ПРОВЕРКА СЛОЖНЫХ ГИПОТЕЗ ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ ПРИ НАЛИЧИИ КЛАССИФИКАЦИИ НАБЛЮДЕНИЙ

Е. С. Агеева

ВВЕДЕНИЕ

На практике достаточно часто вместо точного значения переменной мы наблюдаем только интервал, в который оно попадает. Такое искажение данных будем называть классификацией. Классификация является одним из случаев группированных данных [1]. Эта работа посвящена регрессионной модели при наличии классификации зависимой переменной. Состоятельность оценок максимального правдоподобия исследована в [4].

1. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Рассмотрим модель нелинейной множественной регрессии [2]:

$$Y_t = F(X_t;\theta^0) + \xi_t,\ t = 1,...,n, \tag{1}$$

где n объём выборки; $F(\cdot): \mathrm{X} \times \Theta \to R^1$ – известная с точностью до векторного параметра функция регрессии; $\theta^0 = (\theta_1^0,...,\theta_m^0)^T \in \Theta \subseteq R^m$ – неизвестный истинный вектор-столбец параметров функции регрессии; $X_t = (X_t^1,...,X_t^N)^T \in \mathrm{X} \subseteq R^N$ – наблюдаемый неслучайный вектор-столбец регрессоров; $Y_t \in R^1$ – зависимая переменная; $\xi_t \in R^1$ – ненаблюдаемая случайная величина ошибок с нормальным распределением вероятностей $\boldsymbol{N}(0,\sigma^2)$. Предполагается, что $\{\xi_t\}_{t=1}^n$ независимы в совокупности.

Пусть заданы K непересекающихся интервалов ($K \geq 2$):

$$A_k = (a_{k-1}, a_k],\ k=1,\ldots,K,\ a_0 = -\infty,\ a_K = +\infty. \tag{2}$$

Эта система борелевских множеств задаёт классификацию Y_t:

$$Y_t \text{ относится к классу } \Omega_{\nu_t}, \text{ если } Y_t \in A_{\nu_t},\ \nu_t \in \{1,...,K\}. \tag{3}$$

Вместо точных значений зависимой переменной $Y_1,...,Y_n$ наблюдаются лишь соответствующие номера классов $\nu_1,...,\nu_n \in \{1,...,K\}$. Задача заключается в том, чтобы, зная разбиение на интервалы $A_1,...,A_K$, классифицированные наблюдения $\nu_1,...,\nu_n$ и значения $X_1,...,X_n$, построить статистические тесты для проверки гипотез о значении неизвестного вектора параметров $\gamma^0 = (\gamma_1^0,...,\gamma_{m+1}^0)^T = ((\theta^0)^T,(\sigma^0)^2)^T$.

Дискретные случайные величины $\{\nu_t\}_{t=1}^n$ связаны с Y_t стохастической зависимостью, порождаемой (1)–(3):

$$P_{X_t,\gamma}\{\nu_t \in k\} = P_{X_t}(\nu_t;\gamma) = \Phi\left(\frac{a_k - F(X_t;\theta)}{\sigma}\right) - \Phi\left(\frac{a_{k-1} - F(X_t;\theta)}{\sigma}\right),\ k=1,\ldots,K,$$

где $\Phi(\cdot)$ – функция распределения вероятностей гауссовской стандартной случайной величины. Введем обозначение: $l(\gamma) = \sum_{t=1}^n \ln P_{X_t}(\nu_t;\gamma)$. Информационная матрица Фишера для выборки $\{\nu_t\}_{t=1}^n$ примет вид:

$$\Gamma_n(\gamma) = \left(\sum_{t=1}^n \sum_{k=1}^K \frac{\partial \ln P_{X_t}(k;\gamma)}{\partial \gamma_i} \frac{\partial \ln P_{X_t}(k;\gamma)}{\partial \gamma_j}\right)_{i,j=1}^n.$$

Теорема 1. Пусть ОМП $\hat{\gamma}$ является состоятельной оценкой вектора параметров γ^0; для любого $\theta \in R^m$ функции $F(X;\theta)$, $\frac{\partial F(X;\theta)}{\partial \theta_i}$, $\frac{\partial^2 F(X;\theta)}{\partial \theta_i \partial \theta_j}$, $\frac{\partial^3 F(X;\theta)}{\partial \theta_i \partial \theta_j \partial \theta_s}$, i, j, s=1,…,m, ограничены на $X \subseteq R^N$; $\frac{1}{n}\Gamma_n(\gamma^0) \succ 0$, $\underline{\lim}_{n\to\infty} \frac{1}{n}\Gamma_n(\gamma^0) = b > 0$. Тогда ОМП $\hat{\gamma}$ асимптотически нормально распределена:

$$L\left\{(\Gamma_n(\gamma^0))^{\frac{-1}{2}})^T(\hat{\gamma} - \gamma^0)\right\} \xrightarrow[n\to\infty]{} N_{m+1}(0_{m+1}, I_{m+1}).$$

2. ПРОВЕРКА СЛОЖНЫХ ГИПОТЕЗ

Введем обозначения $\boldsymbol{X} = (X_1,...,X_n) \in \boldsymbol{R}^{nN}$, $\boldsymbol{H} = (\nu_1,...,\nu_n) \in \boldsymbol{K}^n$. Пусть Ω – множество всевозможных параметров. Определим две сложные гипотезы:

$$H_0: \ \gamma^0 = \bar{\gamma},$$
$$H_1: \ \gamma^0 \neq \bar{\gamma}.$$

Статистика отношения правдоподобия для проверки сложных гипотез H_0, H_1 примет вид $\Lambda = \Lambda_{\boldsymbol{X}}(\boldsymbol{H}) = \dfrac{P_{\boldsymbol{X}}(\boldsymbol{H}, \gamma^0)}{\sup_{\gamma \in \Omega} P_{\boldsymbol{X}}(\boldsymbol{H}, \gamma)} \in [0,1]$, где $P_{\boldsymbol{X}}(\boldsymbol{H}, \gamma) = \prod_{t=1}^{n} P_{x_t}(\nu_t; \gamma)$ [2].

Пусть $F_{\chi^2_m}$ – функция распределения вероятностей χ^2_m с m степенями свободы. Обозначим $(A^{-1})_{i,j}$ – i,j элемент матрицы A^{-1}.

Теорема 2. Пусть выполнены условия теоремы 1. Тогда для любого наперед заданного ε, $\varepsilon \in [0,1]$, существует $c^* = F_{\chi^2_m}^{-1}(1-\varepsilon)$, такое, что предел при $n \to \infty$ размера решающего правила

$$d^* = d^*_{\boldsymbol{X}}(\boldsymbol{H}) = \begin{cases} 0, & -2\ln \Lambda_{\boldsymbol{X}}(\boldsymbol{H}) < c^* \\ 1, & -2\ln \Lambda_{\boldsymbol{X}}(\boldsymbol{H}) \geq c^* \end{cases}$$

не превосходит ε. Решающее правило $d^*_{\boldsymbol{X}}(\boldsymbol{H})$ обладает максимальной мощностью среди всех решающих правил $\tilde{d} = \tilde{d}_{\boldsymbol{X}}(\boldsymbol{H})$, для которых $P_{\boldsymbol{X},\bar{\gamma}}\{\tilde{d} = 1\} \leq P_{\boldsymbol{X},\bar{\gamma}}\{d^* = 1\}$.

Теорема 3. Пусть выполнены условия теоремы 1. Тогда для любого наперед заданного ε, $\varepsilon \in [0,1]$, существует $c^* = \max_{i=1,\ldots,n} \sqrt{((\Gamma_n)^{-1})_{i,j}}\, \Phi^{-1}(1-\frac{\varepsilon}{2})$, такое, что предел при $n \to \infty$ размера решающего правила

$$d^* = d^*_{\boldsymbol{X}}(\boldsymbol{H}) = \begin{cases} 0, & \max_{i=1,\ldots,n} |\hat{\gamma}_i - \bar{\gamma}_i| < c^*; \\ 1, & \max_{i=1,\ldots,n} |\hat{\gamma}_i - \bar{\gamma}_i| \geq c^* \end{cases}$$

не превосходит ε. Решающее правило $d^*_{\boldsymbol{X}}(\boldsymbol{H})$ обладает максимальной мощностью среди всех решающих правил $\tilde{d} = \tilde{d}_{\boldsymbol{X}}(\boldsymbol{H})$, для которых $P_{\boldsymbol{X},\bar{\gamma}}\{\tilde{d} = 1\} \leq P_{\boldsymbol{X},\bar{\gamma}}\{d^* = 1\}$.

3. КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

Для компьютерного моделирования в качестве функции нелинейной регрессии использовалась производственная функция Кобба-Дугласа [3]:

$$Y_t = \theta_1^0 (X_t^1)^{\theta_2^0} (X_t^2)^{\theta_3^0} + \xi_t,\ t = 1,\ldots,n,$$

Рассматривались гипотезы H_0, H_1:

$$H_0: \quad \gamma^0 = (1,3,4,1)^T$$
$$H_1: \quad \gamma^0 \neq (1,3,4,1)^T.$$

При моделировании предполагались $A_1 = (-\infty,10]$, $A_2 = (10,40]$, $A_3 = (40,60]$, $A_2 = (60,+\infty)$. Значения регрессоров $\{X_t^1, X_t^2\}_{t=1}^n$ представляют собой узлы равномерной сетки на [0,2]×[0,2]. Проводились две серии экспериментов. В первом случае для моделирования использовался вектор параметров $\gamma^0 = (1,3,4,1)^T$, во втором – $\gamma^0 = (1.1,3.1,3.9,2)^T$. По методу Монте-Карло для каждого значения объема выборки *n* проводилось *Q*=100 экспериментов и строились статистические тесты $d^*_{\boldsymbol{X}}(\boldsymbol{H}_1^q)$ и $d^*_{\boldsymbol{X}}(\boldsymbol{H}_2^q)$ по первому и второму методам для проверки гипотез H_0, H_1. Вычислялись статистики

$$\hat{\alpha} = \sum_{q=1}^n d^*_{\boldsymbol{X}}(\boldsymbol{H}_1^q), \ \hat{w} = 1 - \sum_{q=1}^n d^*_{\boldsymbol{X}}(\boldsymbol{H}_2^q).$$

Статистика $\hat{\alpha}$ является оценкой ошибки первого рода построенного статистического теста, статистика $\hat{w}$ является оценкой мощности построенного статистического теста. На рисунках 1 и 2 изображены графики зависимости $\hat{\alpha}$ и $\hat{w}$ от объема выборки *n* соответственно.

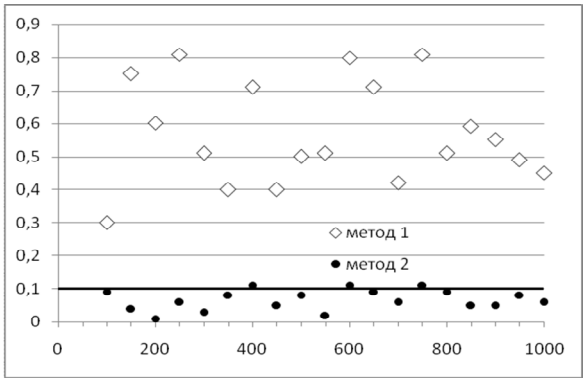

Рис. 1. График зависимости $\hat{\alpha}$ от *n*

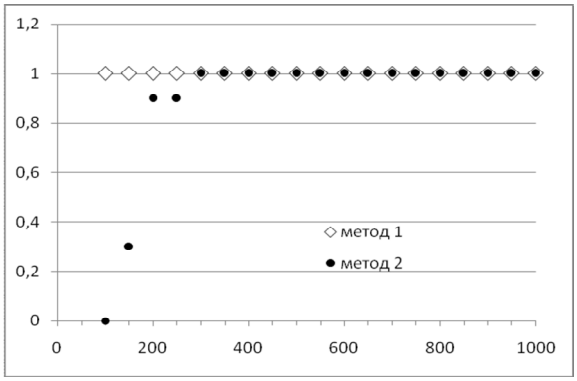

Рис. 2. График зависимости $\hat{w}$ от *n*

ЗАКЛЮЧЕНИЕ

Рассмотрена регрессионная модель, в которой зависимые данные наблюдаются не полностью: вместо точных значений известны только номера классов, в которые они попадают. Построены статистические тесты для проверки сложных гипотез для множественной регрессии при наличии классификации наблюдений. Проведены численные эксперименты, иллюстрирующие теоретические результаты.

Литература

1. *Heitjan D. F.* (1989) Inference from Grouped Continuous Data: A Review // Statistical Science, Vol. 4, no. 2, C. 164–183.
2. *Боровков А. А.* Математическая статистика // М.: Наука, 1984.
3. *Калитин Н.Н.* Численные методы // М.: Наука, 1978.
4. *Агеева Е. С., Харин Ю. С.* (2012) Состоятельность оценки максимального правдоподобия параметров множественной регрессии по классифицированным наблюдениям // Доклады НАН Беларуси Том 56, No.5, С. 11–19.

ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ В АНТИ-СПАМ ТЕХНОЛОГИЯХ

А. О. Варивончик

ВВЕДЕНИЕ

В последние десятилетия в мире бурно развивается новая прикладная область математики, специализирующаяся на искусственных нейронных сетях. Актуальность исследований в этом направлении подтверждается массой различных применений нейронных сетей. Это автоматизация процессов распознавания образов, адаптивное управление, аппроксимация функционалов, прогнозирование, создание экспертных систем, организация ассоциативной памяти и многие другие приложения. С помощью нейронных сетей можно, например, предсказывать показатели биржевого рынка, выполнять распознавание оптических или звуковых сигналов, создавать самообучающиеся системы, способные управлять автомашиной при парковке или синтезировать речь по тексту.

В настоящее время компьютеры достигли необходимой вычислительной мощности для проведения серьёзных исследований. Например, ученые из Google запустили нейронную сеть на 16 тысячах процессоров. В результате они смоделировали систему с примерно миллиардом взаимосвязей между отдельными процессами (нейронами). В качестве материала для работы полученная сеть использовала видеоролики с YouTube. Как следствие, спустя некоторое количество времени, система научилась отличать видео с котами от остальных.

Использование нейронных сетей для распознавания спама является логическим развитием поиска более продуктивных алгоритмов и начало развиваться сравнительно недавно. Как раз то, что на данном этапе компьютеры достигли достаточных вычислительных мощностей для реализации обучения сложных нейронных сетей на огромных базах данных, дало возможность реализации предлагаемых решений.