Journal of Computational and Applied Mathematics 255 (2014) 743–752

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Sample size determination for logistic regression[☆]

Anastasiya Motrenko[a,*], Vadim Strijov[b], Gerhard-Wilhelm Weber[c]

[a] *Moscow Institute of Physics and Technology, Moscow, Russia*
[b] *Computing Center of the Russian Academy of Sciences, Moscow, Russia*
[c] *Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey*

ARTICLE INFO

Article history:
Received 6 December 2012
Received in revised form 7 June 2013

Keywords:
Logistic regression
Sample size
Feature selection
Bayesian inference
Kullback–Leibler divergence

ABSTRACT

The problem of sample size estimation is important in medical applications, especially in cases of expensive measurements of immune biomarkers. This paper describes the problem of logistic regression analysis with the sample size determination algorithms, namely the methods of univariate statistics, logistics regression, cross-validation and Bayesian inference. The authors, treating the regression model parameters as a multivariate variable, propose to estimate the sample size using the distance between parameter distribution functions on cross-validated data sets. Herewith, the authors give a new contribution to data mining and statistical learning, supported by applied mathematics.

1. Introduction

This paper is devoted to logistic regression analysis [1], applied to classification problems in biomedicine. A group of patients is investigated as a sample set; each patient is described with a set of features, named biomarkers, and is classified into two classes.

Since the patient measurement is expensive, the number of patients in the sample studied in this paper is rather small: two classes contain 14 and 17 patients, respectively. In this case, the classification model overfitting is unavoidable. This leads to the problem of the sample size determination. Due to the high cost of examination of each new patient, the estimation of the sample size should be precise. The common practice [2,3] for the logistic regression is to use statistical methods to estimate the sample size. The sample size is estimated with respect to each feature, one by one. However, these methods appear to provide an overestimated sample size. The problem of sample size estimation calls for a new solution. Let us define the instability of the model in relation to the sample size. We will call the model *unstable*, if the model parameters change significantly when the sample is slightly varied. Fig. 2 shows how the position of the hyperplane has changed after two objects were added to the sample. When the sample size is insufficient, the model parameters estimations are unstable. Increasing the sample size we expect to increase stability of the parameters. To measure stability, we propose to compute the averaged Kullback–Leibler divergence between the probability density functions of the model parameters. The parameters are estimated at different subsets of the same size. The divergence should decrease with the increment of the subsets' size, if these subsets belong to the same statistical population. When a threshold value of stability is assigned, one can compute the sample size required to achieve this level of stability.

The paper is organized in the following way. A brief description of the logistic regression and the quality function used in this paper is presented in Section 2. The target variable is assumed to follow a Bernoulli distribution. The parameters of the regression model are estimated [4,5]. The studied sample consists of 31 patients with cardio-vascular system disorder. The

[☆] This project was supported by the Russian Foundation for Basic Research, grant 12-07-31095, 13-07-00709.

[*] Corresponding author.
E-mail addresses: anastasia.motrenko@gmail.com (A. Motrenko), strijov@ccas.ru (V. Strijov).

0377-0427/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cam.2013.06.031

experts name 20 features that describe the sample. With a given set of features, the model is excessively complex. That is why, before estimating the sample size, we select a set of features of a smaller size that will classify patients effectively. In logistic regression, features are selected using stepwise regression procedure [6,7]. In our computational experiment an exhaustive search is implemented. This makes the experts sure that every possible combination of the features is considered. We use the area under the ROC curve [8–10] as the optimum criterion in the feature selection procedure. The feature selection problem is discussed in Section 3. In Section 4 the following methods of minimum sample size determination are discussed:

1. Method of *confidence intervals*: a method of univariate statistics [2]. This method does not consider the model or our assumptions about a probability distribution of the variables. This method is designed for case of a single feature.
2. Method of *sample size evaluation* in *logistic regression* [3]. Unlike the previous one, this method considers the distribution of the responsive variable according to the logistic regression model. Again, this approach is meant to be used when the class variable is described by a single feature.
3. *Cross-validation*: a method which evaluates sample size by observing potential overfitting [11,12]. This method is not associated with any certain model, but can be implemented in case of multiple features.
4. *Comparing different subsets* of the same sample using the *Kullback–Leibler divergence* [13] between probability density functions of the model parameters, evaluated at similar subsets. This approach allows us to estimated the sample size for the multi-feature sample set and takes into account the probabilistic assumptions and the model. This is a new method proposed by the authors.

These methods are tested on real and synthetical data. The results of the experiment are discussed in Section 5.

2. Classification problem statement

Consider the sample set $D = \{(\mathbf{x}_i, y_i) : i = 1, \dots, m\}$ of m objects (patients). Each patient is described by n features (biomarkers), $\mathbf{x}_i \in \mathbb{R}^n$, and belongs to one of two classes: $y_i \in \{0, 1\}$. The logistic regression problem assumes that the vector of responsive variables $\mathbf{y} = [y_1, \dots, y_m]^T$ is a vector of Bernoulli random variables, $y_i \sim \mathcal{B}(\theta_i)$, with the probability density function

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

The probability density function depends on the parameter vector $\boldsymbol{\beta}$. Given $\boldsymbol{\beta}$, the probability θ_i is defined as

$$\theta_i = f(\mathbf{x}_i^T\boldsymbol{\beta}) = \frac{1}{1+\exp(-\mathbf{x}_i^T\boldsymbol{\beta})}. \tag{2}$$

We use the *maximum likelihood* method, write the error function for Eq. (1) as

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

To find the vector of parameters $\hat{\boldsymbol{\beta}}$ of regression function, one has to solve the following optimization problem:

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

Then, the classification algorithm is defined as:

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

where c_0 is a cut-off value of regression function (2), defined by (6).

Classification quality function. Let us use an additional to (1) namely the quality function AUC, or the *area under the ROC-curve*. We introduce TPR(ξ), which stands for true positive rate

$$\text{TPR}(\xi) = \frac{1}{m}\sum_{i=1}^{m}[a(\mathbf{x}_i, \xi) = 1][y_i = 1],$$

and FPR(ξ) means the false positive rate

$$\text{FPR}(\xi) = \frac{1}{m}\sum_{i=1}^{m}[a(\mathbf{x}_i, \xi) = 1][y_i = 0].$$