Раннее прогнозирование достаточного объема выборки для обобщенной линейной модели

Валентин Бучнев

Московский физико-технический институт

группа 694, 2020

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.

Существующие методы

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: $\mathfrak{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{m}$,
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 $\mathfrak{D}$:

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

где $p(y|\mathbf{x}, \mathbf{w})$ — плотность зависимой переменной.

Постановка задачи раннего прогнозирования

Функция ошибки

Будем рассматривать ожидаемое значение функции $e^{-S(\hat{\mathbf{w}}(\mathfrak{D}_\mathcal{L})|\mathfrak{D}_\mathcal{T}))}$ по разным обучающим и тестовым выборкам размера m:

$$I(m) = \mathsf{E}e^{-S(\hat{\mathbf{w}}(\mathfrak{D}_\mathcal{L})|\mathfrak{D}_\mathcal{T}))}.$$

Функция ошибки $S(\mathbf{w}, \mathfrak{D})$ для задач регрессии и классификации

$$S_{\text{reg}}(\mathbf{w}|\mathfrak{D}) = \frac{1}{|\mathfrak{D}|}\sum_{\mathbf{x},y\in\mathfrak{D}}(y - f(\mathbf{x},\mathbf{w}))^2,$$

$$S_{\text{class}}(\mathbf{w}|\mathfrak{D}) = \frac{1}{|\mathfrak{D}|}\sum_{\mathbf{x},y\in\mathfrak{D}}\big(y\ln f(\mathbf{x},\mathbf{w}) + (1-y)\ln(1 - f(\mathbf{x},\mathbf{w}))\big).$$

Постановка задачи раннего прогнозирования

Функция ошибки $I(m)$

$\hat{I}(m)$ — оценка функции $I(m)$, посчитанная с помощью метода бутстреп по разным обучающим и тестовым подвыборкам выборки $\mathfrak{D}$.

Критерий достаточности объема

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

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

где δ — достаточно малое пороговое значение.

Предлагаемый метод решения

Семейство функций Φ

Для предсказания значения функции $I(m)$ при $m > m_0$ введем параметрическое семейство функций:

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

Аппроксимация $\phi(m) \sim I(m)$

Аппроксимация функции $I(m)$ является решением следующей задачи:

$$\hat{\phi} = \arg\min_{\phi \in \Phi} \mathrm{MAE}(I, \phi, 1, m_0),$$

где

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

Критерий достаточности объёма

Оценка $\hat{m}^*$

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

где δ — достаточно малое пороговое значение.

Вычислительный эксперимент

Цель эксперимента

Проверить работоспособность предложенного метода.

Результаты

Выборка	Тип задачи	m^*	n^*
Синтетическая, случайная выборка	регрессия	72	10
Синтетическая, скоррелиро-ванная выборка	регрессия	31	2
Синтетическая, ортогональ-ная выборка	регрессия	45	10
Синтетическая, избыточная выборка	регрессия	22	5

Результаты

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

Заключение

- 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.