Раннее прогнозирование достаточного объема выборки для обобщенной линейной модели

Валентин Бучнев

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов), группа 694, весна 2019
консультант: А. В. Грабовой

Прогнозирование объема выборки

Цель исследования

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

Проблема

Большинство неассимптотических 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: $\mathfrak{D}_m = \{\mathbf{x}_i, y_i\}_{i=1}^m$,
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 $\mathbf{w}$.

Постановка задачи раннего прогнозирования

Логарифмическая функция правдоподобия

Будем рассматривать ожидаемое значение функции l:

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

Ожидаемое значение

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

$$l(m^*) = \mathop{\mathsf{E}}_{\mathfrak{D}_{\mathcal{L}_m}} \tilde{l}(\mathfrak{D}_{\mathcal{L}_m}).$$

Постановка задачи раннего прогнозирования

Критерий достаточности объема

Будем считать, что объем выборки достаточный, если:

$$\forall m_1, m_2 > m^* \quad |l(m_1) - l(m_2)| < \varepsilon,$$

где ε - достаточно малое пороговое значение.

Предлагаемый метод решения

Критерий средней длины

$$A(\mathfrak{D}) = \{\mathbf{w} : ||\mathbf{w} - \hat{\mathbf{w}}|| \leqslant r_m\}$$

$$\mathsf{P}(A(\mathfrak{D})) = 1 - \alpha,$$

где α — некоторое малое значение.
Критерий средней длины выглядит следующим образом:

$$\forall m \geqslant m^* \ \mathsf{E}_{\mathfrak{D}_m} r_m \leqslant l,$$

где r_m — радиус шара $A(\mathfrak{D}_m)$, l — некоторое наперед заданное достаточно малое значение.

Критерий средней длины

Оценка вектора параметров

Для оценки вектора параметров используется принцип максимума правдоподобия:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}\in\mathbb{W}} L(\mathfrak{D}_{\mathcal{L}_m}, \mathbf{w}).$$

Распределение оценки вектора параметров

Далее используется предположение о распределении оценки вектора параметров:

$$\hat{\mathbf{w}} \sim \mathcal{N}(\mathbf{m}, \mathsf{D}\hat{\mathbf{w}}),$$

и с помощью сэмплирования вычисляется приближенное значение r_m.

Модификация критерия средней длины

Ковариационная матрица вектора параметров

Воспользуемся эффективностью оценки $\hat{\mathbf{w}}$:

$$\mathsf{D}\hat{\mathbf{w}} = \mathbf{I}^{-1}(\mathfrak{D}_m),$$

где $\mathbf{I}(\mathfrak{D}_m)$ — информационная матрица Фишера.

Матрица Фишера

Для аппроксимации матрицы Фишера для m наблюдений воспользуемся свойством:

$$\hat{\mathbf{I}}(\mathfrak{D}_m) = \frac{m}{m^*}\,\mathbf{I}(\mathfrak{D}_{m^*}).$$

Вычисление функции эффективности

Таким образом, построена аппроксимация параметров распределения $\hat{\mathbf{w}}$ для m наблюдений при использовании выборки размера m^* для вычисления приближенного значения функции эффективности:

$$\hat{\mathbf{w}} \sim \mathcal{N}(\mathbf{m}, \frac{m}{m^*}\ \mathbf{I}^{-1}(\mathfrak{D}_{m^*})).$$

Вычислительный эксперимент

Цель эксперимента

Проверить работоспособность предложенного метода.

Выборки из UCI репозитория.

Выборка	Тип задачи	Размер выборки	Число признаков
Servo	регрессия	167	4
Boston	регрессия	506	13
Diabetes	регрессия	442	5

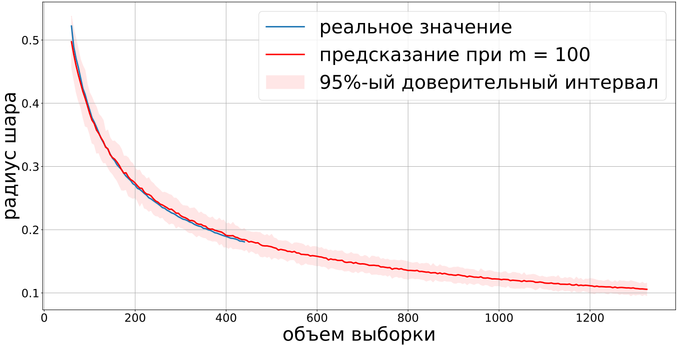

Рис.: ALC метод, выборка Diabetes

Предсказание достаточного объема выборки, ALC метод.

Выборка	Реальное значение	Предсказание
Servo	не хватает данных	450
Boston	не хватает данных	1370
Diabetes	235	240

Заключение

- Задача прогнозирования достаточного объема выборки сведена к задаче аппроксимации корреляционной матрицы вектора параметров.
- Показана работоспособность предложенного метода на тестовых выборках.
- Далее предлагается строить аппроксимацию зависимости ожидаемого значения логарифма правдоподобия от размера выборки.