Байесовский подход к выбору достаточного размера выборки

Киселев Никита
kiselev.ns@phystech.edu

Грабовой Андрей
grabovoy.av@phystech.edu

15 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] рассматриваются различные методы оценки объема выборки в обобщенных линейных моделях, включая статистические, эвристические и байесовские методы. Анализируются такие методы, как тест на множители Лагранжа, тест на отношение правдоподобия, статистика Вальда, кросс-валидация, бутстрап, критерий Куллбэка-Лейблера,

критерий средней апостериорной дисперсии, критерий среднего охвата, критерий средней длины и максимизация полезности. Авторы статьи указывают на возможное развитие темы, которое заключается в поиске метода, сочетающего байесовский и статистический подходы для оценки размера выборки для недостаточного доступного размера выборки.

В [4] рассматривается новый метод определения размера выборки в логистической регрессии. Метод использует кросс-валидацию и дивергенцию Кульбака-Лейблера между апостериорными распределениями параметров модели на схожих подвыборках.

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 $\mathfrak{D}_m$:

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

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?

Предположим, что $\mathbb{W} = \mathbb{R}^n$, т.е. параметры $\mathbf{w}$ представляются в виде вектора. Напомним, что информацией Фишера называется матрица

$$[\mathcal{I}(\mathbf{w})]_{ij} = -\mathbb{E}\left[\frac{\partial^2 \log p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{\partial w_i \partial w_j}\right].$$

Достаточно известным результатом является асимптотическая нормальность оценки максимума правдоподобия.

Утверждение 1. *Пусть $\hat{\mathbf{w}}_k$ — оценка максимума правдоподобия $\mathbf{w}$. Тогда при определенных условиях регулярности (которые на практике чаще всего выполнены) имеет место следующая сходимость по распределению:*

$$\sqrt{m}\,(\hat{\mathbf{w}}_k - \mathbf{w}) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}^{-1}(\mathbf{w})\right),$$

или, что равносильно,

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

Из сходимости по распределению в общем случае не следует сходимость моментов случайного вектора. Тем не менее, если предположить последнее, то в некоторых моделях

можно доказать корректность предложенного нами определения M-достаточного размера выборки.

Для удобства обозначим параметры распределения $\hat{\mathbf{w}}_k$ следующим образом: математическое ожидание $\mathbb{E}_{\mathfrak{D}_k}\hat{\mathbf{w}}_k = \mathbf{m}_k$ и матрица ковариации $\mathrm{cov}(\hat{\mathbf{w}}_k) = \mathbf{\Sigma}_k$. Тогда имеет место следующая лемма.

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* $M(k) \leqslant \varepsilon$.

По условию задана одна выборка. Поэтому в эксперименте нет возможности посчитать указанные в определениях математическое ожидание и дисперсию. Для их оценки воспользуемся техникой бутстрап. 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 бутстрап-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 и M-достаточного размеров выборки. Использовалось $B = 1000$ бутстрап-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.

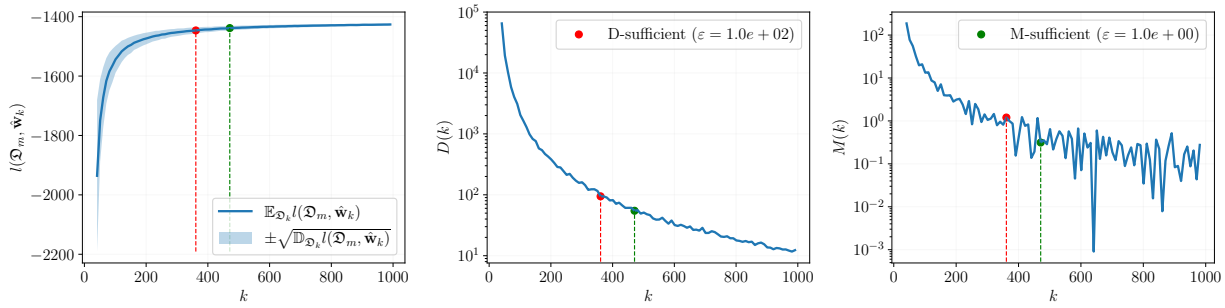

Рис. 1: Синтетическая выборка (линейная регрессия) при $m^* \leqslant m$

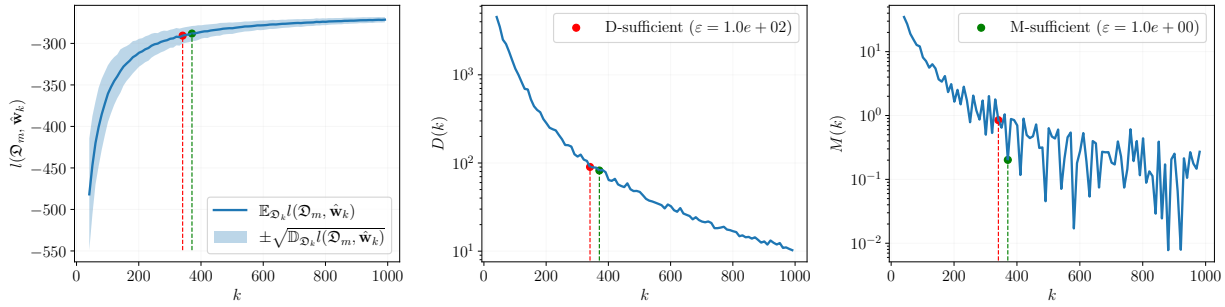

Рис. 2: Синтетическая выборка (логистическая регрессия) при $m^* \leqslant m$

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.

6 Заключение

Основные результаты данной работы заключаются в следующем.

Бла-бла-бла.

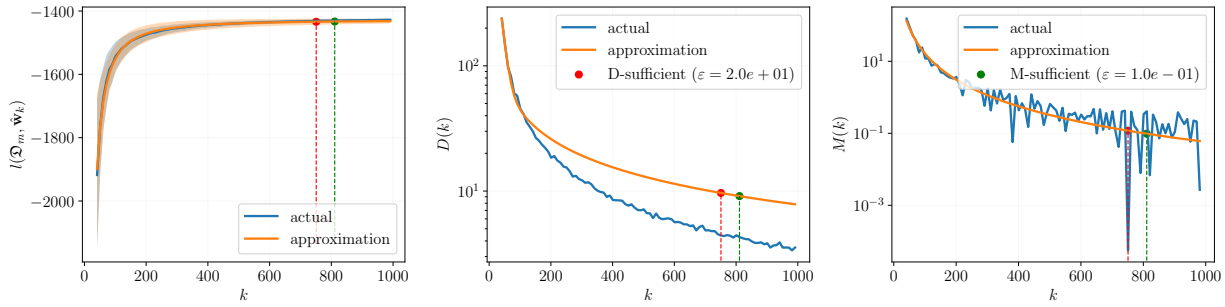

Рис. 3: Синтетическая выборка (линейная регрессия) при $m^* > m$

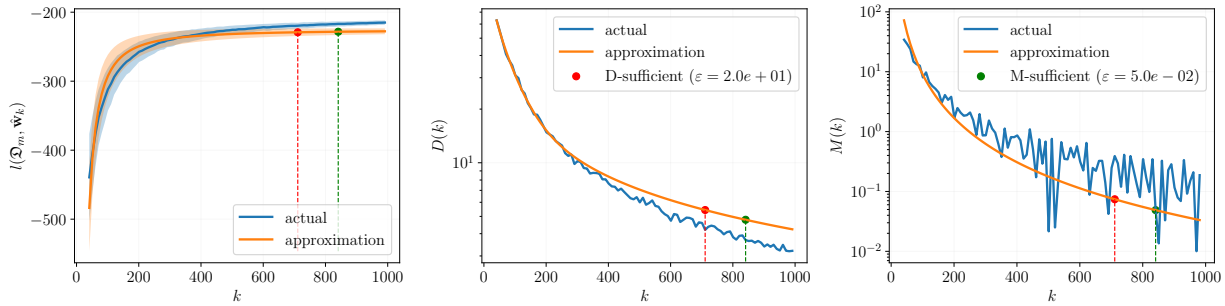

Рис. 4: Синтетическая выборка (логистическая регрессия) при $m^* > m$

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.

[4] A. Motrenko, V. Strijov, and G.-W. Weber. Sample size determination for logistic regression. *Journal of Computational and Applied Mathematics*, 255:743–752, 2014.

7 Приложение

Доказательство (Лемма 1). Рассмотрим определение М-достаточного размера выборки в терминах логарифма функции правдоподобия. В модели линейной регрессии

$$L\left(\mathfrak{D}_m, \hat{\mathbf{w}}_k\right) = p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}_k) = \prod_{i=1}^{m} p(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_k) = \prod_{i=1}^{m} \mathcal{N}\left(y_i|\hat{\mathbf{w}}_k^{\mathsf{T}}\mathbf{x}_i, \sigma^2\right) =$$

$$= \left(2\pi\sigma^2\right)^{-m/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_k\|_2^2\right).$$

Прологарифмируем:

$$l\left(\mathfrak{D}_m, \hat{\mathbf{w}}_k\right) = \log p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}_k) = -\frac{m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_k\|_2^2.$$

Возьмем математическое ожидание по $\mathfrak{D}_k$, учитывая, что $\mathbb{E}_{\mathfrak{D}_k}\hat{\mathbf{w}}_k = \mathbf{m}_k$ и $\operatorname{cov}(\hat{\mathbf{w}}_k) = \boldsymbol{\Sigma}_k$:

$$\mathbb{E}_{\mathfrak{D}_k} l\left(\mathfrak{D}_m, \hat{\mathbf{w}}_k\right) = -\frac{m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\Big(\|\mathbf{y} - \mathbf{X}\mathbf{m}_k\|_2^2 + \operatorname{tr}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\Sigma}_k\right)\Big).$$

Запишем выражение для разности математических ожиданий:

$$\mathbb{E}_{\mathfrak{D}_{k+1}} l(\mathfrak{D}_m, \hat{\mathbf{w}}_{k+1}) - \mathbb{E}_{\mathfrak{D}_k} l(\mathfrak{D}_m, \hat{\mathbf{w}}_k) =$$

$$= \frac{1}{2\sigma^2}\Big(\|\mathbf{y} - \mathbf{X}\mathbf{m}_k\|_2^2 - \|\mathbf{y} - \mathbf{X}\mathbf{m}_{k+1}\|_2^2\Big) + \frac{1}{2\sigma^2}\operatorname{tr}\Big(\mathbf{X}^{\mathsf{T}}\mathbf{X}\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\right)\Big) =$$

$$= \frac{1}{2\sigma^2}\Big(2\mathbf{y}^{\mathsf{T}}\mathbf{X}(\mathbf{m}_{k+1} - \mathbf{m}_k) + (\mathbf{m}_k - \mathbf{m}_{k+1})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\Big) +$$

$$+ \frac{1}{2\sigma^2}\operatorname{tr}\Big(\mathbf{X}^{\mathsf{T}}\mathbf{X}\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\right)\Big).$$

Значение функции $M(k)$ есть модуль от вышеприведенного выражения. Применим неравенство треугольника для модуля, а затем оценим каждое слагаемое.
Первое слагаемое оценим, используя неравенство Коши-Буняковского:

$$\left|\mathbf{y}^{\mathsf{T}}\mathbf{X}(\mathbf{m}_{k+1} - \mathbf{m}_k)\right| \leqslant \|\mathbf{X}^{\mathsf{T}}\mathbf{y}\|_2\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2.$$

Второе слагаемое оценим, используя неравенство Коши-Буняковского, свойство согласованности спектральной матричной нормы, а также ограниченность последовательности векторов $\mathbf{m}_k$, которая следует из предъявленной в условии сходимости:

$$\left|(\mathbf{m}_k - \mathbf{m}_{k+1})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\right| \leqslant \|\mathbf{X}(\mathbf{m}_k - \mathbf{m}_{k+1})\|_2\|\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\|_2 \leqslant$$

$$\leqslant \|\mathbf{X}\|_2^2\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2\|\mathbf{m}_k + \mathbf{m}_{k+1}\|_2 \leqslant C\|\mathbf{X}\|_2^2\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2.$$

Последнее слагаемое оценим, используя суб-мультипликативность нормы Фробениуса:

$$\left|\operatorname{tr}\Big(\mathbf{X}^{\mathsf{T}}\mathbf{X}\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\right)\Big)\right| \leqslant \|\mathbf{X}^{\mathsf{T}}\mathbf{X}\|_F\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\|_F.$$

Наконец, из приведенных в условии сходимостей следует, что $\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \to 0$ и $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\|_F \to 0$ при $k \to \infty$. Таким образом, $M(k) \to 0$ при $k \to \infty$, что доказывает лемму.

□