Байесовский подход к выбору достаточного размера выборки

Киселев Никита Сергеевич

Научный руководитель:
к.ф.-м.н. А. В. Грабовой

Московский физико-технический институт
(национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Москва — 2023

Байесовский выбор достаточного размера выборки

Исследуется задача выбора достаточного размера выборки.

Проблема

Большинство подходов используют распределение параметров модели. Статистические методы требуют для оценки избыточный размер доступной выборки.

Цель

Требуется предложить метод, не использующий напрямую параметры модели. Необходимо учесть недостаточный размер доступной выборки.

Решение

Предлагается использовать функцию правдоподобия выборки. Рассматривается подход в случаях избыточного и недостаточного размеров доступной выборки.

Постановка задачи выбора размера выборки

Выборка

$$\mathfrak{D}_m = \{\mathbf{x}_i, y_i\}_{i=1}^m, \ \mathbf{x}_i \in \mathbb{X}, \ y_i \in \mathbb{Y}.$$

Параметризация распределения

$$p(y|\mathbf{x}) \quad \longrightarrow \quad p(y|\mathbf{x}, \mathbf{w}), \ \mathbf{w} \in \mathbb{W}.$$

Функция правдоподобия выборки

$$L(\mathfrak{D}_m, \mathbf{w}) = \prod_{i=1}^m p(y_i|\mathbf{x}_i, \mathbf{w}), \qquad l(\mathfrak{D}_m, \mathbf{w}) = \sum_{i=1}^m \log p(y_i|\mathbf{x}_i, \mathbf{w}).$$

Оценка максимального правдоподобия

$$\hat{\mathbf{w}}_m = \arg\max_{\mathbf{w}} L(\mathfrak{D}_m, \mathbf{w}).$$

Цель

Требуется определить достаточный размер выборки m^*.

Достаточный размер выборки не превосходит доступный

Рассмотрим выборку $\mathfrak{D}_k$ размера $k \leqslant m$. Оценим на ней параметры, используя метод максимума правдоподобия:

$$\hat{\mathbf{w}}_k = \arg\max_{\mathbf{w}} L(\mathfrak{D}_k, \mathbf{w}).$$

Зафиксируем некоторое положительное число $\varepsilon > 0$.

Определение (D-достаточный размер выборки)

*Размер выборки m^* называется **D-достаточным**, если для любого $k \geqslant m^*$*

$$D(k) = \mathbb{D}_{\mathfrak{D}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k) \leqslant \varepsilon.$$

Определение (M-достаточный размер выборки)

*Размер выборки m^* называется **M-достаточным**, если для любого $k \geqslant m^*$*

$$M(k) = \left| \mathbb{E}_{\mathfrak{D}_{k+1}} L(\mathfrak{D}_m, \hat{\mathbf{w}}_{k+1}) - \mathbb{E}_{\mathfrak{D}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k) \right| \leqslant \varepsilon.$$

Корректность M-определения

Утверждение 1 (асимптотическая нормальность)

Пусть $\hat{\mathbf{w}}_k$ — оценка максимума правдоподобия $\mathbf{w}$. Тогда при определенных условиях регулярности имеет место следующая сходимость по распределению:

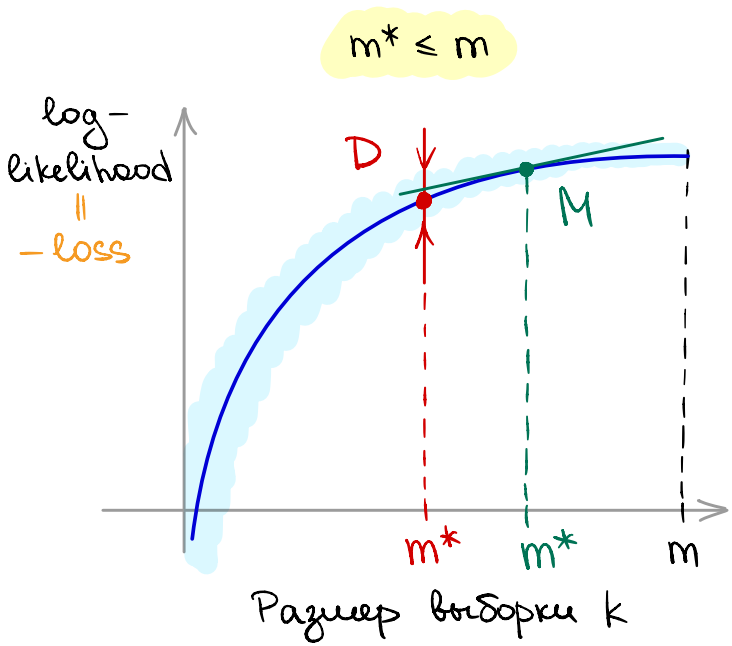
$$\hat{\mathbf{w}}_k \xrightarrow{d} \mathcal{N}\left(\mathbf{w}, [m\mathcal{I}(\mathbf{w})]^{-1}\right).$$

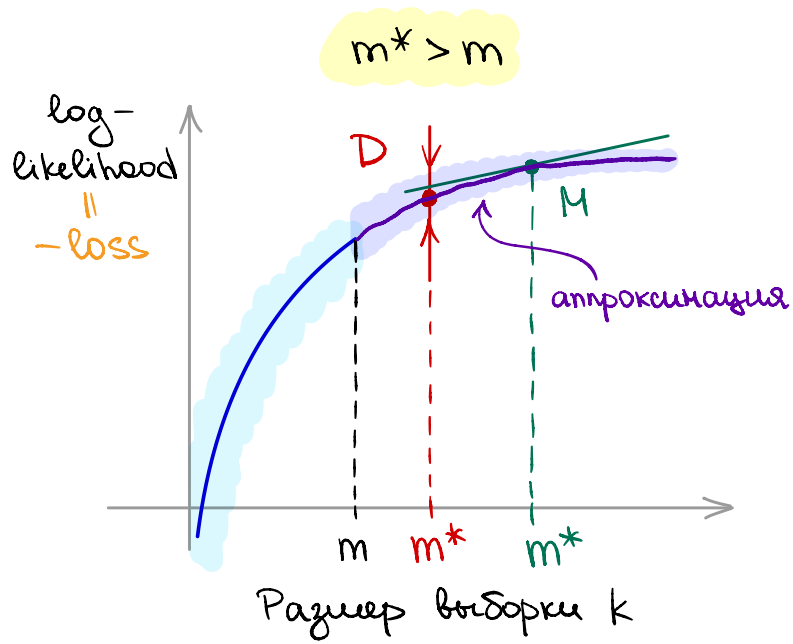
Лемма 1

Пусть $\|\mathbf{m}_k - \mathbf{w}\|_2 \to 0$ и $\|\mathbf{\Sigma}_k - [m\mathcal{I}(\mathbf{w})]^{-1}\|_F \to 0$ при $k \to \infty$. Тогда в модели линейной регрессии определение M-достаточного размера выборки является корректным. А именно, найдется такой m^*, что для всех $k \geqslant m^*$ выполнено $M(k) \leqslant \varepsilon$.

Достаточный размер выборки больше доступного

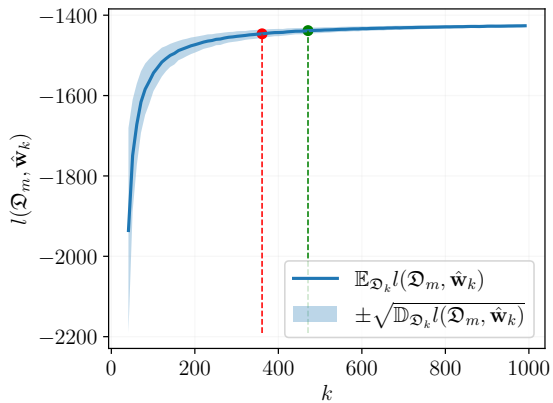
Возникает задача прогнозирования математического ожидания и функции правдоподобия при $k > m$.

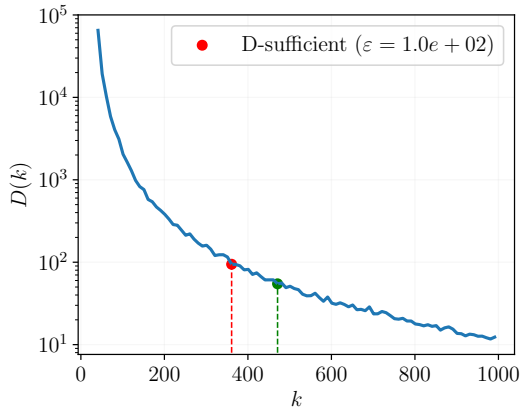


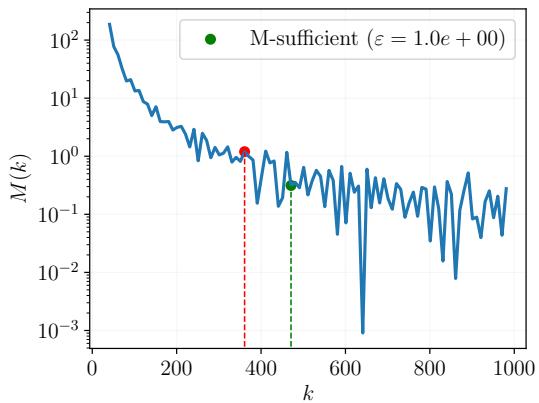


Синтетическая выборка при $m^* \leqslant m$

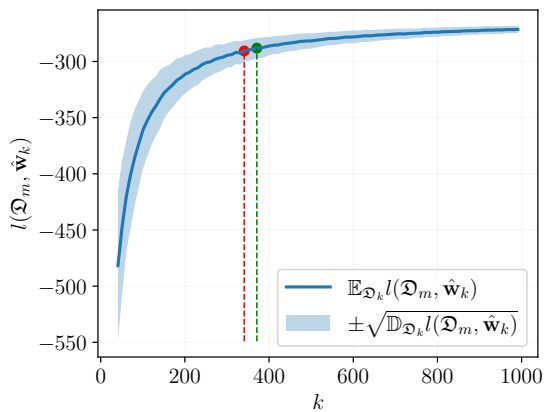
Линейная регрессия

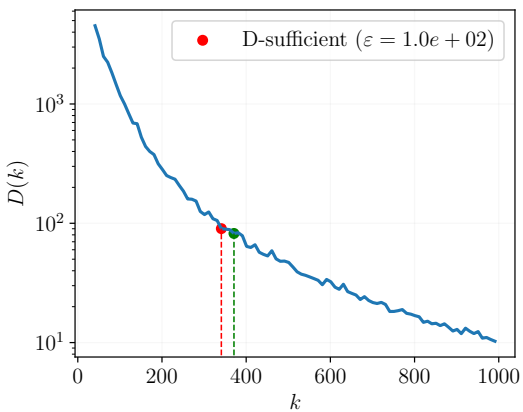


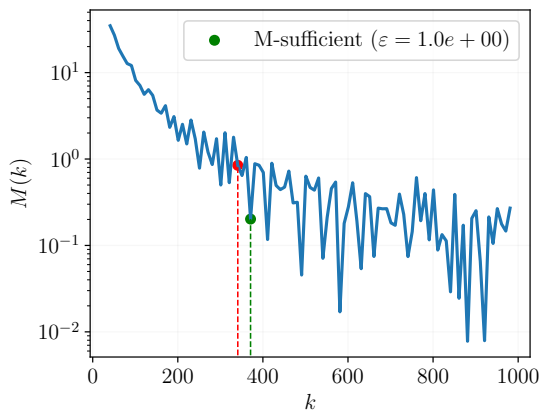




Логистическая регрессия







Синтетическая выборка при $m^* > m$

Для синтетических выборок проведена аппроксимация функций правдоподобия. Среднее значение и дисперсия аппроксимированы соответственно функциями

$$\varphi(m) = a_1 - a_2^2 \exp\left(-a_3^2 m\right) - \frac{a_4^2}{m^{3/2}}$$

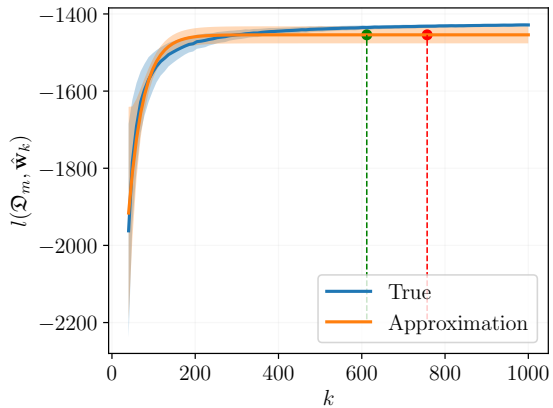
и

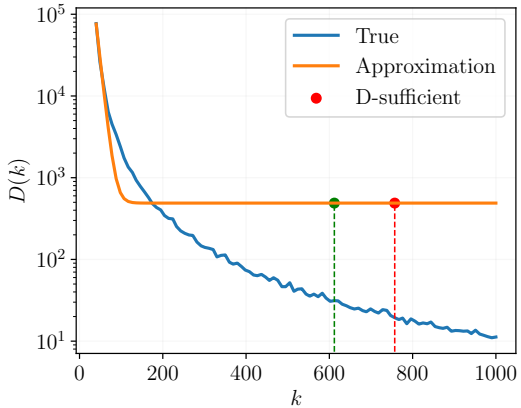
$$\psi(m) = b_1^2 \exp\left(-b_2^2 m\right) + \frac{b_3^2}{m^{3/2}},$$

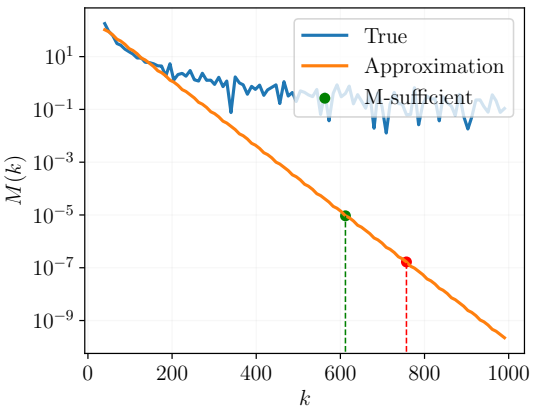
где $\mathbf{a}$ и $\mathbf{b}$ — вектора параметров.

Синтетическая выборка при $m^* > m$

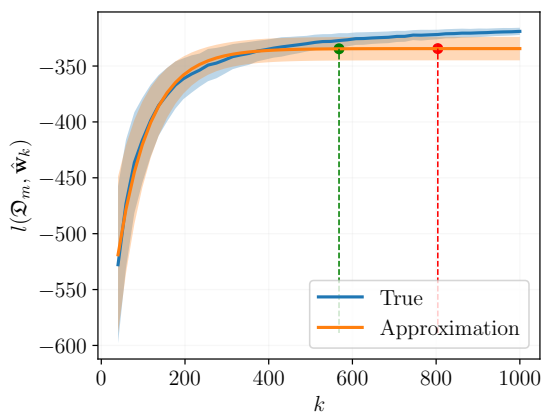
Линейная регрессия

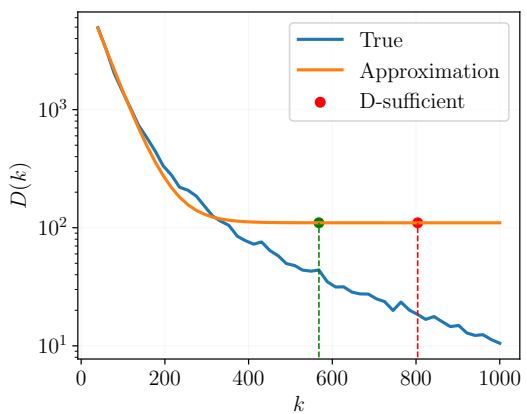


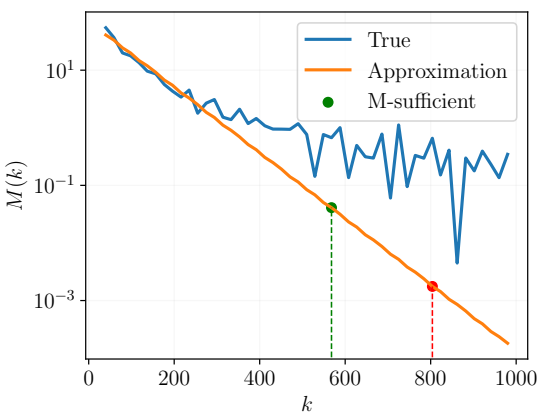




Логистическая регрессия







Дальнейшие цели

- Доказать корректность предложенных определений.
- Доказать «хорошие» свойства бутстрап-оценок математического ожидания и дисперсии функции правдоподобия выборки.
- Улучшить подход к прогнозированию функции правдоподобия при $m^* > m$.