

Раннее прогнозирование достаточного объема выборки для обобщенной линейной модели

Валентин Бучнев

Московский физико-технический институт

группа 694, 2020

Прогнозирование объема выборки

Цель исследования

Предложить метод предсказания достаточного объема выборки для обобщенной линейной модели на ранних этапах сбора данных.

Проблема

Большинство методов требуют заведомо избыточного объема выборки.

Метод решения

Оценка объема строится по собранной выборке путем анализа свойств функции ошибки обобщенной линейной модели.

Ассимптотические методы

- S. G. Self and R. H., Mauritsen Power/sample size calculations for generalized linear models // Biometrics, 1988
- G. Shieh, On power and sample size calculations for likelihood ratio tests in generalized linear models // Biometrics, 2000.
- G. Shieh On power and sample size calculations for Wald tests in generalized linear models // Journal of Statistical Planning and Inference, 2005.

Байесовские методы

- D. B. Rubin and H. S. Stern Sample size determination using posterior predictive distributions // Sankhya : The Indian Journal of Statistics Special Issue on Bayesian Analysis, 1998.

Постановка задачи раннего прогнозирования

Дано

Выборка размера m : $\mathfrak{D}_m = \{\mathbf{x}_i, y_i\}_{i=1}^m$,
где $\mathbf{x}_i \in \mathbb{R}^n$ - вектор признаков, $y_i \in \mathbb{Y}$.

Функция правдоподобия

Определим функцию правдоподобия и логарифмическую функцию правдоподобия выборки \mathfrak{D} :

$$L(\mathfrak{D}_m, \mathbf{w}) = \prod_{y, \mathbf{x} \in \mathfrak{D}_m} f(y, \mathbf{x}, \mathbf{w}), \quad l(\mathfrak{D}_m, \mathbf{w}) = \sum_{y, \mathbf{x} \in \mathfrak{D}_m} \log f(y, \mathbf{x}, \mathbf{w}),$$

где $f(y, \mathbf{x}, \mathbf{w})$ - аппроксимация плотности апостериорной вероятности выборки \mathfrak{D}_m при заданном векторе параметров \mathbf{w} .

Логарифмическая функция правдоподобия

Будем рассматривать ожидаемое значение функции l :

$$\tilde{l}(\mathfrak{D}) = \mathbb{E}_{y, \mathbf{x} \in \mathfrak{D}} l(\{y, \mathbf{x}\}, \mathbf{w}).$$

Ожидаемое значение

Рассмотрим ожидаемое значение логарифма правдоподобия по разным обучающим выборкам $\mathfrak{D}_{\mathcal{L}_m}$ размера m :

$$l(m) = \mathbb{E}_{\mathfrak{D}_{\mathcal{L}_m}} \tilde{l}(\mathfrak{D}_{\mathcal{L}_m}).$$

Критерий достаточности объема

Будем считать, что объем выборки m^* достаточен, если:

$$\forall m' > m^* \quad I(m') > (1 - \delta) \max_{m > m^*} I(m),$$

где δ — достаточно малое пороговое значение.

Предлагаемый метод решения

Семейство функций Φ

Для предсказания значения функции $\hat{l}(m)$ при $m > m_0$ введем параметрическое семейство функций

$$\Phi = \{\phi(m) = a + b \cdot e^{c \cdot m} \mid a, b \in \mathbb{R}, c \in (-\infty, 0)\} \quad (1)$$

Аппроксимация $\phi(m) \sim l(m)$

Аппроксимация функции $\hat{l}(m)$ является решением следующей задачи:

$$\hat{\phi} = \arg \min_{\phi \in \Phi} \text{MAE}(\hat{l}, \phi, 1, m_0), \quad (2)$$

$\text{MAE}(\psi, \phi, m_1, m_2)$

$$\text{MAE}(\psi, \phi, m_1, m_2) = \frac{1}{m_2 - m_1 + 1} \sum_{i=m_1}^{m_2} |\phi(i) - \psi(i)|. \quad (3)$$

Оценка \hat{m}^*

$$\hat{m}^* = \min_m \max_{m' > m} \hat{\phi}(m') > (1 - \delta)\hat{\phi}(m), \quad (4)$$

где δ — достаточно малое пороговое значение.

Цель эксперимента

Проверить работоспособность предложенного метода.

Выборки

Выборка	Тип задачи	m^*	n^*
Синтетическая, случайная выборка	регрессия	72	10
Синтетическая, скоррелированная выборка	регрессия	31	2
Синтетическая, ортогональная выборка	регрессия	45	10
Синтетическая, избыточная выборка	регрессия	22	5

Выборка	Тип задачи	m^*	n^*
UCI repo, Diabetes	регрессия	442	11
UCI repo, Boston	регрессия	506	14
UCI repo, Wine	классификация	130	14
UCI repo, Nba	классификация	400	20

- Задача прогнозирования достаточного объема выборки сведена к задаче аппроксимации функции ошибок.
- Показана работоспособность предложенного метода на синтетических выборках, а также на выборках из UCI репозитория.