Раннее прогнозирование достаточного объема выборки для обобщенной линейной модели.

Валентин Бучнев

МФТИ
ФИВТ

buchnev.valentin@gmail.com

21 марта 2019 г.

Цели исследования

Цель работы

Предложить метод предсказания достаточного объема выборки для обобщенной линейной модели.

Проблема

Большинство неассимптотических методов требуют заведомо большого объема выборки.

Метод решения

Модификация таких методов путем введения нескольких аппроксимаций для получения возможности прогнозирования изменений зависимостей при увеличении размера выборки.

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 $\mathfrak{D}$:

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

где $f(y, \mathbf{x}, \mathbf{w})$ - аппроксимация плотности апостериорной вероятности выборки $\mathfrak{D}_{\mathcal{L}_m}$ 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:

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

где ε - достаточно малое пороговое значение.

Предлагаемый метод решения

Критерий средней длины

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

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

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

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

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

Предлагаемый метод решения

Модификация критерия

Предлагается заменить ковариационную матрицу вектора параметров на её аппроксимацию через матрицу информации Фишера, далее строить аппроксимацию зависимости значения функции эффективности от размера выборки.

Вычислительный эксперимент

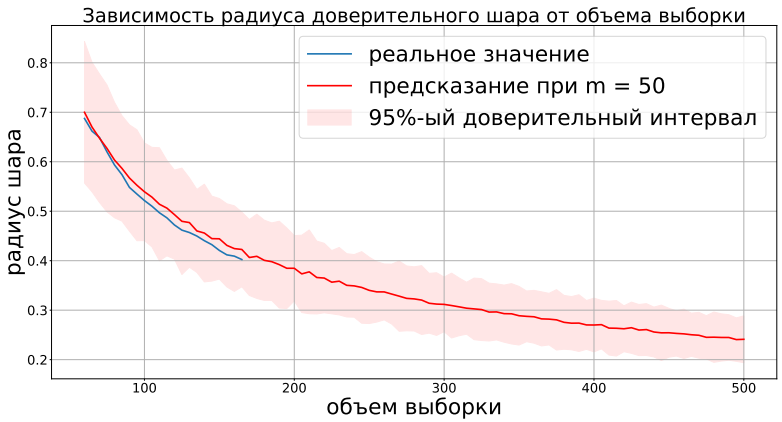
Цель эксперимента

Проверить работоспособность предложенного метода.

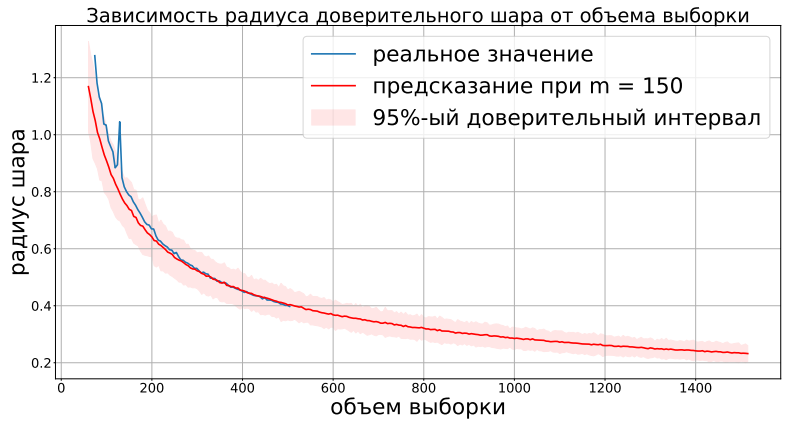
Выборки из UCI репозитория.

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

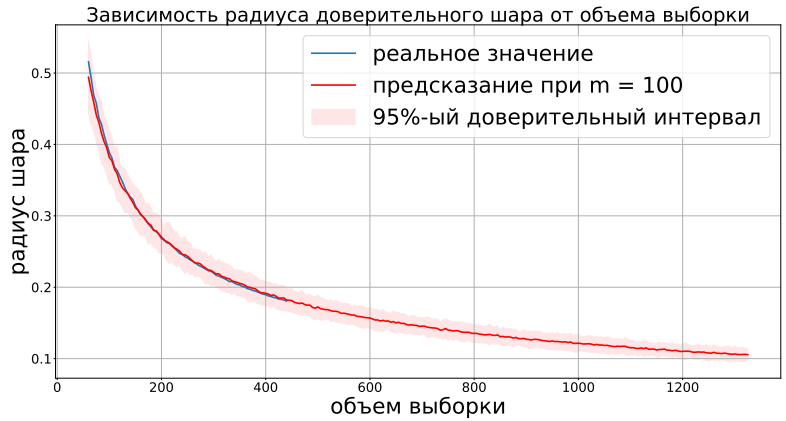
Результаты



(a) Servo



(b) Boston



(c) Diabetes

Заключение

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