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 — 2024

Байесовский подход к выбору достаточного размера выборки

Исследуется задача выбора достаточного размера выборки.

Проблема

Определение достаточного размера выборки без постановки статистической гипотезы о распределении параметров модели.

Цель

Предложить критерий определения достаточного размера выборки. Построить метод, реализующий этот критерий на практике.

Решение

Предлагаются подходы к определению достаточного размера выборки

- По значениям функции правдоподобия на бутстрапированных подвыборках;
- По близости апостериорных распределений параметров модели на схожих подвыборках.

Обозначения

Выборка

$$\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \dots, m\},$$

где $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^n$ есть вектор признакового описания объекта, а $y \in \mathbb{Y}$ есть значение целевой переменной.

Модель

$$f : \mathbb{X} \times \mathbb{W} \to \mathbb{Y},$$

где $\mathbb{W}$ есть пространство параметров модели.

Вероятностная модель

$$p(y, \mathbf{w}|\mathbf{x}) = p(y|\mathbf{x}, \mathbf{w})p(\mathbf{w}) : \mathbb{Y} \times \mathbb{W} \times \mathbb{X} \to \mathbb{R}^+,$$

где $p(y|\mathbf{x}, \mathbf{w})$ задает правдоподобие объекта, а $p(\mathbf{w})$ задает априорное распределение параметров.

Обозначения

Функция правдоподобия

$$L(\mathfrak{D}_m, \mathbf{w}) = p(\mathbf{y}_m | \mathbf{X}_m, \mathbf{w}) = \prod_{i=1}^{m} p(y_i | \mathbf{x}_i, \mathbf{w})$$

Логарифм функции правдоподобия

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

Оценка максимума правдоподобия

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

Постановка задачи

Ставится задача определения достаточного размера выборки m^*. Пусть задан некоторый критерий T. Он может быть построен, например, на основе эвристик о поведении параметров модели.

Определение

Размер выборки m^* называется **достаточным** согласно критерию T, если T выполняется для всех $k \geqslant m^*$.

Требуется

Предложить критерий T определения достаточного размера выборки m^*. Построить метод, реализующий критерий T на практике.

Анализ поведения функции правдоподобия

Определение (D-достаточный размер выборки)

Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется **D-достаточным**, если для всех $k \geqslant m^*$

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

Определение (M-достаточный размер выборки)

Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется **M-достаточным**, если для всех $k \geqslant m^*$

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

Замечание

В определениях выше вместо функции правдоподобия $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$ можно рассматривать ее логарифм $l(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$. На практике вместо функции правдоподобия можно использовать функцию ошибки. Математическое ожидание и дисперсия оцениваются по значениям на бутстрапированных подвыборках.

Корректность определения M-достаточности

Обозначим $\mathbb{E}_{\mathfrak{D}_k} \hat{\mathbf{w}}_k = \mathbf{m}_k$ и $\mathbb{D}_{\mathfrak{D}_k} \hat{\mathbf{w}}_k = \mathbf{\Sigma}_k$.

Теорема (Киселев, 2023)

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

Следствие

Пусть $\|\mathbf{m}_k - \mathbf{w}\|_2 \to 0$ и $\|\mathbf{\Sigma}_k - [k\mathcal{I}(\mathbf{w})]^{-1}\|_F \to 0$ при $k \to \infty$. Тогда в модели линейной регрессии определение M-достаточного размера выборки является корректным.

Анализ апостериорного распределения параметров модели

Рассмотрим две подвыборки $\mathfrak{D}^1 \subseteq \mathfrak{D}_m$ и $\mathfrak{D}^2 \subseteq \mathfrak{D}_m$. Пусть $\mathcal{I}_1 \subseteq \mathcal{I} = \{1, \ldots, m\}$ и $\mathcal{I}_2 \subseteq \mathcal{I} = \{1, \ldots, m\}$ — соответствующие им подмножества индексов.

Определение

Подвыборки $\mathfrak{D}^1$ и $\mathfrak{D}^2$ называются **схожими**, если

$$|\mathcal{I}_1 \triangle \mathcal{I}_2| = |(\mathcal{I}_1 \setminus \mathcal{I}_2) \cup (\mathcal{I}_2 \setminus \mathcal{I}_1)| = 1.$$

Рассмотрим две схожие подвыборки $\mathfrak{D}_k = (\mathbf{X}_k, \mathbf{y}_k)$ и $\mathfrak{D}_{k+1} = (\mathbf{X}_{k+1}, \mathbf{y}_{k+1})$ размеров k и $k+1$ соответственно. Найдем апостериорное распределение параметров модели по этим подвыборкам:

$$p_k(\mathbf{w}) = p(\mathbf{w}|\mathfrak{D}_k) = \frac{p(\mathfrak{D}_k|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D}_k)} \propto p(\mathfrak{D}_k|\mathbf{w})p(\mathbf{w}),$$

$$p_{k+1}(\mathbf{w}) = p(\mathbf{w}|\mathfrak{D}_{k+1}) = \frac{p(\mathfrak{D}_{k+1}|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D}_{k+1})} \propto p(\mathfrak{D}_{k+1}|\mathbf{w})p(\mathbf{w}).$$

Анализ апостериорного распределения параметров модели

Определение (KL-достаточный размер выборки)

Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется **KL-достаточным**, если для всех $k \geqslant m^*$

$$KL(k) = D_{KL}(p_k \| p_{k+1}) = \int p_k(\mathbf{w}) \log \frac{p_k(\mathbf{w})}{p_{k+1}(\mathbf{w})} d\mathbf{w} \leqslant \varepsilon.$$

Определение (S-достаточный размер выборки)

Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется **S-достаточным**, если для всех $k \geqslant m^*$

$$S(k) = \text{s-score}(p_k, p_{k+1}) \geqslant 1 - \varepsilon.$$

$$\text{s-score}(g_1, g_2) = \frac{\int_{\mathbf{w}} g_1(\mathbf{w}) g_2(\mathbf{w}) d\mathbf{w}}{\max_{\mathbf{b}} \int_{\mathbf{w}} g_1(\mathbf{w} - \mathbf{b}) g_2(\mathbf{w}) d\mathbf{w}}$$

Motrenko A., Strijov V., Weber G-W. Sample size determination for logistic regression. Journal of Computational and Applied Mathematics, 2014.
Адуенко А. Выбор мультимоделей в задачах классификации. PhD thesis, МФТИ, 2017.

Анализ апостериорного распределения параметров модели

Предположим, что апостериорное распределение является нормальным, то есть $p_k(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_k, \mathbf{\Sigma}_k)$.

Теорема (Киселев, 2024)

Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ и $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_F \to 0$ при $k \to \infty$. Тогда в модели с нормальным апостериорным распределением параметров определение KL-достаточного размера выборки является корректным. А именно, для любого $\varepsilon > 0$ найдется такой m^*, что для всех $k \geqslant m^*$ выполнено $KL(k) \leqslant \varepsilon$.

Теорема (Киселев, 2024)

Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ при $k \to \infty$. Тогда в модели с нормальным апостериорным распределением параметров определение S-достаточного размера выборки является корректным. А именно, для любого $\varepsilon > 0$ найдется такой m^*, что для всех $k \geqslant m^*$ выполнено $S(k) \geqslant 1 - \varepsilon$.

Анализ апостериорного распределения параметров модели

Вероятностная модель линейной регрессии

$$p(\mathbf{y}, \mathbf{w}|\mathbf{X}) = p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}) = \mathcal{N}\left(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}\right)\mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right)$$

Апостериорное распределение

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}, \boldsymbol{\Sigma}\right),$$

$$\boldsymbol{\Sigma} = \left(\alpha\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}, \qquad \mathbf{m} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

Теорема (Киселев, 2024)

Пусть множества значений признаков и целевой переменной ограничены, то есть $\exists M \in \mathbb{R} : \|\mathbf{x}\|_2 \leqslant M$ и $|y| \leqslant M$. Если $\lambda_{\min}\left(\mathbf{X}_k^\mathsf{T}\mathbf{X}_k\right) = \omega(\sqrt{k})$ при $k \to \infty$, то в модели линейной регрессии с нормальным априорным распределением параметров $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ и $\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_k\|_F \to 0$ при $k \to \infty$.

Генетический алгоритм в задаче аппроксимации набора функций

Дано: зависимость среднего значения функции правдоподобия от используемого размера выборки для N различных датасетов.

Найти: параметрическое семейство функций, аппроксимирующее заданные зависимости.

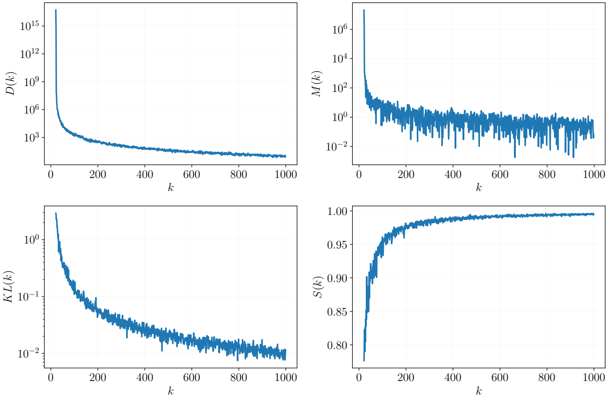
Критерий: минимизация среднеквадратичного отклонения.

Решение: генетический алгоритм, где

- Особь есть параметрическое семейство функций, представленное в виде дерева синтаксического разбора, причем каждому узлу дерева сопоставляется компонента вектора параметров;
- Популяция есть набор особей;
- Приспособленность есть среднее значение MSE по всем N зависимостям;
- Кроссинговер есть замена случайного поддерева.

Сходимость предложенных функций

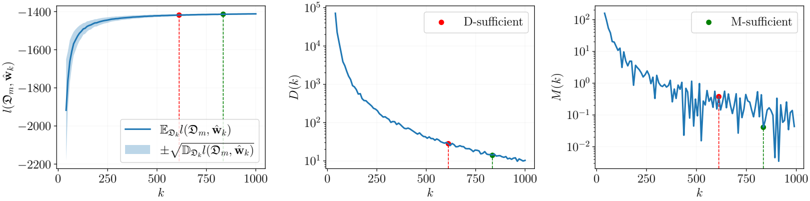
Синтетические данные сгенерированы из модели линейной регрессии. Число объектов 1000, число признаков 20. Из данной выборки последовательно удаляется по одному объекту. Такой процесс повторяется $B = 1000$ раз.



Функции стремятся к своим теоретическим пределам.

Определение достаточного размера выборки

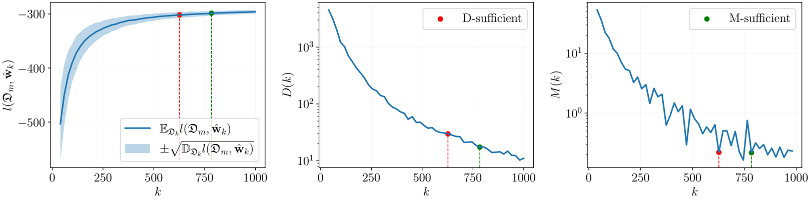
Синтетические данные сгенерированы из модели линейной регрессии. Число объектов 1000, число признаков 20. Из данной выборки последовательно удаляется по одному объекту. Такой процесс повторяется $B = 1000$ раз.



На графиках указаны D-достаточный и M-достаточный размеры выборки. Для D-достаточности выбрано $\varepsilon = 3 \cdot 10^1$, для M-достаточности $\varepsilon = 4 \cdot 10^{-1}$.

Определение достаточного размера выборки

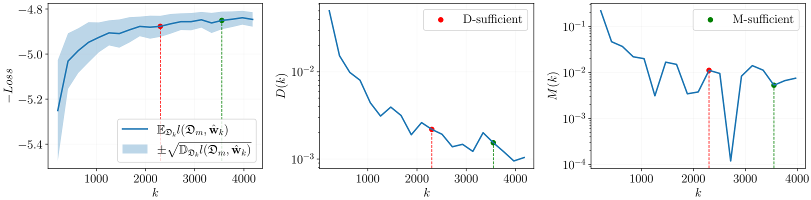
Синтетические данные сгенерированы из модели логистической регрессии. Число объектов 1000, число признаков 20. Из данной выборки последовательно удаляется по одному объекту. Такой процесс повторяется $B = 1000$ раз.



На графиках указаны D-достаточный и M-достаточный размеры выборки. Для D-достаточности выбрано $\varepsilon = 3 \cdot 10^1$, для M-достаточности $\varepsilon = 6 \cdot 10^{-1}$.

Определение достаточного размера выборки

Используется датасет Abalone с задачей регрессии из открытой библиотеки UCI[1]. Число объектов 4177, число признаков 8. Из данной выборки последовательно удаляется по одному объекту. Такой процесс повторяется $B = 1000$ раз.

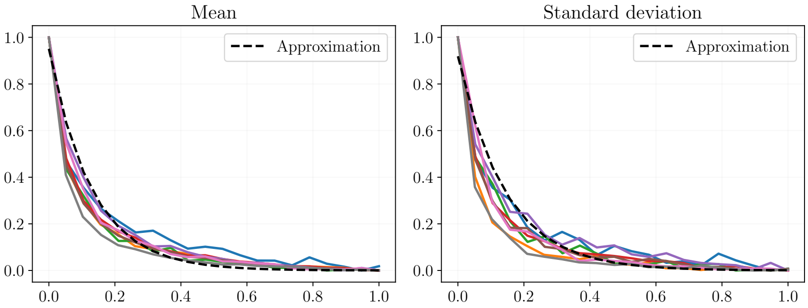


На графиках указаны D-достаточный и M-достаточный размеры выборки. Для D-достаточности выбрано $\varepsilon = 2.5 \cdot 10^{-3}$, для M-достаточности $\varepsilon = 8 \cdot 10^{-3}$.

[1] Markelle Kelly, Rachel Longjohn, Kolby Nottingham, The UCI Machine Learning Repository.

Определение параметрического семейства функций с помощью генетического алгоритма

Анализируются датасеты с задачей регрессии из открытой библиотеки UCI[2]. Усреднение проводится по $B = 100$ бутстрап-подвыборкам.



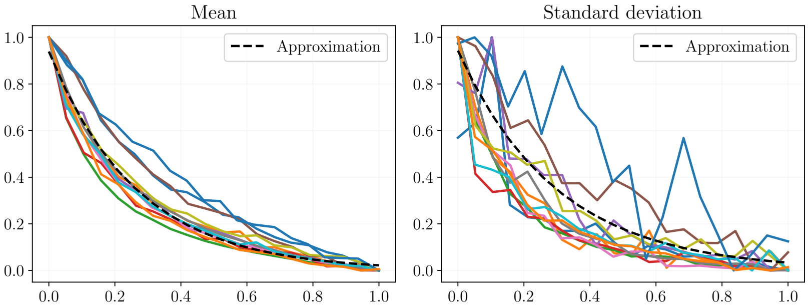
Применение генетического алгоритма приводит к семейству функций

$$w_0 + w_1 \cdot \exp(w_2 \cdot x).$$

[2] Markelle Kelly, Rachel Longjohn, Kolby Nottingham, The UCI Machine Learning Repository.

Определение параметрического семейства функций с помощью генетического алгоритма

Анализируются датасеты с задачей классификации из открытой библиотеки UCI[3]. Усреднение проводится по $B = 100$ бутстрап-подвыборкам.



Применение генетического алгоритма приводит к семейству функций

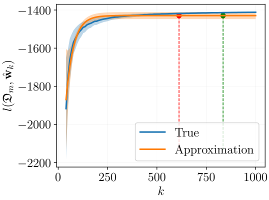
$$w_0 + w_1 \cdot \exp(w_2 \cdot x).$$

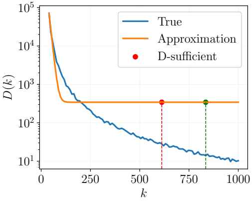
[3]Markelle Kelly, Rachel Longjohn, Kolby Nottingham, The UCI Machine Learning Repository.

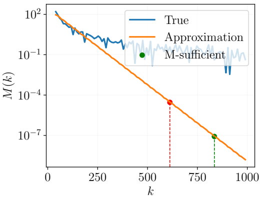
Прогнозирование функции правдоподобия

Среднее значение и дисперсия аппроксимированы параметрическим семейством функций, полученным ранее. Производилось разделение на обучающую и тестовую выборки в соотношении 70:30. Аппроксимация производилась только на обучающей части. Достаточный размер выборки находился в тестовой части.

Синтетическая выборка (линейная регрессия)

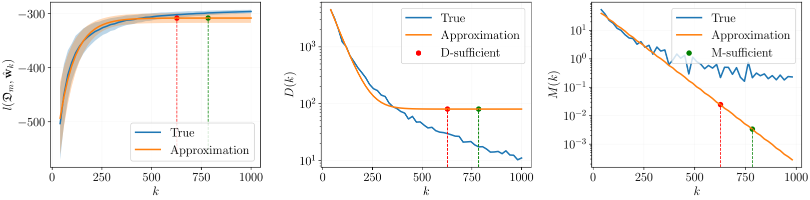




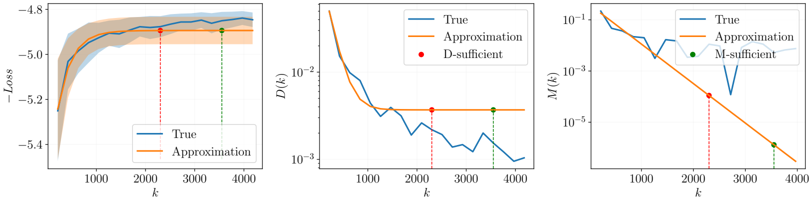


Прогнозирование функции правдоподобия

Синтетическая выборка (логистическая регрессия)



Выборка Abalone (регрессия)



Заключение

- Предложены подходы к определению достаточного размера выборки на основе функции правдоподобия и апостериорных распределений.
- Доказана корректность подходов при определенных ограничениях на модель.
- Предложен метод прогнозирования функции правдоподобия при недостаточном размере выборки.
- Проведен вычислительный эксперимент для анализа методов.
- Определено параметрическое семейство функций, аппроксимирующее функцию ошибки для различных датасетов.