Байесовский подход к выбору достаточного размера выборки

Киселев Никита
kiselev.ns@phystech.edu

Грабовой Андрей
grabovoy.av@phystech.edu

13 декабря 2023 г.

Аннотация

Исследуется задача выбора достаточного размера выборки. Рассматривается проблема определения достаточного размера выборки без учета природы параметров используемой модели. Предлагается использовать функцию правдоподобия выборки. Используются подходы на основе эвристик о поведении функции правдоподобия при достаточном количестве объектов в выборке. Проводится вычислительный эксперимент для анализа свойств предложенных методов.

Ключевые слова: определение размера выборки · байесовский подход

1 Введение

Задача машинного обучения с учителем предполагает выбор предсказательной модели из некоторого параметрического семейства. Обычно такой выбор связан с некоторыми статистическими гипотезами, например, максимизацией некоторого функционала качества.

Определение 1. *Модель прогнозирования, которая соответствует этим статистическим гипотезам, называется* ***адекватной*** *моделью.*

При проведении эксперимента зачастую дана конечная обучающая выборка.

Определение 2. *Размер выборки, необходимый для построения адекватной модели прогнозирования, называется* ***достаточным****.*

В работе [1] представлены десять методов для оценки достаточного размера выборки. Среди них есть как статистические, так и байесовские подходы.

2 Постановка задачи

Задана выборка размера m:

$$\mathfrak{D}_m = \{\mathbf{x}_i, y_i\}_{i=1}^m,$$

где $\mathbf{x}_i \in \mathbb{X}, y_i \in \mathbb{Y}$.

Введем параметрическое семейство $p(y|\mathbf{x}, \mathbf{w})$ для аппроксимации неизвестного апостериорного распределения $p(y|\mathbf{x})$ целевой переменной y при известных признаковом описании объекта $\mathbf{x}$ и параметрах $\mathbf{w} \in \mathbb{W}$.

Определим функцию правдоподобия и логарифмическую функцию правдоподобия выборки $\mathfrak{D}_m$:

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

Оценим параметры, используя метод максимума правдоподобия:

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

Требуется определить достаточный размер выборки m^*. При этом понятие достаточности может определяться различными способами. Часто оно дается в терминах функции правдоподобия и полученной из ее максимизации оценки параметров. Также стоит учесть, что возможно $m^* \leqslant m$ или $m^* > m$. Эти два случая будут отдельно рассмотрены далее.

3 Достаточный размер выборки не превосходит доступный

В этой главе будем считать, что достоверно $m^* \leqslant m$.

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

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

Поскольку природа $\mathbf{w}$ нам неизвестна, для определения достаточности будем использовать функцию правдоподобия.

Когда в наличии имеется достаточно объектов, вполне естественно ожидать, что от одной реализации выборки к другой полученная оценка параметров не будет сильно меняться [2, 3]. То же можно сказать и про функцию правдоподобие. Таким образом, сформулируем, какой размер выборки можно считать достаточным.

Определение 3. *Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется* ***D-достаточным****, если для любого $k \geqslant m^*$*

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

Замечание. *В определении 3 вместо функции правдоподобия* $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$ *можно рассматривать ее логарифм* $l(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$.

С другой стороны, когда в наличии имеется достаточно объектов, также вполне естественно, что при добавлении очередного объекта в рассмотрение полученная оценка параметров не будет сильно меняться. Сформулируем еще одно определение.

Определение 4. *Зафиксируем некоторое положительное число* $\varepsilon > 0$. *Размер выборки* m^* *называется* ***M-достаточным****, если для любого* $k \geqslant m^*$

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

Замечание. *В определении 4 вместо функции правдоподобия* $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$ *можно рассматривать ее логарифм* $l(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$.

Как доказать корректность этих определений? А именно, почему такой размер выборки существует?

По условию задана одна выборка. Поэтому в эксперименте нет возможности посчитать указанные в определениях математическое ожидание и дисперсию. Для их оценки воспользуемся техникой бутстрэп. А именно, сгенерируем из заданной $\mathfrak{D}_m$ некоторое число B подвыборок размера k с возвращением. Для каждой из них получим оценку параметров $\hat{\mathbf{w}}_k$ и посчитаем значение $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$. Для оценки будем использовать выборочное среднее и несмещенную выборочную дисперсию (по бутстрэп-выборкам). Как доказать «хорошие» свойства этих оценок?

4 Достаточный размер выборки больше доступного

В этой главе будем считать, что достоверно $m^* > m$.

Возникает задача прогнозирования математического ожидания и функции правдоподобия при $k > m$. Как определить характер этой зависимости?

5 Вычислительный эксперимент

Проводится эксперимент для анализа свойств предложенных методов оценки достаточного размера выборки. Эксперимент состоит из двух частей. В первой части рассматриваются оценки достаточного размера выборки в случае, когда достаточный размер выборки не превосходит доступный. Во второй части исследуются результаты, полученные в условиях того, что достаточный размер выборки больше доступного.

5.1 Достаточный размер выборки не превосходит доступный

Синтетические данные сгенерированы из модели линейной регрессии. Число объектов 1000, число признаков 20. Далее приведены графики логарифма функции правдоподобия выборки, а также функций $D(k)$ и $M(k)$, определенных в Главе 3 (здесь используется логарифм функции правдоподобия). Выполнено определение D-достаточного

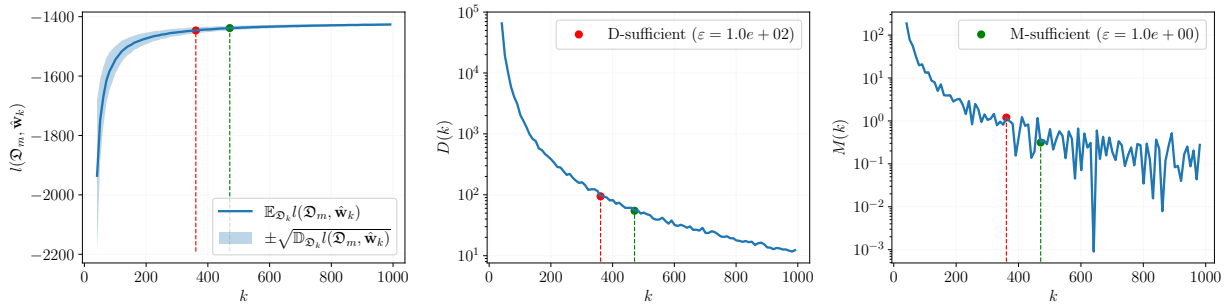

Рис. 1: Синтетическая выборка (линейная регрессия) при $m^* \leqslant m$

и M-достаточного размеров выборки. Использовалось $B = 1000$ бутстрэп-выборок. Результаты представлены на Рис. 1.

Вторая синтетическая выборка сгенерирована из модели логистической регрессии. Число объектов 1000, число признаков 20. Аналогичные графики приведены на Рис. 2.

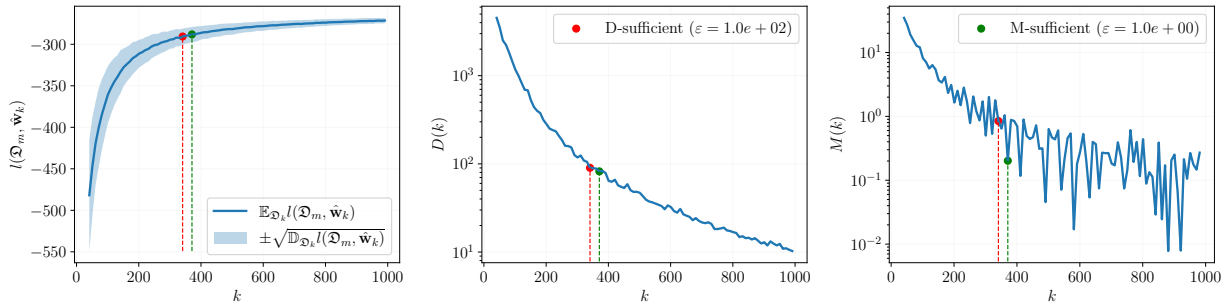

Рис. 2: Синтетическая выборка (логистическая регрессия) при $m^* \leqslant m$

5.2 Достаточный размер выборки больше доступного

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.

Производилось разделение на обучающую и тестовую выборки в соотношении 70:30. Аппроксимация производилась только на обучающей части. Достаточный размер выборки находился в тестовой части. На Рис. 3 и Рис. 4 представлены истинные и восстановленные зависимости. Там же указаны определенные D-достаточный и M-достаточный размеры выборки.

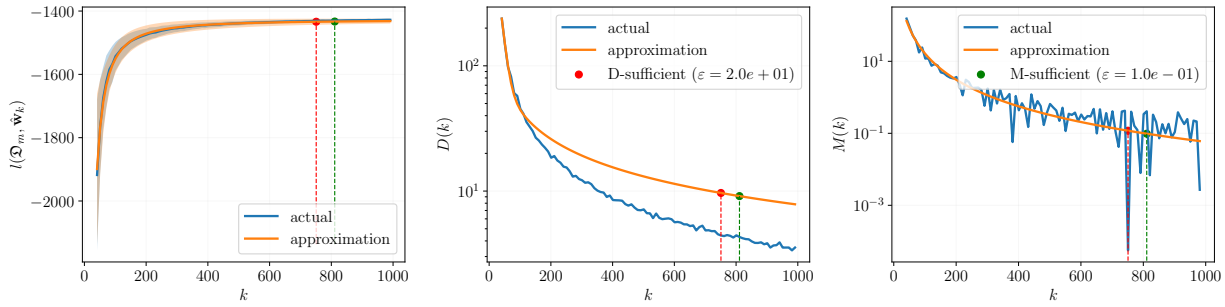

Рис. 3: Синтетическая выборка (линейная регрессия) при $m^* > m$

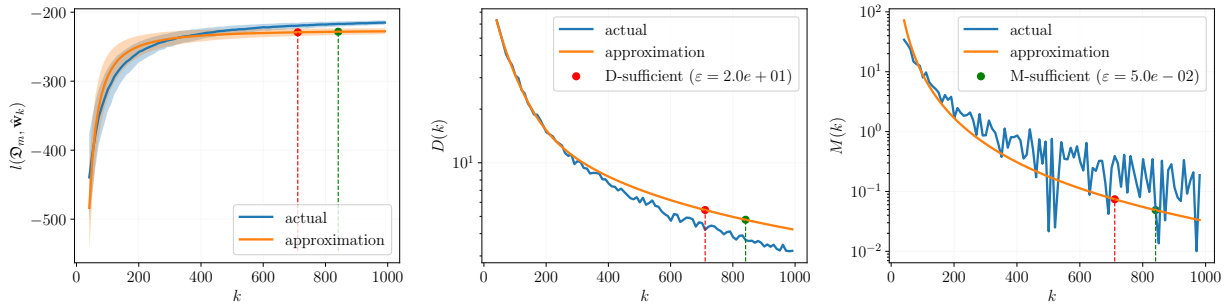

Рис. 4: Синтетическая выборка (логистическая регрессия) при $m^* > m$

6 Заключение

Основные результаты данной работы заключаются в следующем.

Бла-бла-бла.

Список литературы

[1] A. V. Grabovoy, T. T. Gadaev, A. P. Motrenko, and V. V. Strijov. Numerical methods of sufficient sample size estimation for generalised linear models. *Lobachevskii Journal of Mathematics*, 43(9):2453–2462, Sept. 2022.

[2] L. Joseph, R. D. Berger, and P. Bélisle. Bayesian and mixed bayesian/likelihood criteria for sample size determination. *Statistics in Medicine*, 16(7):769–781, 1997.

[3] L. Joseph, D. B. Wolfson, and R. D. Berger. Sample size calculations for binomial proportions via highest posterior density intervals. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 44(2):143–154, 1995.