Байесовский подход к выбору оптимального размера выборки

Выпускная квалификационная работа бакалавра

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

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных
Направление: 03.03.01 Прикладные математика и физика

2024

Байесовский подход к определению достаточного размера выборки

Исследуется задача определения 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

Предлагается провести исследование

1. Значений функции правдоподобия на бутстрапированных подвыборках;
2. Расстояния между апостериорными распределениями параметров модели на схожих подвыборках.

Постановка задачи определения достаточного размера выборки

Выборка

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

- $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^n$ — вектор признакового описания объекта;
- $y \in \mathbb{Y}$ — значение целевой переменной.

Вероятностная модель

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

- $p(y|\mathbf{x}, \mathbf{w})$ — правдоподобие;
- $p(\mathbf{w})$ — априорное распределение.

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

Функция правдоподобия

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

Оценка максимума правдоподобия

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

Определение (D-достаточный размер выборки)

$$\forall k \geqslant m^* : D(k) = \mathbb{D}_{\hat{\mathbf{w}}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k) \leqslant \varepsilon$$

Определение (M-достаточный размер выборки)

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

Корректность определения M-достаточного размера выборки

Обозначим параметры распределения $\hat{\mathbf{w}}_k$ следующим образом:

- Математическое ожидание $\mathbb{E}\hat{\mathbf{w}}_k = \mathbf{m}_k$;
- Матрица ковариации $\mathbb{D}\hat{\mathbf{w}}_k = \mathbf{\Sigma}_k$.

Теорема 1 (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 $M(k) \leqslant \varepsilon$.

Определение[1] (Схожие подвыборки)

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.

Подвыборки $\mathfrak{D}^1$ и $\mathfrak{D}^2$ называются схожими, если $|\mathcal{I}_1 \triangle \mathcal{I}_2| = 1$.

Апостериорные распределения на схожих подвыборках

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

$$\mathfrak{D}_{k+1} = (\mathbf{X}_{k+1}, \mathbf{y}_{k+1}) \to p_{k+1}(\mathbf{w}) = p(\mathbf{w}|\mathfrak{D}_{k+1}) \propto p(\mathfrak{D}_{k+1}|\mathbf{w})p(\mathbf{w})$$

Функция близости s-score[2]

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

[1] Anastasiya Motrenko, Vadim Strijov и Gerhard-Wilhelm Weber. **"Sample size determination for logistic regression".** В: *Journal of Computational and Applied Mathematics* 255 (2014), с. 743—752.

[2] Адуенко Александр Александрович. **"Выбор мультимоделей в задачах классификации".** Кандидатская диссертация. Москва: МФТИ, 2017.

Близость апостериорных распределений на схожих подвыборках

Определение (KL-достаточный размер выборки)

$$\forall k \geqslant m^* : KL(k) = D_{KL}(p_k \| p_{k+1}) = \int p_k(\mathbf{w}) \log \frac{p_k(\mathbf{w})}{p_{k+1}(\mathbf{w})} d\mathbf{w} \leqslant \varepsilon$$

Определение (S-достаточный размер выборки)

$$\forall k \geqslant m^* : S(k) = \text{s-score}(p_k, p_{k+1}) \geqslant 1 - \varepsilon$$

Корректность определений KL- и S-достаточного размера выборки

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

Теорема 2 (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 3 (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 $S(k) \geqslant 1 - \varepsilon$.

Модель линейной регрессии с нормальным априорным распределением

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

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

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

Теорема 4 (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 $\lambda_{\min}\left(\mathbf{X}_k^\intercal\mathbf{X}_k\right) = \omega(\sqrt{k})$ 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 $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_F \to 0$ при $k \to \infty$.

Разделение на обучающую и тестовую выборки

$$\mathfrak{D}_m = \mathfrak{D}_{m_1}^{\text{train}} \sqcup \mathfrak{D}_{m_2}^{\text{test}}$$

Подвыборка обучающей выборки

$$(\mathbf{X}_k, \mathbf{y}_k) \subset \mathfrak{D}_{m_1}^{\text{train}}$$

Теорема 5 (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

$$\left\|\mathbb{E}\left[\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{k+1}, \mathbf{y}_{k+1}\right] - \mathbb{E}\left[\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_k, \mathbf{y}_k\right]\right\|_2 \to 0,$$

$$\left\|\mathbb{D}\left[\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{k+1}, \mathbf{y}_{k+1}\right] - \mathbb{D}\left[\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_k, \mathbf{y}_k\right]\right\|_F \to 0,$$

$$D_{\text{KL}}\left(p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_k, \mathbf{y}_k) \| p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{k+1}, \mathbf{y}_{k+1})\right) \to 0.$$

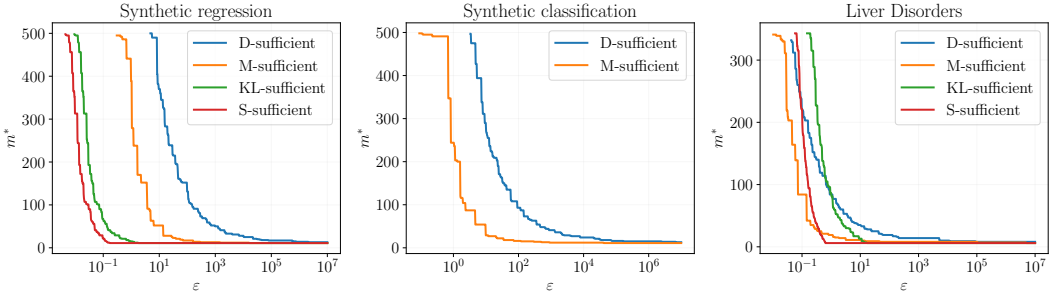
Вычислительный эксперимент

Достаточный размер выборки в зависимости от гиперпараметра ε

Используются выборки

- Синтетическая регрессия: 500 объектов, 10 признаков;
- Синтетическая классификация: 500 объектов, 10 признаков;
- Liver Disorders с задачей регрессии: 345 объектов, 5 признаков.

Для каждого значения ε определяется достаточный размер выборки.



Сравнение подходов на множестве выборок с задачей регрессии

Определяется размер выборки, при котором значение метрики уменьшается в

- 1000 раз для D- и M-достаточного размера выборки;
- 2 раза для KL- и S-достаточного размера выборки.

Пропуски означают, что первоначальный размер выборки недостаточен.

Название выборки	Объектов m	Признаков n	D	M	KL	S
Abalone	4177	8	96	96	3921	4091
Auto MPG	392	8	15	15	62	—
Automobile	159	25	70	156	156	—
Liver Disorders	345	6	12	19	—	—
Servo	167	4	41	—	163	163
Forest fires	517	12	208	—	507	—
Wine Quality	6497	12	144	144	5305	6099
Energy Efficiency	768	9	24	442	—	—
Student Performance	649	32	129	177	636	—
Facebook Metrics	495	18	31	388	475	—
Real Estate Valuation	414	7	15	23	—	—
Heart Failure Clinical Records	299	12	63	224	276	293
Bone marrow transplant: children	142	36	—	—	109	—

Выносится на защиту

1. Подходы к определению достаточного размера выборки по
   - Сходимости функции правдоподобия на бутстрапированных подвыборках;
   - Близости апостериорных распределений параметров на схожих подвыборках;
2. Теоремы о корректности определений
   - M-достаточного размера выборки в модели линейной регрессии;
   - KL-достаточного размера выборки в модели с нормальным апостериорным распределением параметров;
   - S-достаточного размера выборки в модели с нормальным апостериорным распределением параметров;
3. Теорема о близости моментов предельного апостериорного распределения в модели линейной регрессии с нормальным априорным распределением параметров;
4. Теорема о близости полных байесовских прогнозов в модели линейной регрессии с нормальным априорным распределением параметров.

Список работ автора по теме диплома

Публикации ВАК

1. ***N. Kiselev***, *A. Grabovoy.* Sample Size Determination: Posterior Distributions Proximity // Computational Management Science (на рецензировании).

Выступления с докладом

1. Определение достаточного размера выборки по апостериорному распределению параметров модели // 66-я Всероссийская научная конференция МФТИ, 2024.