# Оценка оптимального объёма выборки для задач классификации*

***Харатян А. С., Катруца А. М.***[1]***, Стрижов В. В.***[2]
haratyan.as@phystech.edu; aleksandr.katrutsa@phystech.edu;
strijov@phystech.edu
[1]Московский физико-технический институт, Москва, Россия;[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН, Москва, Россия

В статье рассматривается задача выбора оптимального числа объектов выборки для их классификации. Исследуется использование порождающих и разделяющих вероятностных моделей бинарной классификации. Обсуждается проблема медицинской диагностики пациентов. Определяется понятие достаточности объёма выборки. Показывается, какими методами возможно выбрать оптимальное количество объектов, обеспечивающее необходимую точность классификации объектов . В работе рассматривается, применение каких критериев выявляет наилучшее качество классификации. Приводится теоретическое и практическое обоснование предложенных критериев. Используется модель логистической регрессии.

**Ключевые слова**: *определение оптимального объёма выборки, логистическая регрессия, расстояние Кульбака-Лейблера.*

## 1 Введение

Работа посвящена оценке оптимального объёма исследуемой выборки применительно к проблемам медицинской диагностики. Рассматриваются биомедицинские данные пациентов как выборка. Каждый пациент набором признаков. Получение данных о пациентах требует немалых средст. В случае, если количество данных избыточно, то их измерения приносят крайне неоправданные расходы. В связи с этим поднимается вопрос оптимального количества измерений. Ввиду дороговизны анализов всех признаков оценка измерений должна быть точной.

Для нахождения оценки используется модель логистической регрессии [1]. Стандартной практикой является использование статистических методов [2] для оценивания объёма данных при помощи логистической регрессии. Введём понятие устойчивости модели в отношении объёма выборки. Будем называть модель устойчивой, если при изменении малом изменении объёма параметры модели меняются незначительно. Если размер выборки крайне мал и недостаточен, то параметры модели меняются скачкообразно при увеличении объёма.Соответственно, увеличивая объём выборки, мы повышаем устойчивость модели. В качестве показателя устойчивости для моделей будем использовать расстояние Кульбака-Лейблера. Для того чтобы показать отличие моделей в устойчивости будем использовать разность усредненных значений расстояний К-Л, вычисленных на разных выборках одного и того же объёма. Если объекты порождены одинаковым распределением, то при росте объёма выборок разность расстояний К-Л между моделями падает. Достигнув необходимого показателя устойчивости, можно легко вычислить оптимальный размер данных.

*Работа выполнена при финансовой поддержке РФФИ, проект № 00-00-00000. Научный руководитель: Стрижов В. В. Консультант: Катруца А. М.

Для вычислительного экперимента используются реальные данные 569 пациентов с 30 признаками и метками об опухоли молочной железы: доброкачественная или злокачественная. Как описано выше, будем обучать модели на разных подвыборках и после достижения необходимого уровня устойчивости получим оптимальный объём данных.

## 2 Постановка задачи классификации

Пусть у нас задана выборка $D = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, m\}$ с объёмом $m$ объектов(пациентов), каждый из которых описывается $n$ признаками, $\mathbf{x}_i \in \mathbb{R}^n$ и принадлежит одному из двух классов: $y_i \in \{0, 1\}$ . Модель логистической регрессии предполагает, что вектор целевой переменной $\mathbf{y} = [y_1, \ldots, y_m]^T$ имеет распределение Бернулли, $y_i \sim \mathscr{B}(\theta_i)$ с плотностью распределения

$$p(y|\boldsymbol{\omega}) = \prod_{i=1}^{m} \theta_i^{y_i} (1 - \theta_i)^{1-y_i} \tag{1}$$

Плотность вероятности зависит от вектора параметров $\omega$. Зная $\omega$, можно вычислить вероятность принадлежности к классу

$$\theta_i = f\left(\mathbf{x}_i^T \omega\right) = \frac{1}{1 + \exp\left(-\mathbf{x}_i^T \omega\right)} \tag{2}$$

Пользуясь принципом максимального правдоподобия, можем вычислить функцию ошибки для уравнения (1)

$$E(\boldsymbol{\omega}) = -\ln p(\mathbf{y}|\boldsymbol{\omega}) = -\sum_{i=1}^{m} \left(y_i \ln \theta_i + (1 - y_i) \ln (1 - \theta_i)\right) \tag{3}$$

Чтобы найти вектор параметров логистической регрессии $\widehat{\omega}$, необходимо решить оптимизационную задачу:

$$\hat{\omega} = \arg\min_{\omega \in \mathbb{R}^n} E(\omega) \tag{4}$$

Тогда алгоритм классификации определяется следующим образом:

$$a(\mathbf{x}, c_0) = \operatorname{sign}\left(f(\mathbf{x}, \boldsymbol{\omega}) - c_0\right) \tag{5}$$

где $c_0$ - пороговое значение функции активации [3] .

## 3 Постановка задачи определения оптимального объёма выборки

Допустим, что задано $m$ объектов выборки $D$. Количество этих объектов недостаточно для обучения устойчивой модели логистической регрессии. Необходимо, имея $m$ объектов, найти такое число $m^*$, что показатель устойчивости модели при увеличении числа $m^*$ меняется незначительно. Задача нахождения данного числа объектов представляется следующим образом:

$$m^* = \min m \in \mathbb{N} \;\; \forall k \in \mathbb{N} \to \rho(\mathrm{p}(\boldsymbol{w}|m^*), \mathrm{p}(\boldsymbol{w}|m^* + k)) < \varepsilon \tag{6}$$

где $\rho$ - некоторая функция расстояния между распределениями, $\varepsilon$ - заранее заданный порог устойчивости, $\mathrm{p}(\boldsymbol{w}|m^*)$ - распределение весов модели, обученной на выборке размером $m^*$.

## Литература

[1] Hosmer Jr, D. W., S. Lemeshow, and R. X. Sturdivant. 2013. *Applied logistic regression.* John Wiley & Sons. Vol. 398.

[2] Demidenko, E. 2007. Sample size determination for logistic regression revisited. *Statistics in medicine* 26(18):3385–3397.

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

*Received*