ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2025, том 65, № 2, с. 235–242

ИНФОРМАТИКА

УДК 519.21

ДОСТАТОЧНЫЙ РАЗМЕР ВЫБОРКИ: БУТСТРАПИРОВАНИЕ ПРАВДОПОДОБИЯ

© 2025 г. Н. С. Киселев[1,*], А. В. Грабовой[1,]**

[1] 141701 Долгопрудный, М.о., Институтский пер., 9, МФТИ, Россия
**e-mail: kiselev.ns@phystech.edu*
***e-mail: grabovoy.av@phystech.edu*

Поступила в редакцию 06.10.2024 г.
Переработанный вариант 06.10.2024 г.
Принята к публикации 10.11.2024 г.

Определение подходящего размера выборки имеет решающее значение для построения эффективных моделей машинного обучения. Существующие методы часто либо не имеют строгого теоретического обоснования, либо привязаны к конкретным статистическим гипотезам о параметрах модели. В настоящей работе представляются два новых метода, основанных на значениях правдоподобия на бутстрапированных подвыборках. Демонстрируется корректность одного из этих методов на в модели линейной регрессии. Вычислительные эксперименты как с синтетическими, так и с реальными наборами данных показывают, что предложенные функции сходятся по мере увеличения размера выборки, что подчеркивает практическую полезность подхода. Библ. 13. Фиг. 4. Табл. 1.

Ключевые слова: достаточный размер выборки, бутстрапирование правдоподобия, линейная регрессия, вычислительная линейная алгебра.

DOI: 10.31857/S0044466925020094, **EDN**: CBDKTA

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. Модель, которая соответствует этим статистическим гипотезам, называется *адекватной* моделью [1, 2].

При планировании вычислительного эксперимента требуется оценить минимальный размер выборки — количество объектов, необходимое для построения адекватной модели. Размер выборки, необходимый для построения адекватной модели прогнозирования, называется *достаточным* [3–5].

В работе рассматривается проблема определения достаточного размера выборки. Этой теме посвящено большое число работ. Используемые в них подходы можно разделить на статистические, байесовские и эвристические.

Одни из первых исследований по данной теме [6, 7] формулируют определенный статистический критерий, где связанный с данным критерием метод оценки размера выборки гарантирует достижение фиксированной статистической мощности с величиной ошибки I рода, не превышающей заданного значения. К статистическим методам относятся метод множителей Лагранжа [8], метод проверки отношения правдоподобия [9], метод Вальда [10]. Статистические методы имеют ряд ограничений, которые связаны с их применением на практике. Они позволяют оценить размер выборки, исходя из предположений о распределении данных и информации о соответствии наблюдаемых величин предположениям нулевой гипотезы.

Байесовский подход тоже имеет место в данной проблеме. В работе [11] достаточный размер выборки определяется исходя из максимизации ожидаемой функции полезности. Она может включать в себя в явном виде функции распределения параметров и штрафы за увеличение размера выборки. Также в этой работе рассматриваются альтернативные подходы, основанные на ограничении некоторого критерия качества оценки параметров модели. Среди таких критериев можно выделить критерий средней апостериорной дисперсии (APVC), критерий среднего покрытия (ACC), критерий средней длины (ALC) и критерий эффективного объема выборки (ESC). Эти критерии получили свое развитие в других работах, например, [12] и [13]. Спустя время, в [14] было проведено теоретическое и практическое сравнение методов из [6, 7, 11].

В работе [15], как и в [16], рассматриваются различия между байесовским и частотным подходами при определении размера выборки. Также предлагаются робастные методы для байесовского подхода и приводятся наглядные примеры для некоторых вероятностных моделей.

В работе [17] рассматриваются различные методы оценки размера 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 Кульбака–Лейблера, критерий средней апостериорной дисперсии, критерий среднего охвата, критерий средней длины и максимизация полезности. Указывается на возможное развитие темы, которое заключается в поиске метода, сочетающего байесовский и статистический подходы для оценки размера выборки для недостаточного доступного размера выборки.

В [18] рассматривается метод определения размера выборки в логистической регрессии, использующий кросс-валидацию и дивергенцию Кульбака–Лейблера между апостериорными распределениями параметров модели на схожих подвыборках. Под схожими подвыборками понимают такие подвыборки, которые могут быть получены друг из друга добавлением, удалением или заменой одного объекта.

В настоящей работе рассматриваются несколько подходов к определению достаточного размера выборки. Предлагается оценивать математическое ожидание и дисперсию функции правдоподобия на бутстрапированных подвыборках. Малое изменение этих величин при добавлении очередного объекта свидетельствует о достижении достаточного числа объектов в выборке. Доказывается корректность определения в модели линейной регрессии. Представленный метод легко использовать и на практике. Для этого предлагается подсчитывать значение функции ошибки, а не правдоподобия.

2. ПОСТАНОВКА ЗАДАЧИ

Объектом называется пара $(\mathbf{x}, y)$, где $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^n$ есть вектор признакового описания объекта, а $y \in \mathbb{Y}$ есть значение целевой переменной. В задаче регрессии $\mathbb{Y} = \mathbb{R}$, а в задаче K-классовой классификации $\mathbb{Y} = \{1, \ldots, K\}$.

Матрицей объекты-признаки для выборки $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$, размера m называется матрица $\mathbf{X}_m = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^\intercal \in \mathbb{R}^{m \times n}$.

Вектором ответов (вектором значений целевой переменной) для выборки $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$, размера m называется вектор $\mathbf{y}_m = \left[y_1, \ldots, y_m\right]^\intercal \in \mathbb{Y}^m$.

Моделью называется параметрическое семейство функций f, отображающих декартово произведение множества значений признакового описания объектов $\mathbb{X}$ и множества значений параметров $\mathbb{W}$ во множество значений целевой переменной $\mathbb{Y}$:

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

Вероятностной моделью называется совместное распределение вида

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

где $\mathbf{w} \in \mathbb{W}$ есть набор параметров модели, $p(y|\mathbf{x}, \mathbf{w})$ задает правдоподобие объекта, а $p(\mathbf{w})$ задает априорное распределение параметров.

Функцией правдоподобия простой выборки $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$, размера m называется функция

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

Ее логарифм

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

называется *логарифмической функцией правдоподобия*. Далее, если не оговорено противное, будем считать выборку простой.

Оценкой максимума правдоподобия набора параметров $\mathbf{w} \in \mathbb{W}$ по подвыборке $\mathcal{D}_k$ размера k называется

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

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

3. ПРЕДЛАГАЕМЫЕ МЕТОДЫ ОПРЕДЕЛЕНИЯ ДОСТАТОЧНОГО РАЗМЕРА ВЫБОРКИ

В этом разделе будем считать, что достоверно $m^* \leqslant m$. Это означает, что нам нужно просто формализовать, какой размер выборки можно считать достаточным. 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 [7, 19]. То же можно сказать и про функцию правдоподобия. Таким образом, сформулируем, какой размер выборки можно считать достаточным.

Определение 2. Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется *D-достаточным*, если для всех $k \geqslant m^*$

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

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

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

В определениях выше вместо функции правдоподобия $L(\mathcal{D}_m, \hat{\mathbf{w}}_k)$ можно рассматривать ее логарифм $l(\mathcal{D}_m, \hat{\mathbf{w}}_k)$.

Предположим, что $\mathbb{W} = \mathbb{R}^n$. Напомним, что информацией Фишера называется матрица

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

Известным результатом является асимптотическая нормальность оценки максимума правдоподобия, то есть $\sqrt{k}\,(\hat{\mathbf{w}}_k - \mathbf{w}) \xrightarrow{d} \mathcal{N}\left(0, \mathbf{I}^{-1}(\mathbf{w})\right)$. 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 определения M-достаточного размера выборки.

Для удобства обозначим параметры распределения $\hat{\mathbf{w}}_k$ следующим образом: математическое ожидание $\mathbb{E}\hat{\mathbf{w}}_k = \mathbf{m}_k$ и матрица ковариации $\mathbb{D}\hat{\mathbf{w}}_k = \mathbf{\Sigma}_k$. Тогда имеет место следующая теорема, доказательство которой приведено в Приложении.

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

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 $\mathcal{D}_m$ некоторое число B подвыборок размера k с возвращением. Для каждой из них получим оценку параметров $\hat{\mathbf{w}}_k$ и посчитаем значение $L(\mathcal{D}_m, \hat{\mathbf{w}}_k)$. Для оценки будем использовать выборочное среднее и несмещенную выборочную дисперсию (по бутстрап-выборкам).

Предложенные выше определения можно применять и в тех задачах, когда минимизируется произвольная функция потерь, а не максимизируется функция правдоподобия. Мы не приводим никаких теоретических обоснований этого, однако на практике такая эвристика оказывается достаточно удачной.

4. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

В данном разделе проводится эмпирическое исследование предлагаемых методов. Эксперименты проводятся на синтетических данных и на выборке Liver Disorders из библиотеки [20]. Полностью воспроизводимый код экспериментов доступен в GitHub репозитории (https://github.com/kisnikser/Likelihood-Bootstrapping).

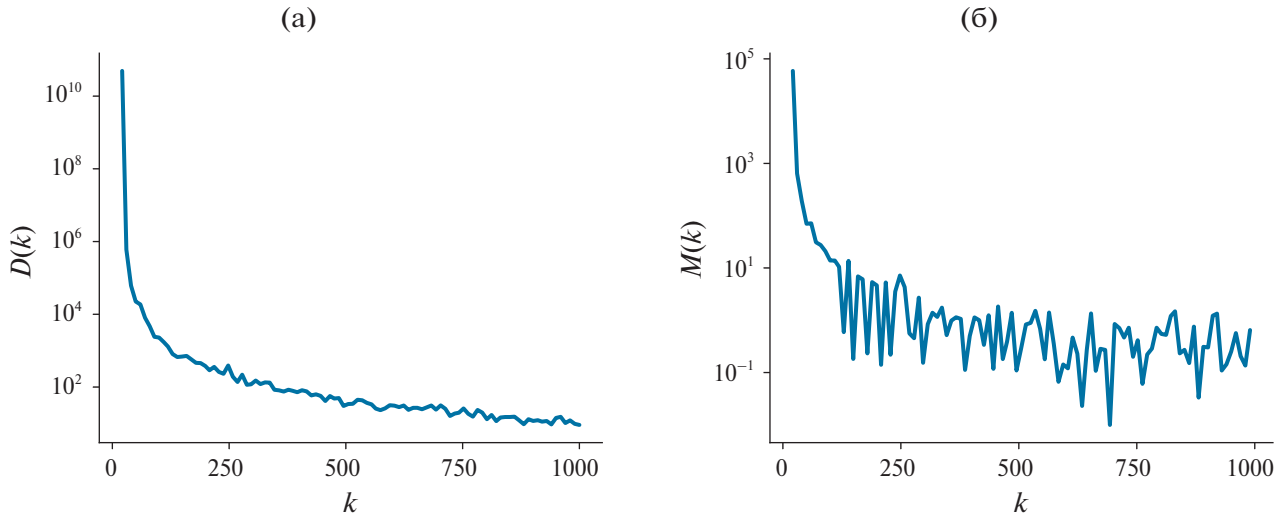
Синтетические данные сгенерированы из моделей линейной и логистической регрессий. Число объектов 1000, число признаков 20. Используется $B = 1000$ бустрапированных подвыборок. Подсчитываются значения функций D и M. Датасет с задачей регрессии Liver Disorders содержит 345 объектов и 5 признаков. Мы

также используем $B = 1000$ бутстрапированных подвыборок для оценки математического ожидания и дисперсии функции ошибки.

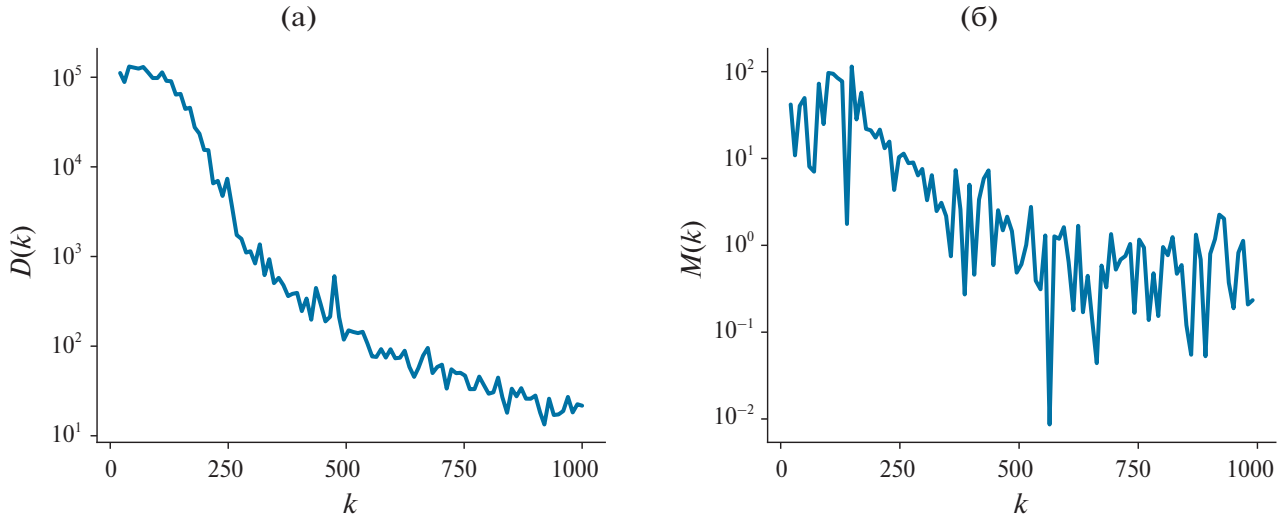
На фиг. 1 можно видеть полученные зависимости между используемым размером подвыборки k и рассматриваемыми функциями D и M для синтетической выборки с задачей регрессии. Результаты для синтетической выборки с задачей классификации представлены на фиг. 2. В то же время, на фиг. 3 мы видим аналогичные графики для датасета Liver Disorders. Видно, что во всех случаях значения функций D и M стремятся к нулю при увеличении размера выборки. Эти эмпирические результаты подтверждают теоретические, полученные ранее.

В определениях D-достаточности и M-достаточности участвует гиперпараметр ε, который отвечает за порог для достаточного размера выборки m^*. С целью изучения зависимости между ними, мы представляем фиг. 4, где указано, какой размер выборки следует выбрать, чтобы обеспечить определенный уровень уверенности.

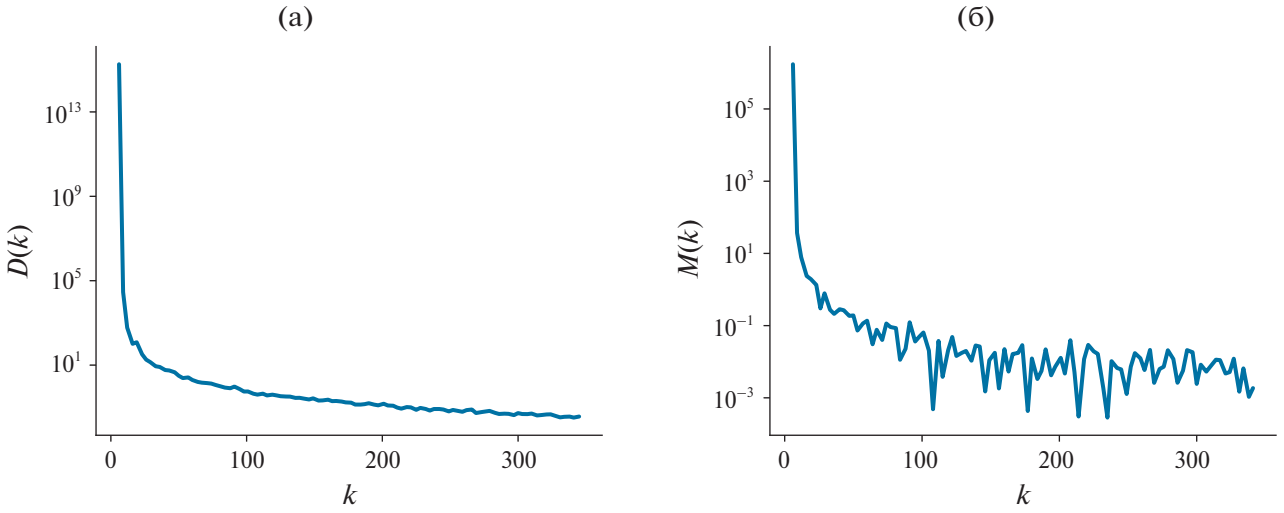
Чтобы сравнить эффективность предложенных методов на разных наборах данных, были выбраны выборки из открытой библиотеки [20]. Подробная информация о каждом наборе данных, количество наблюдений и количество признаков представлены в табл. 1. Для демонстрационных целей были выбраны такие значения гиперпараметра ε, при которых значения функций D и M уменьшаются в два раза. Соответствующие результаты приведены в табл. 1. Пропуски означают, что первоначальный размер выборки недостаточен.



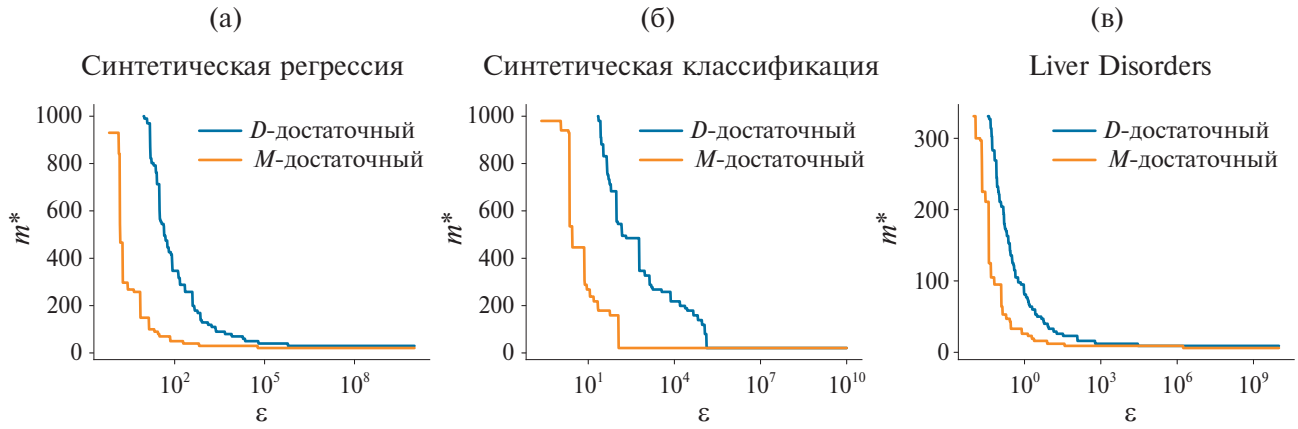
Фиг. 1. Сходимость предлагаемых функций D и M для выборки синтетической регрессии, т.е. модели линейной регрессии. Обе функции стремятся к нулю при увеличении размера выборки.



Фиг. 2. Сходимость предлагаемых функций D и M для выборки синтетической классификации, т.е. модели логистической регрессии. Обе функции стремятся к нулю при увеличении размера выборки.



Фиг. 3. Сходимость предлагаемых функций D и M для выборки Liver Disorders. Обе функции стремятся к нулю при увеличении размера выборки.



Фиг. 4. Зависимость достаточного размера выборки от значения порога на трех наборах данных: синтетическая регрессия, синтетическая классификация и Liver Disorders. При увеличении значения порога ε достаточный размер уменьшается. Это означает, что можно выбирать меньше объектов для удовлетворения желаемых значений предлагаемых функций D и M.

5. ОБСУЖДЕНИЕ

В статье предлагаются два новых метода определения достаточного размера выборки, основанные на значениях правдоподобия на бутстрапированных подвыборках. Первый метод, называемый D-достаточностью, основан на дисперсии функции правдоподобия, в то время как второй, M-достаточность, фокусируется на разности в математическом ожидании функции правдоподобия при добавлении одного объекта в выборку. Демонстрируется корректность определения M-достаточности в модели линейной регрессии при определенных условиях на параметры модели.

Вычислительные эксперименты, проведенные как на синтетических, так и на реальных выборках, показывают, что предлагаемые функции D и M стремятся к нулю при увеличении размера выборки. Эксперименты также подчеркивают практическую значимость методов, поскольку они могут быть легко применены к различным наборам данных.

Предлагаемые методы потенциально могут быть применены к широкому спектру моделей и наборов данных, помимо линейной регрессии. Хотя мы доказали корректность определения M-достаточного размера выборки только для линейной регрессии, эмпирические результаты показывают, что эти методы могут быть эффективны и для других моделей. Будущая работа должна быть сосредоточена на распространении теоретического анализа на другие модели, включая, вероятно, нейронные сети.

Таблица 1. Сравнение предлагаемых методов определения достаточного размера выборки: на основе D и M. Для каждой из предлагаемых функций подбирается такой значение порога, что ее изначальное значение уменьшается вдвое. Результаты представлены для набора выборок с задачей регрессии. Пропуски в данных означают, что исходный размер выборки недостаточен.

Название	Объекты m	Признаки n	D	M
Abalone	4177	8	96	96
Auto MPG	392	8	15	15
Automobile	159	25	70	156
Liver Disorders	345	6	12	19
Servo	167	4	41	—
Forest fires	517	12	208	—
Wine Quality	6497	12	144	144
Energy Efficiency	768	9	24	442
Student Performance	649	32	129	177
Facebook Metrics	495	18	31	388
Real Estate Valuation	414	7	15	23
Heart Failure Clinical Records	299	12	63	224
Bone marrow transplant: children	142	36	—	—

6. ЗАКЛЮЧЕНИЕ

В статье представлены два новых метода, D-достаточность и M-достаточность, для определения достаточного размера выборки на основе значений правдоподобия на бутстрапированных подвыборках. Корректность определения M-достаточного размера выборки продемонстрирована в модели линейной регрессии, а вычислительные эксперименты на синтетических и реальных наборах данных показывают, что предложенные функции сходятся к нулю по мере увеличения размера выборки, что подчеркивает практичность методов.

СПИСОК ЛИТЕРАТУРЫ

1. *Robert R Bies, Matthew F Muldoon, Bruce G Pollock et al.* A genetic algorithm-based, hybrid machine learning approach to model selection // J. Pharmacokinet. Pharmacodyn. 2006. V. 33. № 2. P. 195.

2. *Cawley, Gavin C.* On over-fitting in model selection and subsequent selection bias in performance evaluation // J. Mach. Learn. Res. 2010. V. 11. № 1. P. 2079–2107.

3. *Richard H Byrd, Gillian M Chin, Jorge Nocedal, Yuchen Wu.* Sample size selection in optimization methods for machine learning // Math. Program. 2012. V. 134. № 1. P. 127–155.

4. *Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, Long H Ngo.* Predicting sample size required for classification performance // BMC Med. Inf. Decis. Making. 2012. V. 12. № 1. P. 1–10.

5. *Indranil Balki, Afsaneh Amirabadi, Jacob Levman et al.* Sample-size determination methodologies for machine learning in medical imaging research: a systematic review // Can. Assoc. Radiol. J. 2019. V. 70. № 4. P. 344–353.

6. *Adcock, C. J.* A Bayesian Approach to Calculating Sample Sizes // J. R. Stat. Soc. D. 1988. V. 37. № 4. P. 433.

7. *Lawrence Joseph, David B. Wolfson, Roxane Du Berger.* Sample Size Calculations for Binomial Proportions via Highest Posterior Density Intervals // J. R. Stat. Soc. D. 1995. V. 44. № 2. P. 143–154.

8. *Steven G Self, Robert H Mauritsen.* Power/sample size calculations for generalized linear models // Biometrics. 1988. V. 44. № 1. P. 79–86.

9. *Gwowen Shieh*. On power and sample size calculations for likelihood ratio tests in generalized linear models // Biometrics. 2000. V. 56. № 4. P. 1192–1196.

10. *Gwowen Shieh*. On power and sample size calculations for Wald tests in generalized linear models // J. Stat. Plann. Inference . 2005. V. 128. № 1. P. 43–59.

11. *Dennis V. Lindley*. The choice of sample size // J. R. Stat. Soc. D. 1997. V. 46. № 2. P. 129–138.

12. *Dennis V. Lindley*. On Bayesian analysis, Bayesian decision theory and the sample size problem // J. R. Stat. Soc. D. 1997. V. 46. № 2. P. 139–144.

13. *Alan E. Gelfand, Fei Wang*. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models // Stat. Sci. 2002. V. 17. № 2. P. 192-208.

14. *Jing Cao, J. Jack Lee, Susan Alber*. Comparison of Bayesian sample size criteria: ACC, ALC, and WOC // J. Stat. Plann. Inference. 2009. V. 139. № 12. P. 4111–4122.

15. *Pierpaolo Brutti, Fulvio De Santis, Stefania Gubbiotti*. Bayesian-frequentist sample size determination: a game of two priors // METRON 2014. V. 72. № 2. P. 133–151.

16. *Hamid Pezeshk, Nader Nematollahi, Vahed Maroufy, John Gittins*. The choice of sample size: a mixed Bayesian / frequentist approach // Stat. Methods Med. Res. 2008. V. 18. № 2. P. 183–194.

17. *A. V. Grabovoy, T. T. Gadaev, A. P. Motrenko, V. V. Strijov*. Numerical Methods of Sufficient Sample Size Estimation for Generalised Linear Models // Lobachevskii J. Math. 2022. V. 43. № 9. P. 2453–2462.

18. *Anastasiya Motrenko, Vadim Strijov, Gerhard-Wilhelm Weber*. Sample size determination for logistic regression // J. Comput. Appl. Math. 2014. V. 255. № 2. P. 743–752.

19. *Lawrence Joseph, Roxane Du Berger, Patrick Belisle*. Bayesian and mixed Bayesian/likelihood criteria for sample size determination // Stat. Med. 1997. V. 16. № 7. P. 769–781.

20. *Markelle, Kelly*. The UCI Machine Learning Repository. https://archive.ics.uci.edu.

ПРИЛОЖЕНИЕ

Доказательство теоремы 1

Доказательство. Рассмотрим определение M-достаточного размера выборки в терминах логарифма функции правдоподобия. В модели линейной регрессии

$$L(\mathcal{D}_m, \hat{\mathbf{w}}_k) = p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}_k) = \prod_{i=1}^{m} p(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_k) = \prod_{i=1}^{m} \mathcal{N}\left(y_i|\hat{\mathbf{w}}_k^\top \mathbf{x}_i, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-m/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_k\|_2^2\right).$$

Прологарифмируем:

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

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

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

Запишем выражение для разности математических ожиданий:

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

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

$$= \frac{1}{2\sigma^2}\left(2\mathbf{y}^\top\mathbf{X}(\mathbf{m}_{k+1} - \mathbf{m}_k) + (\mathbf{m}_k - \mathbf{m}_{k+1})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\right) +$$

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

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:

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

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:

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

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

Последнее слагаемое оценим, используя неравенство Гёльдера для нормы Фробениуса:

$$\left|\mathrm{tr}\left(\mathbf{X}^\top \mathbf{X}\,(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1})\right)\right| \leqslant \|\mathbf{X}^\top \mathbf{X}\|_F \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\|_F.$$

Наконец, поскольку $\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \to 0$ и $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\|_F \to 0$ при $k \to \infty$, то $M(k) \to 0$ при $k \to \infty$, что доказывает теорему.

Доказательство следствия 1

Доказательство. 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 $k \to \infty$. Применение теоремы 1 заканчивает доказательство.

SUFFICIENT SAMPLE SIZE: LIKELIHOOD BOOTTRAPPING

N. S. Kiselev[a,*], A. V. Grabovoi[a,]**

[a] *MIPT, Institutskii per. 9, Moscow region, 141701 Dolgoprudny, Russia*
**e-mail: kiselev.ns@phystech.edu*
***e-mail: grabovoy.av@phystech.edu*

Received October 6, 2024
Revised October 6, 2024
Accepted November 10, 2024

Abstract. Determining the appropriate sample size is crucial for building effective machine learning models. Existing methods often either lack a rigorous theoretical basis or are tied to specific statistical hypotheses about the model parameters. In this paper, we present two new methods based on likelihood values on bootstrapped subsamples. We demonstrate the correctness of one of these methods in a linear regression model. Computational experiments with both synthetic and real datasets show that the proposed functions converge as the sample size increases, highlighting the practical usefulness of the approach.

Keywords: sufficient sample size, likelihood bootstrapping, linear regression, computational linear algebra