N e^{-\theta_i^{(2)}} \cdot \frac{(\theta_i^{(2)})^{x_i}}{x_i!} \right).$$

Подстановочное БРП принимает вид:

$$\hat{d}_0(x) = \arg \max_{l \in D} \left(\ln \hat{\pi}_l + \sum_{i=1}^N \left(-\hat{\theta}_i^{(l)} + x_i \ln \hat{\theta}_i^{(l)} \right) \right),$$

где $\hat{\pi}_l = M_l / \sum_{i=1}^L M_i$ – оценка априорной вероятности, а $\hat{\theta}^{(l)} = (\hat{\theta}_1^{(l)}, \dots, \hat{\theta}_n^{(l)})$ – оценки максимального правдоподобия параметров распределения $p^{(l)}(x)$:

$$\hat{\theta}_i^{(l)} = \frac{\sum_{j=1}^{M_l} x_{ji}^{(l,j)}}{M}, \quad i = 1, \dots, N; \quad M = \sum_{k=1}^L M_k.$$

Рассмотрим следующий численный пример: $L = 2$; $N = 10$; $\theta^{(1)} = (7, 2, 6, 5, 2, 6, 7, 3, 6, 1)$, $\theta^{(2)} = (8, 2, 8, 9, 4, 5, 2, 5, 7, 5)$; $\pi_1 = 0.5$, $\pi_2 = 1 - \pi_1 = 0.5$. Число экзаменационных реализаций $M_1^* = M_2^* = 50$.

График зависимости оценки $\hat{r}$ вероятности ошибки классификации r от объема $M = M_1 + M_2$ обучающей выборки представлен на рис. 3.

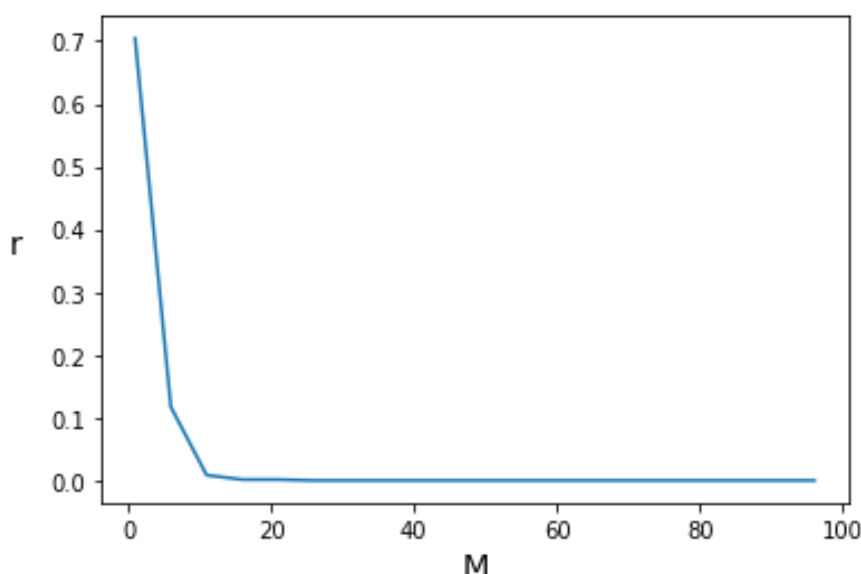


Рис.3. Зависимость оценки $\hat{r}$ вероятности ошибки классификации r от объема M обучающей выборки из многомерного распределения Пуассона

7. БРП для многомерного логарифмического распределении

Многомерное логарифмическое распределение – совместное распределение случайных величин $\xi_1, \dots, \xi_N$, принимающих значения $x_1, \dots, x_N$, $x_i \in \{0, 1, \dots\}$, с вероятностями [5]:

$$p(x_1, x_2, \dots, x_N) = \frac{\left(\sum_{i=1}^N x_i - 1\right)!}{\left(\sum_{i=1}^N x_i!\right) \cdot \left(-\ln\left(1 - \sum_{i=1}^N \theta_i\right)\right)} \cdot \prod_{i=1}^N \theta_i^{x_i},$$

где $x = (x_1, x_2, \dots, x_N)$ – N -мерный вектор,

$x \in A = \{(x_1, x_2, \dots, x_N) : x_i \in \{0, 1, 2, \dots\}\}$; $\theta = (\theta_i)$ – вектор параметров,

$\theta_i > 0$, $\sum_{i=1}^N \theta_i < 1$.

Имеется $2 \leq L < +\infty$ различных классов распределений вероятностей

$$p^{(l)}(x_1, x_2, \dots, x_N) = \frac{\left(\sum_{i=1}^N x_i - 1\right)!}{\left(\sum_{i=1}^N x_i!\right) \cdot \left(-\ln\left(1 - \sum_{i=1}^N \theta_i^{(l)}\right)\right)} \cdot \prod_{i=1}^N (\theta_i^{(l)})^{x_i}.$$

Следствие 4. Байесовское решающее правило для L классов многомерного логарифмического распределения вероятностей имеет вид:

$$d_0(x) = \arg \max_{l \in D} \left(\ln \pi_l - \ln \left(-\ln \left(1 - \sum_{i=1}^N \theta_i^{(l)} \right) \right) + \sum_{i=1}^N x_i \ln \theta_i^{(l)} \right), \quad x \in A,$$

а вероятность ошибки этого байесовского решающего правила:

$$r_0 = 1 - \sum_{x \in A} \frac{\left(\sum_{i=1}^N x_i - 1\right)!}{\left(\prod_{i=1}^N x_i!\right)} \max_{l \in D} \frac{-\pi_l \cdot \prod_{i=1}^N (\theta_i^{(l)})^{x_i}}{\ln \left(1 - \sum_{i=1}^N \theta_i^{(l)} \right)}.$$

В случае двух классов ($L = 2$) БРП можно записать в следующем виде:

$$d_0(x) = \begin{cases} 1, & B(x) \geq 0, \\ 2, & B(x) < 0, \end{cases}$$

где

$$B(x) = \ln \frac{\pi_1}{\pi_2} - \ln \frac{\ln \left(1 - \sum_{i=1}^N \theta_i^{(1)} \right)}{\ln \left(1 - \sum_{i=1}^N \theta_i^{(2)} \right)} + \sum_{i=1}^N x_i \ln \frac{\theta_i^{(1)}}{\theta_i^{(2)}}.$$

Вероятность ошибки этого решающего правила

$$r_0 = 1 - \sum_{x \in A} \frac{\left(\sum_{i=1}^N x_i - 1\right)!}{\left(\prod_{i=1}^N x_i!\right)} \max \left(\frac{-\pi_1 \cdot \prod_{i=1}^N (\theta_i^{(1)})^{x_i}}{\ln \left(1 - \sum_{i=1}^N \theta_i^{(1)} \right)}, \frac{-\pi_2 \cdot \prod_{i=1}^N (\theta_i^{(2)})^{x_i}}{\ln \left(1 - \sum_{i=1}^N \theta_i^{(2)} \right)} \right).$$

Подстановочное БРП представимо в виде:

$$\hat{d}_0(x) = \arg \max_{l \in D} \left(\ln \hat{\pi}_l - \ln \left(\ln \left(1 - \sum_{i=1}^N \hat{\theta}_i^{(l)} \right) \right) + \sum_{i=1}^N x_i \ln \hat{\theta}_i^{(l)} \right),$$

где $\hat{\pi}_l = M_l / \sum_{i=1}^L M_i$; $\hat{\theta}^{(l)} = (\hat{\theta}_1^{(l)}, \dots, \hat{\theta}_N^{(l)})$ – оценка максимального правдоподобия параметров распределения $p^{(l)}(x)$, которая является решением системы N уравнений:

$$\frac{1}{M_l} \sum_{j=1}^{M_l} x_i^{(l,j)} = \frac{\hat{\theta}_i^{(l)}}{\left(1 - \sum_{k=1}^N \hat{\theta}_k^{(l)}\right)}, \quad i=1, \dots, N.$$

Рассмотрим следующий численный пример: $L=2$; $N=10$; $\theta^{(1)} = (0.05, 0.004, 0.15, 0.08, 0.12, 0.15, 0.14, 0.02, 0.06, 0.02)$, $\theta^{(2)} = (0.04, 0.03, 0.15, 0.03, 0.09, 0.18, 0.11, 0.09, 0.04, 0.02)$; $\pi_1 = 0.5$, $\pi_2 = 1 - \pi_1 = 0.5$. Число экзаменационных реализаций $M_1^* = M_2^* = 50$.

График зависимости оценки $\hat{r}$ вероятности ошибки классификации r от объема $M = M_1 + M_2$ обучающей выборки представлен на рис. 4.

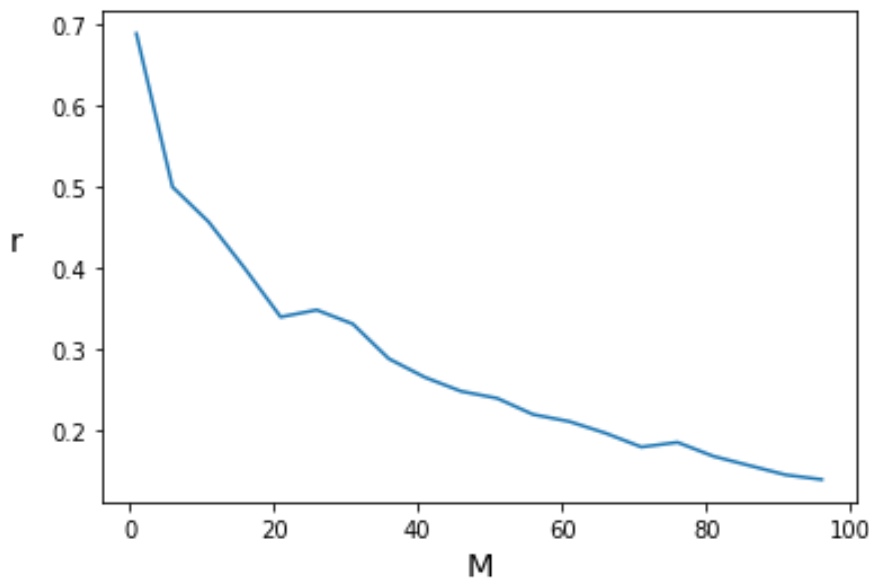


Рис.4. Зависимость оценки $\hat{r}$ вероятности ошибки классификации r от объема M обучающей выборки из многомерного логарифмического распределения

8. БРП для многомерного гипергеометрического распределения

Многомерное гипергеометрическое распределение – совместное распределение случайных величин $\xi_1, \dots, \xi_N$, принимающих значения $x_1, \dots, x_N$, $x_i \in A = \{0, 1, \dots, n\}$ с вероятностями [5]:

$$p(x_1, x_2, \dots, x_N) = \frac{\prod_{i=1}^N \binom{m_i}{x_i}}{\binom{m}{n}},$$

где $\sum_{i=1}^N m_i = m$; $\sum_{i=1}^N x_i = n$; $x = (x_1, x_2, \dots, x_N)$ – N -мерный вектор, $x \in A = \{(x_1, x_2, \dots, x_N): 0 \leq x_i \leq n, x_1 + x_2 + \dots + x_N = n\}$.

Имеется $2 \leq L < +\infty$ различных классов распределений вероятностей:

$$p^{(l)}(x) = p^{(l)}(x_1, x_2, \dots, x_N) = \left(\prod_{i=1}^N \binom{m_i^{(l)}}{x_i} \right) / \binom{m^{(l)}}{n}, \quad x \in A.$$

Следствие 5. Байесовское решающее правило для L классов многомерного гипергеометрического распределения вероятностей имеет вид:

$$d_0(x) = \arg \max_{l \in D} \left(\ln \pi_l + \sum_{i=1}^N \ln \binom{m_i^{(l)}}{x_i} - \ln \binom{m^{(l)}}{n} \right),$$

а его вероятность ошибки

$$r_0 = 1 - \sum_{x \in A} \max_{l \in D} \left(\pi_l \cdot \left(\prod_{i=1}^N \binom{m_i^{(l)}}{x_i} \right) / \binom{m^{(l)}}{n} \right).$$

В случае двух классов ($L = 2$) БРП можно записать в следующем виде:

$$d_0(x) = \begin{cases} 1, & B(x) \geq 0, \\ 2, & B(x) < 0, \end{cases}$$

$$B(x) = \left(\pi_1 \prod_{i=1}^n \binom{m_i^{(1)}}{x_i} \right) / \binom{m^{(1)}}{n} - \left(\pi_2 \prod_{i=1}^n \binom{m_i^{(2)}}{x_i} \right) / \binom{m^{(2)}}{n},$$

а его вероятность ошибки

$$r_0 = 1 - \sum_{x \in A} \max \left(\pi_1 \left(\prod_{i=1}^n \binom{m_i^{(1)}}{x_i} \right) / \binom{m^{(1)}}{n}, \pi_2 \left(\prod_{i=1}^n \binom{m_i^{(2)}}{x_i} \right) / \binom{m^{(2)}}{n} \right).$$

Подстановочное БРП принимает вид:

$$\hat{d}_0(x) = \arg \max_{l \in D} \left(\hat{\pi}_l \cdot \left(\prod_{i=1}^n \binom{\hat{m}_i^{(l)}}{\hat{x}_i} \right) / \binom{\hat{m}^{(l)}}{n} \right),$$

где $\hat{\pi}_l = M_l / \sum_{i=1}^L M_i$; $\hat{m}^{(l)} = (\hat{m}_1^{(l)}, \dots, \hat{m}_n^{(l)})$ – оценки параметров распределения $p^{(l)}(x)$, вычисленные по методу моментов в предположении, что $m^{(l)}$ – известно априорно:

$$\hat{m}_i^{(l)} = \frac{m^{(l)}}{n} \sum_{j=1}^{M_l} x_i^{(l,j)}, \quad i=1, \dots, n; \quad l \in D.$$

Приведем прикладной пример из области контроля качества, когда размер популяции $m^{(l)}$ известен, но количество объектов $m_i^{(l)}$ типа i неизвестно. Такая ситуация возникает, если у нас есть партия из $m^{(l)}$ произведенных изделий, содержащая неизвестное количество $m_i^{(l)}$ дефектных изделий i -типа. Было бы слишком дорого тестировать все $m^{(l)}$ элементов (возможно, даже разрушительно), поэтому мы могли бы вместо этого выбрать $n < m$ элементов случайным образом и протестировать их.

Рассмотрим следующий численный пример: $L = 2$; $N = 10$; $n = 10$; $m^{(1)} = (5, 1, 6, 2, 5, 7, 6, 2, 5, 7)$, $m^{(2)} = (6, 8, 9, 6, 4, 7, 2, 3, 9, 9)$; $\pi_1 = 0.5$, $\pi_2 = 1 - \pi_1 = 0.5$. Число экзаменационных реализаций $M_1^* = M_2^* = 50$.

График зависимости оценки $\hat{r}$ вероятности ошибки классификации r от объема $M = M_1 + M_2$ обучающей выборки представлен на рис. 5.

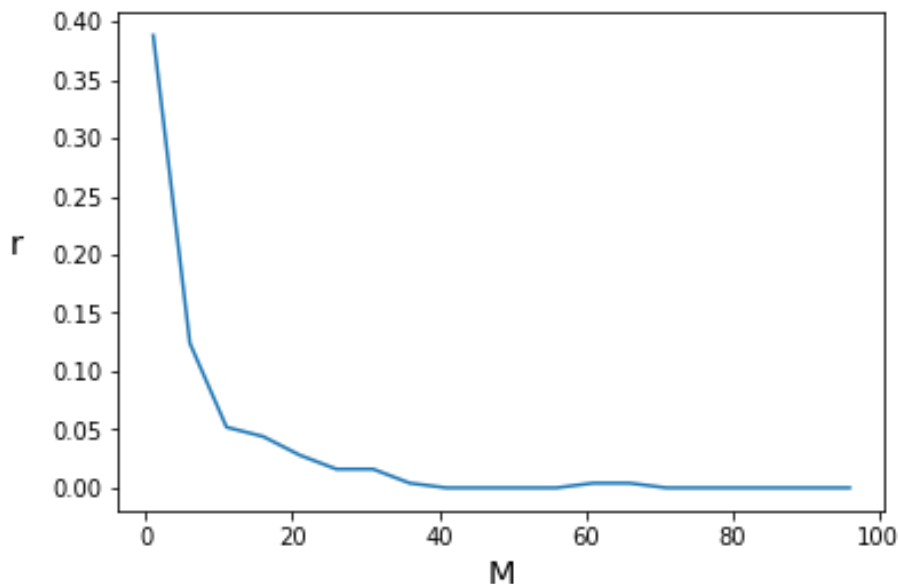


Рис.5. Зависимость оценки $\hat{r}$ вероятности ошибки классификации r от объема M обучающей выборки из многомерного гипергеометрического распределения

9. Заключение

В статье получены следующие основные результаты:

- разработана общая математическая модель дискриминантного анализа дискретно-значных случайных векторов из $L \geq 2$ классов, и в рамках этой модели построены байесовское решающее правило (БРП), минимизирующее вероятность ошибки, когда условные распределения вероятностей априорно известны, и подстановочное байесовское решающее правило (ПБРП) когда параметры условных распределений вероятностей оцениваются по обучающей выборке;
- построены БРП и ПБРП для пяти основных семейств (стандартных) многомерных дискретных распределений вероятностей;
- алгоритмы БРП и ПБРП реализованы в виде Python-библиотеки программ, и проведены серии компьютерных экспериментов, иллюстрирующих теоретические результаты.

Библиографические ссылки

1. Харин Ю. С. Оптимальность и робастность в статистическом прогнозировании. Минск: БГУ, 2008. 263 с.
2. Харин Ю. С., Зуев Н. М., Жук Е. Е. Теория вероятностей, математическая и прикладная статистика. Минск: БГУ, 2011. 463 с.
3. Evans M., Hastings N., Peacock B. Statistical Distribution. N.Y.: Wiley, 2000. Ch. 4. P. 31-33.
4. Holgate P. Miscellanea: Estimation for the Bivariate Poisson Distribution // Biometrika. 1964. Vol. 51. P. 241-245.
5. Johnson N. L., Kotz S., Balakrishnan N. Discrete Multivariate Distributions. N.Y.: Wiley, 1997.
6. Huiming Zhang, Yunxiao Liu, Li Bo. Notes on discrete compound Poisson model with applications to risk theory // Insurance: Mathematics and Economics. 2014. Vol. 59. P. 325-336.