

Байесовский подход к выбору оптимального размера выборки

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

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

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

Вероятностная модель

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

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

Определение (Достаточный размер выборки)

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

Требуется

Предложить критерий T определения достаточного размера выборки.

Анализ поведения функции правдоподобия

Функция правдоподобия

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

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)

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

Определение (**M**-достаточный размер выборки)

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

Корректность определения М-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 $\mathbb{D}\hat{\mathbf{w}}_k = \Sigma_k$.

Теорема 1 (Киселев, 2023)

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

Анализ апостериорного распределения параметров модели

Определение¹ (Схожие подвыборки)

Подвыборки $\mathfrak{D}^1 \subseteq \mathfrak{D}_m$ и $\mathfrak{D}^2 \subseteq \mathfrak{D}_m$ называются схожими, если $|\mathfrak{D}^1 \triangle \mathfrak{D}^2| = 1$, то есть они отличаются на один объект.

Апостериорные распределения на схожих подвыборках

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

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

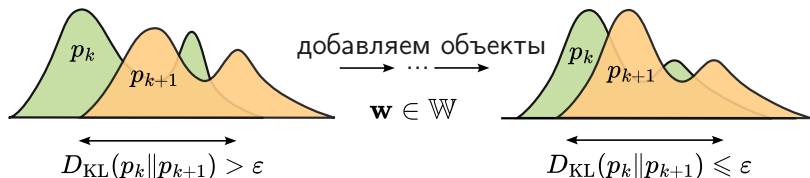
Функция близости s-score²

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

¹Anastasiya 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.

²Александр Александрович Адуенко. “Выбор мультимodelей в задачах классификации”. Кандидатская диссертация. Москва: МФТИ, 2017.

Близость апостериорных распределений на схожих подвыборках



Определение (KL-достаточный размер выборки)

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

Определение (S-достаточный размер выборки)

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

Корректность определений KL- и S-достаточного размера выборки

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

Теорема 2 (Киселев, 2024)

Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$ и $\|\Sigma_{k+1} - \Sigma_k\|_F \rightarrow 0$ при $k \rightarrow \infty$. 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 $k \geq m^*$ выполнено $KL(k) \leq \varepsilon$.

Теорема 3 (Киселев, 2024)

Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$ при $k \rightarrow \infty$. 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 $k \geq m^*$ выполнено $S(k) \geq 1 - \varepsilon$.

Модель линейной регрессии с нормальным априорным распределением

Вероятностная модель линейной регрессии

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

Апостериорное распределение

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

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

Теорема 4 (Киселев, 2024)

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

Близость полных байесовских прогнозов для линейной регрессии

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 $\exists M \in \mathbb{R} : \|\mathbf{x}\|_2 \leq M$ и $|y| \leq M$. Если $\lambda_{\min}(\mathbf{X}_k^T \mathbf{X}_k) = \omega(\sqrt{k})$ при $k \rightarrow \infty$, то в модели линейной регрессии с нормальным априорным распределением параметров

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

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

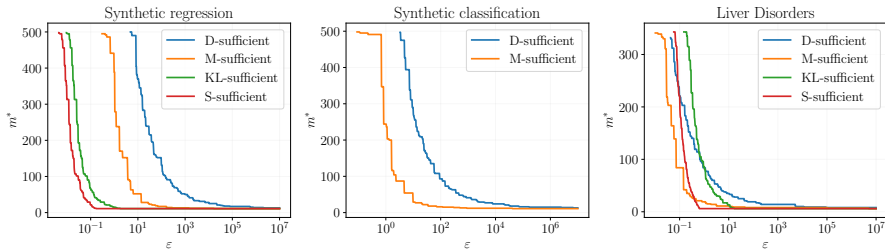
$$D_{\text{KL}}(p(\mathbf{y}_{\text{test}} | \mathbf{X}_{\text{test}}, \mathbf{X}_k, \mathbf{y}_k) \| p(\mathbf{y}_{\text{test}} | \mathbf{X}_{\text{test}}, \mathbf{X}_{k+1}, \mathbf{y}_{k+1})) \rightarrow 0.$$

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.



В синтетической регрессии D- и M-достаточность требуют больших значений ε . Тем не менее, сравнивать между собой подходы можно только попарно.

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 в таблице 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 | — |

При попарном сравнении M- и S-достаточные размеры выборки в среднем оказываются больше.

Выносятся на защиту

1. Подходы к определению достаточного размера выборки по значениям функции правдоподобия и близости апостериорных распределений;
2. Теоремы о корректности предложенных определений;
3. Теоремы о близости моментов и полных байесовских прогнозов предельного апостериорного распределения.

Список работ автора по теме диплома

1. **N. Kiselev**, A. Grabovoy. Sample Size Determination: Posterior Distributions Proximity // Computational Management Science (на рецензировании).
2. Определение достаточного размера выборки по апостериорному распределению параметров модели // 66-я Всероссийская научная конференция МФТИ, 2024.