Раннее прогнозирование достаточного объема выборки для обобщенно линейной модели

Владимир Александрович Жолобов

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 774, весна 2020

Цель исследования

Цель работы

Предложить метод оценки достаточного объема выборки на ранних этапах сбора данных.

Проблема

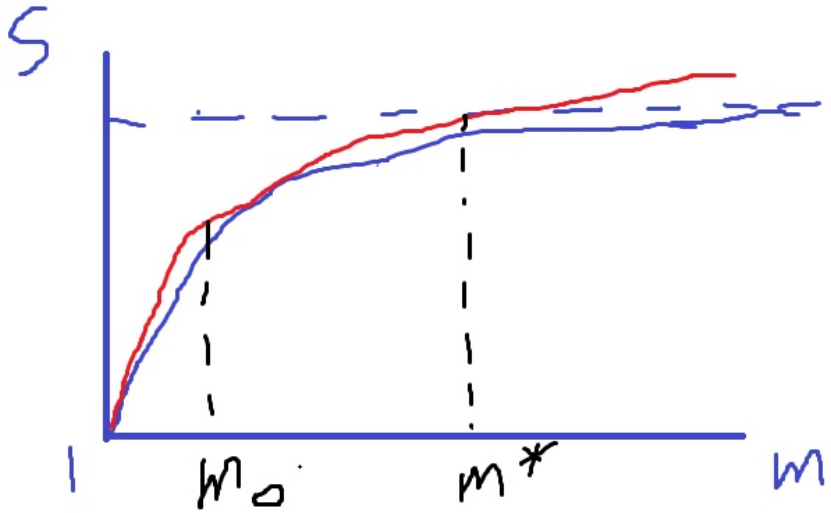
В некоторых ситуациях, таких как крупное медицинское исследование, сбор данных являетя дорогостоящим.

Метод решения

Аппроксимация эмпирической функции ошибки из параметрического семейства функций.

Иллюстрация проблемы

Рис.: Эмпирическая функция ошибки и ее аппроксимация



Задана выборка $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$. Объем выборки $m_{\mathfrak{D}} = |\mathfrak{D}|$. Задана модель $f(\mathbf{w}, x)$ с параметрами из распределения $P(\mathbf{w}|f, \mathfrak{D}, m)$. Функция ошибки точности $S(\mathbf{w})$ имеет вид

$$S = S(\mathfrak{D}, f, \mathbf{w}^*, m) \tag{1}$$

Требуется найти параметры регрессионной модели $\mathbf{w}^*$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^n} S(w|\mathfrak{D}, f) \tag{2}$$

В задаче поиска оптимального состава признаков $\mathbf{x}_{\mathcal{A}} = [x_{1\mathcal{A}} \dots x_{n\mathcal{A}}]^T$ требуется оптимизировать набор признаков $n^* = |\mathcal{A}|$

$$\mathcal{A}^* = \arg\min_{\mathcal{A}\subset\{1,\dots,n\}} S(\mathcal{A}|\mathbf{w}, \mathfrak{D}) \tag{3}$$

Функции ошибки для обобщенно-линейной модели

Для линейной регрессии функция ошибки

$$S(\mathbf{w}) = ||\mathbf{f} - y||_2^2. \tag{4}$$

Для логистической регрессии функция ошибки в общем случае

$$S(\mathbf{w}) = \sum_{i,k} y_{i,k} \ln f_{i,k} \tag{5}$$

Функции ошибки

- Теоретическая функция ошибки S — дважды дифференцируемая по m функция при заданном распределении $p(y)$.
- Эмпирическая функция ошибки S — реализация случайной величины с математическим ожиданием $\mathsf{E}S$ и дисперсией $\mathsf{D}S$ при заданной выборке.
- Аппроксимация функции ошибки $\varphi^*(m)$ — функция из семейства E.

Эмпирическая функция ошибки $S(m)$

Реализация случайной величины с матемматическим ожиданием

$$\mathrm{E}S(m) = \frac{1}{m}\sum_{i=1}^{m} S_i. \tag{6}$$

и дисперсией

$$\mathrm{D}S(m) = \frac{1}{m}\sum_{i=1}^{m}(S_i - \mathrm{E}S(m))^2. \tag{7}$$

Предполагается, что эмпирическая функция ошибки является сходящейся к значению r, то есть

$$\exists m^* < +\infty : \quad \frac{dS}{dm} = 0, \quad \forall m \geq m^*$$

$$\forall \varepsilon > 0 \quad \exists m' > 1 : \quad ||S(m) - r||_1 < \varepsilon, \quad \forall m \geq m'. \tag{8}$$

Аппроксимация эмпирической функции ошибки $S(m)$

Семейство параметризированных функций задается в виде

$$E = \{\exp(w_1 + w_2 \ln m) + w_3 + \frac{w_4}{m^2 + 1} | w_2 > 0\} \tag{9}$$

Задача оптимизация ставится

$$\min_{\mathbf{w}} || \exp(w_1 + w_2 \ln m) + w_3 + \frac{w_4}{m^2 + 1} - \mathrm{E}S(m)||_1$$

$$\text{s.t. } w_2 > 0 \tag{10}$$

Искомая функция $\varphi^*(m)$

$$\varphi^*(m) = \arg\min_{w} || \exp(w_1 + w_2 \ln m) + w_3 + \frac{w_4}{m^2 + 1} - \mathrm{E}S(m)||_1 \tag{11}$$

Цель эксперимента

Проверить работоспособность предложенного метода на выборках

Таблица: Описание выборок

Выборка	Тип задачи	Размер выборки	Число признаков
Boston	Регрессия	506	13
Diabets	Регрессия	442	10
Synthetic 1	Регрессия	50000	4

Рис.: Визуализация в случае отсутствия отбора параметров

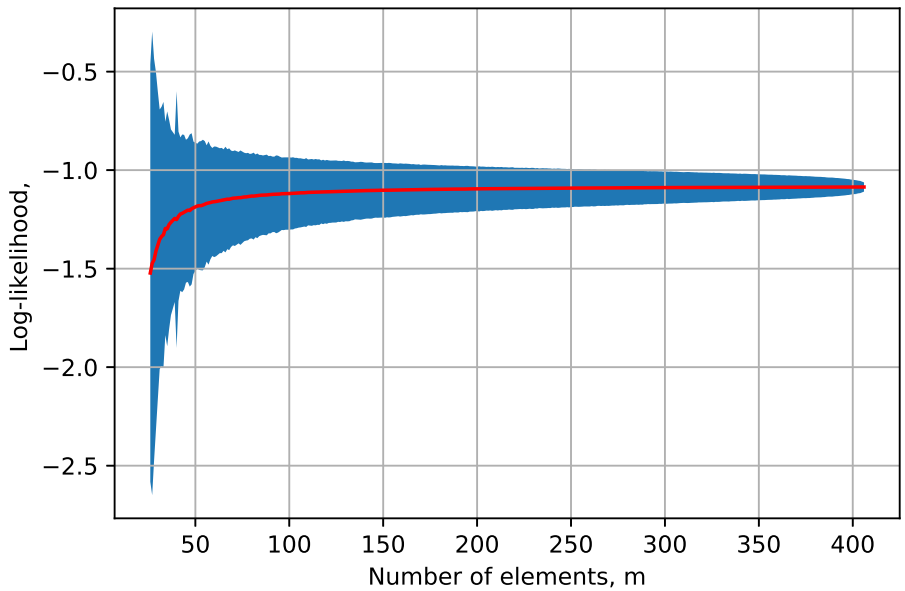


Рис.: Эмпирическая функция ошибки и ее аппроксимация из параметрического семейства

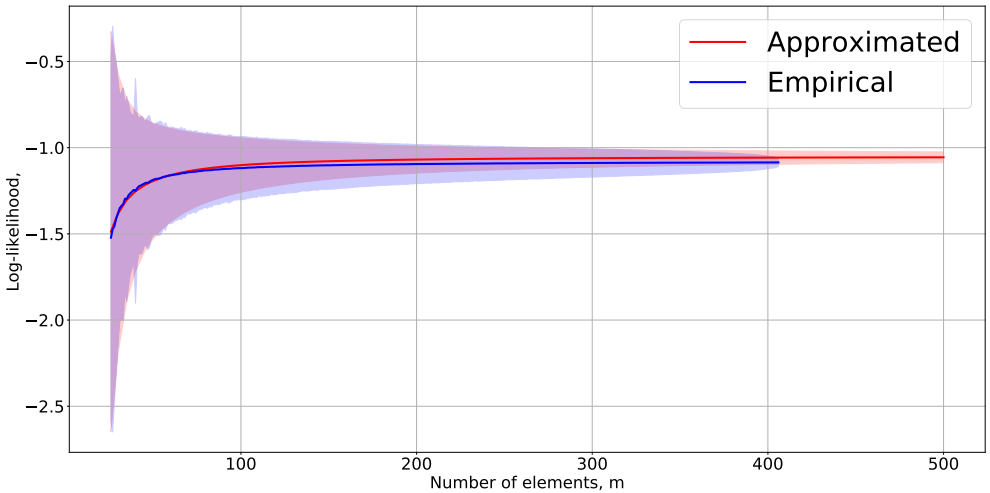


Рис.: Визуализация в случае отсутствия отбора параметров

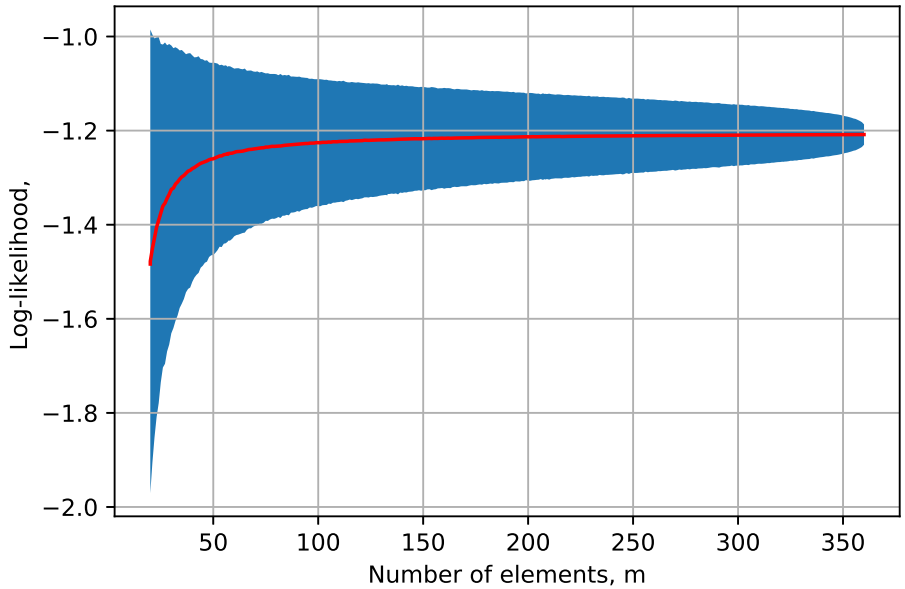
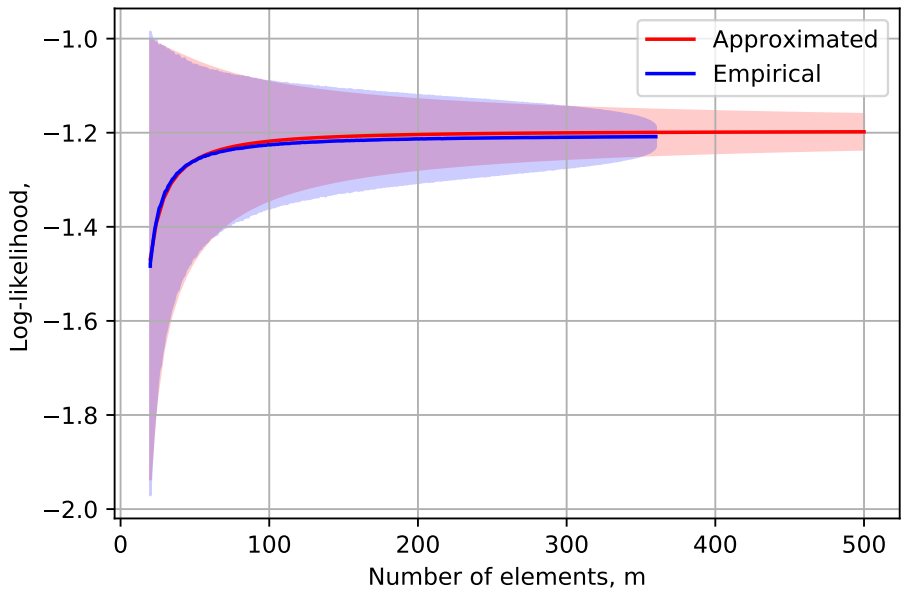


Рис.: Эмпирическая функция ошибки и ее аппроксимация из параметрического семейства



Заключение

- Задача аппроксимации эмпирической функции ошибки сформулирована как задача условной оптимизации
- Показана эффективность предложенного метода. Имеет место дальнейшее улучшения метода с целью уменьшения дисперсии аппроксимации.

Список литературы:


Anastasiya Motrenko, Vadim Strijov, and Gerhard-Wilhelm Weber. Sample size determination for logistic regression. Journal of computational and applied mathematics, ISSN 0377-0427, 255(1):743– 752, 2014.


A. S. Kulunchakov and Vadim V. Strijov. Generation of simple structured information retrieval functions by genetic algorithm without stagnation. Expert Syst. Appl, 85:221–230, 2017.


Parantapa Goswami, Simon Moura, Eric Gaussier, Massih-Reza Amini, and Francis Maes. Exploring the space of ir functions. pages 372–384, April 14 2014.


C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.