

МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ  
(национальный исследовательский университет)  
ФИЗТЕХ-ШКОЛА ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ  
КАФЕДРА ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ

Киселев Никита Сергеевич

## БАЙЕСОВСКИЙ ПОДХОД К ВЫБОРУ ДОСТАТОЧНОГО РАЗМЕРА ВЫБОРКИ

03.03.01 — Прикладные математика и физика

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

**Научный руководитель:**  
к.ф.-м.н. А. В. Грабовой

Москва — 2023

## 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.*

## Оглавление

	Стр.
Введение . . . . .	4
Глава 1. Постановка задачи	5
Глава 2. Достаточный размер выборки не превосходит доступный	6
Глава 3. Достаточный размер выборки больше доступного	8
Глава 4. Вычислительный эксперимент	9
4.1. Достаточный размер выборки не превосходит доступный . . . . .	9
4.2. Достаточный размер выборки больше доступного . . . . .	10
Заключение . . . . .	12
Список иллюстраций . . . . .	13
Список литературы . . . . .	14

## 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.

# Глава 1

## 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  $\mathfrak{D}_m$ :

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

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  $m^* \leq m$  или  $m^* > m$ . Эти два случая будут отдельно рассмотрены далее.

## Глава 2

### Достаточный размер выборки не превосходит доступный

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

Рассмотрим выборку  $\mathfrak{D}_k$  размера  $k \leq m$ . 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 правдоподобия. 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  $k \geq m^*$

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

**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  $k \geq m^*$

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

**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?

## Глава 3

### 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?



## Глава 4

### 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.

#### 4.1. 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 2 (здесь используется логарифм функции правдоподобия). Выполнено определение D-достаточного и M-достаточного размеров выборки. Использовалось  $B = 1000$  бутстрэп-выборок. Результаты представлены на Рис. 4.1.

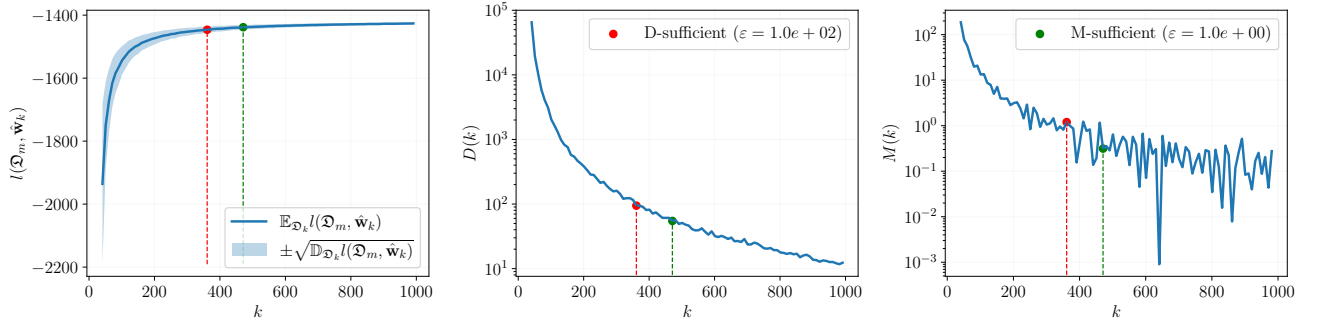


Рис. 4.1: Синтетическая выборка (линейная регрессия) при  $m^* \leq m$

Вторая синтетическая выборка сгенерирована из модели логистической регрессии. Число объектов 1000, число признаков 20. Аналогичные графики приведены на Рис. 4.2.

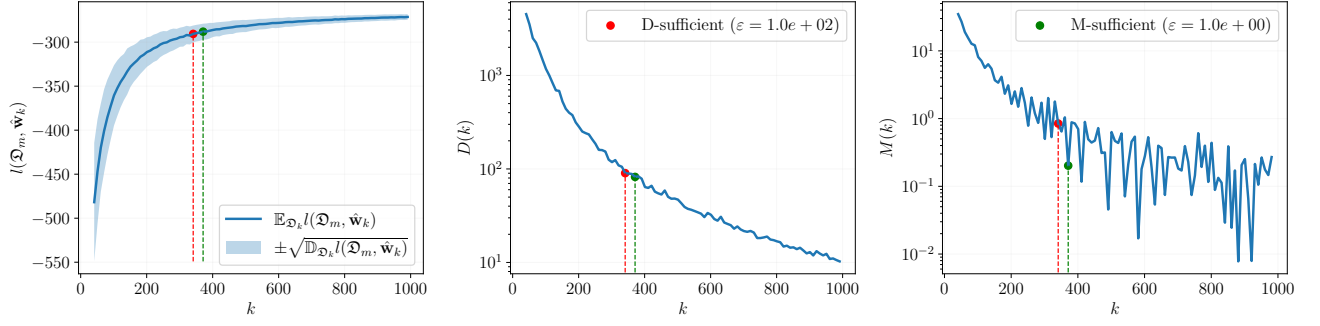


Рис. 4.2: Синтетическая выборка (логистическая регрессия) при  $m^* \leq m$

## 4.2. Достаточный размер выборки больше доступного

Для синтетических выборок проведена аппроксимация функций правдоподобия. Среднее значение и дисперсия аппроксимированы соответственно функциями

$$\varphi(m) = a_1 - a_2^2 \exp(-a_3^2 m) - \frac{a_4^2}{m^{3/2}}$$

и

$$\psi(m) = b_1^2 \exp(-b_2^2 m) + \frac{b_3^2}{m^{3/2}},$$

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. 4.3 и Рис. 4.4 представлены истинные и восстановленные зависимости. Там же указаны определенные D-достаточный и M-достаточный размеры выборки.

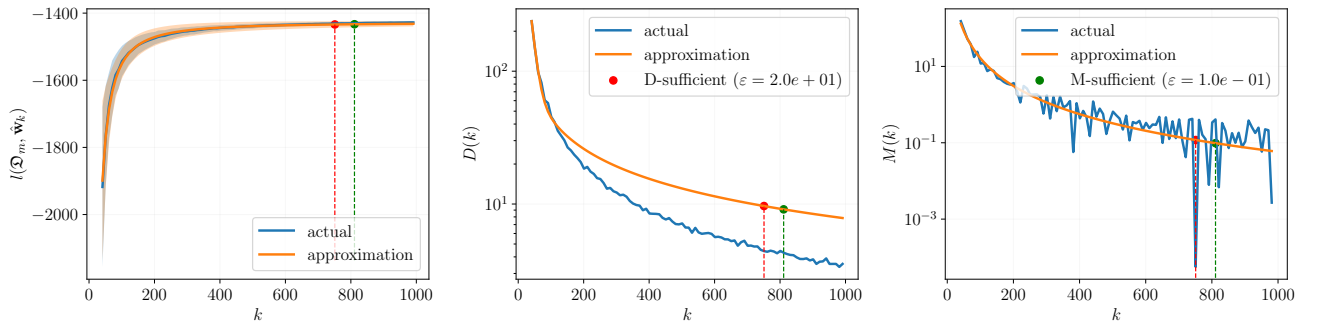


Рис. 4.3: Синтетическая выборка (линейная регрессия) при  $m^* > m$

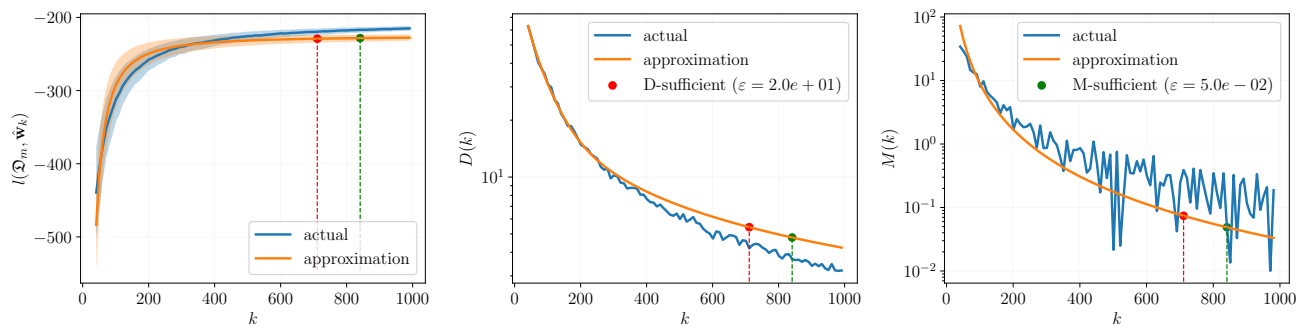


Рис. 4.4: Синтетическая выборка (логистическая регрессия) при  $m^* > m$

## **Заключение**

Основные результаты данной работы заключаются в следующем.

Бла-бла-бла.

## Список иллюстраций

4.1	Синтетическая выборка (линейная регрессия) при $m^* \leq m$ . . . .	9
4.2	Синтетическая выборка (логистическая регрессия) при $m^* \leq m$ .	10
4.3	Синтетическая выборка (линейная регрессия) при $m^* > m$ . . . .	10
4.4	Синтетическая выборка (логистическая регрессия) при $m^* > m$ .	11

## 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.