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:

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

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

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

and FPR(ξ) means the false positive rate

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

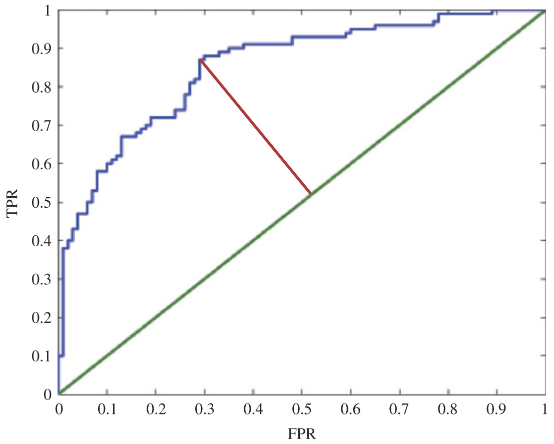


Fig. 1. Sample size m^*, estimated by confidence interval method and method for logistic regression.

The parameter ξ covers the values from zero to one and is considered as the cut-off value to construct the ROC-curve. Here, the following denotation for the indicator function is used:

$$[y=1]=\begin{cases}1, & \text{if } y=1,\\ 0, & \text{if } y\neq 1.\end{cases}$$

Thus, the bigger the AUC value the better the classifier.

Defining c_0 value. Every point $(\mathrm{FPR}(c_0), \mathrm{TPR}(c_0))$ of the ROC-curve corresponds to some $c_0 \in [0, 1]$ value. As shown in Fig. 1, the most distant from segment $[(0, 0); (1, 1)]$ point of the ROC-curve corresponds to the value c_0 used in Eq. (5):

$$\hat{c}_0=\arg\max_{\xi\in[0,1]}\|(\mathrm{TPR}(\xi),\mathrm{FPR}(\xi))-(\xi,\xi)\|=\arg\max_{\xi\in[0,1]}\sqrt{\left(\mathrm{TPR}(\xi)-\xi^2-\mathrm{FPR}(\xi)-\xi\right)^2}. \tag{6}$$

Defining $\hat{c}_0$ includes computing the AUC value and, therefore, computation of (5) and iterative estimation of the parameters in $\boldsymbol{\beta}$ according to Eq. (4).

3. Feature selection problem

Let $\mathcal{A}$ be a subset of the indexes of the features, $\mathcal{A} \subseteq \mathcal{J} = \{1, \dots, n\}$, and $\hat{\mathcal{A}}$ be the optimal subset of the indexes. Denote by $\mathbf{X}_{\mathcal{A}}$ the matrix subsequently composed of the columns of the matrix $\mathbf{X}$ with indexes in $\mathcal{A}$, and $\boldsymbol{\beta}_{\mathcal{A}}$ be the corresponding vector of parameters. Thus, the feature selection problem is a maximization one:

$$\hat{\mathcal{A}}=\arg\max_{\mathcal{A}\subseteq\mathcal{I}}\mathrm{AUC}(\mathcal{A}),\qquad\text{subject to } |\mathcal{A}|=\text{const}. \tag{7}$$

The value of AUC $(\mathcal{A}) \equiv \mathrm{AUC}(\mathbf{X}_{\mathcal{A}}, \hat{\boldsymbol{\beta}}_{\mathcal{A}}, \hat{c}_0, \mathbf{y})$ is computed for a set $\mathcal{A}$ of indexes and the parameters $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ and c_0 are defined by Eqs. (4) and (6), respectively.

The maximization problem (7) is solved in the computational experiment by exhaustive search. This approach is possible due to a relatively small amount of features and it is required by experts.

As the cardinality of $\mathcal{A}$ is unknown, the set of indexes of objects $\mathcal{I}$ is divided into two disjoint subsets, $\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$, the *learning set* and the *test set*: the parameters $\boldsymbol{\beta}$ are estimated at $D_{\mathcal{L}} = \{(\mathbf{x}_i, y_i) : i \in \mathcal{L}\}$, while the classification quality is computed at $D_{\mathcal{T}} = \{(\mathbf{x}_i, y_i) : i \in \mathcal{T}\}$. The maximum cardinality of $\mathcal{A}$ is limited by experts: $|\mathcal{A}|$ shall not exceed the number four. We refer to the feature sets, obtained by solving problem (7), as *optimal sets*, and name the features included into optimal sets as the *most informative features*.

4. Sample size determination

The investigated data describe patients of two classes: (i) those who have already experienced a heart attack, and (ii) patients that might experience it in the future. Concentrations of proteins in blood cells are used as 20 features. There are 17 patients in the first class and 14 in the second. Having these few observations, we must estimate the minimum sample size m^* required to obtain adequate results of classification. In this section, four methods of sample size determination are presented. The last one is a new method, introduced by the authors. The results of implementing this methods are described and analyzed in Section 5.

4.1. Method of confidence intervals

Consider the data set $D = \{(x_i, y_i) : i \in \mathcal{I} = \{1, \dots, m\}\}$ in which every responsive variable y_i depends on a single independent variable $x_i \sim \mathcal{N}(\mu, \sigma^2)$. Suppose $\Delta = \bar{x} - \mu$ is the difference between the average

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

and the known expected value μ of the random variable x_i. Given the variance σ^2, we obtain a standard normally distributed variable

$$Z = \frac{\bar{x} - \mu}{\sigma}\sqrt{m} = \frac{\Delta}{\sigma}\sqrt{m} \sim \mathcal{N}(0, 1). \tag{8}$$

Then m^* can be computed with significance level α as

$$m^* = \left(\frac{z_{\alpha/2}\sigma}{\Delta}\right)^2, \tag{9}$$

where $z_{\alpha/2}$ is defined by $P\left\{|Z| \geq z_{\alpha/2}\right\} = \alpha$.

In this paper, a multi-feature problem is considered and every responsive variable y_i is described by the vector of independent variables $\mathbf{x}_i$. Nevertheless, formula (9) can be used for each feature separately as the components of $\mathbf{x}_i$ are assumed to be independent.

This method only helps to obtain a rough estimation of m^*. The reason is that neither μ nor σ^2 are known. Also, it is more likely that x_i is distributed as a mixture of distributions:

$$x_i \sim \begin{cases} \mathcal{N}(\mu_1, \sigma_1^2), & \text{with probability } \theta_i, \\ \mathcal{N}(\mu_2, \sigma_2^2), & \text{with probability } 1 - \theta_i, \end{cases} \tag{10}$$

where θ_i is defined by Eq. (2).

4.2. Method of sample size evaluation in logistic regression

Let us fixate some index set $\mathcal{A}$. For every feature in the set, defined by $\mathcal{A}$, we can compute the sample size m^*, required to include this feature into the model feature set. We consider the hypothesis

$$H_0 : \beta_j = 0, \quad j \notin \mathcal{A},$$

with β_j being the jth element of the vector $\boldsymbol{\beta}$ of logistic regression parameters. In this way, we assume that the jth feature is not included in the model. Having estimated the vector of parameters under H_0, we obtain the vector $\boldsymbol{\beta}_{\mathcal{A}}$, and under alternative $H_1 : \beta_j \neq 0$ we get $\boldsymbol{\beta}_{\mathcal{A}^*}$, where the index set $\mathcal{A}^*$ is composed of $\mathcal{A}$ and index j. Then H_0 and H_1 can be reformulated in terms of parameters θ_i of Bernoulli distribution $\mathcal{B}(\theta)$ and rewritten as

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_{\mathcal{A}}, \qquad H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_{\mathcal{A}^*}.$$

We note that the exact values of θ_i in each case are not important, we are only interested in the cut-off value c_0. Finally, we have:

$$H_0 : 1 - c_0 = p_0, \qquad H_1 : 1 - c_0 = p_1.$$

To test the hypothesis H_0, we calculate statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0 c_0 / m}}, \quad \hat{p} = \frac{1}{m}\sum_{i=1}^{m} y_i,$$

where $\hat{p}$ is the maximum likelihood estimator for θ. Under H_0,

$$Z \sim \mathcal{N}\left(p_1 - p_0, \sqrt{\frac{p_1 c_1}{p_0 c_0}}\right).$$

Then

$$Z\sqrt{\frac{p_0 c_0}{p1 c_1}} + \frac{p_0 - p_1}{\sqrt{p_1 c_1 / m}} = \sqrt{\frac{p_0 c_0}{p_1 c_1}}\left(Z + \frac{p_0 - p_1}{\sqrt{p_0 c_0}}\sqrt{m}\right) \sim \mathcal{N}(0, 1).$$

With significance level α the power of the criterion can be computed:

$$1 - \beta = P\{|Z| > Z_{\alpha/2} | H_1\} = \Phi\left(\sqrt{\frac{p_0 c_0}{p_1 c_1}}\left(Z_{\alpha/2} + \frac{p_0 - p_1}{\sqrt{p_0 c_0 / m}}\right)\right).$$

Thus we obtain the following formula for m^*:

$$m^* = \frac{p_0 c_0 \left(Z_{1-\alpha/2} + Z_{1-\beta}\sqrt{\frac{p_1 c_1}{p_0 c_0}} \right)^2}{(p_1 - p_0)^2}. \tag{11}$$

We note that m^*, given by Eq. (11), depends on index j of a feature appearing in H_0.

4.3. Cross-validation

This method provides a minimum sample size estimation, based on observing overfitting. When using this approach, the data sample is divided into learning $D_{\mathcal{L}} = \{(\mathbf{x}_i, y_i) : i \in \mathcal{L}\}$ and test set $D_{\mathcal{T}} = \{(\mathbf{x}_i, y_i) : i \in \mathcal{T}\}$, where $\mathcal{I} = \mathcal{L} \bigsqcup \mathcal{T}$. We fixate a set $\mathcal{A}$ of indexes of model features, and by AUC$(\mathcal{A}, \mathcal{D})$ we denote the quality function value computed based on the data set $\mathcal{D}$. A decrease of the quality function AUC$(\mathcal{A}, D_{\mathcal{T}})$ value computed on the basis of the training set and compared to AUC$(\mathcal{A}, D_{\mathcal{L}})$ might indicate overfitting. We define *overfitting* as the following ratio:

$$\mathrm{RS}(m) = \frac{\mathrm{AUC}(\mathcal{A}, D_{\mathcal{T}(m)})}{\mathrm{AUC}(\mathcal{A}, D_{\mathcal{L}(m)})}. \tag{12}$$

In this case, the model f approximates the learning set, but it cannot be used to describe the test set. Overfitting might occur when the sample size m is too small. To reasonably assess m^*, we consequentially increase the sample size m while splitting the data set into learning and test sets under a given ratio:

$$|\mathcal{T}(m)|/|\mathcal{L}(m)| = \text{const} \leq 0.5.$$

With the increase of m, the value of RS(m) approaches one. We find the sample size m^* adequate, if for every $m \geq m^*$ the RS(m) ratio is more than a given value $1 - \varepsilon_1$.

4.4. Using Kullback–Leibler divergence to estimate sample size

The presented approach is based on comparing probability density functions of the model parameters. Consider two "similar" sets of indexes of objects $\mathcal{B}_1 \in \mathcal{J}$ and $\mathcal{B}_2 \in \mathcal{J}$. The index sets $\mathcal{B}_1$ and $\mathcal{B}_2$ are "similar" if

$$|(\mathcal{B}_1 \setminus \mathcal{B}_2) \cup (\mathcal{B}_2 \setminus \mathcal{B}_1)| = 1.$$

In this way, $\mathcal{B}_2$ can be obtained from $\mathcal{B}_1$ by deleting, replacing or adding one element. Parameters, evaluated at different samples $\mathcal{B}_1 \neq \mathcal{B}_2$, do also differ. Fig. 2 shows how the separating hyperplane given by

$$\mathbf{x}^T \boldsymbol{\beta} = \ln\left(\frac{c_0}{1 - c_0}\right)$$

changes when two elements are added to the sample. If the sample $D_{\mathcal{B}_1}$ is large enough, the parameter vector $\boldsymbol{\beta}_1$ evaluated based on $D_{\mathcal{B}_1}$ should not be significantly different from $\boldsymbol{\beta}_2$ obtained with a "similar" sample $D_{\mathcal{B}_2}$. The simplest way to compare them is to compute the Euclidean distance between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$:

$$\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| = \sqrt{\sum_{i=1}^{|\mathcal{A}|} \left(\beta_i^1 - \beta_i^2\right)^2}.$$

In this paper, probability density functions of parameters at $D_{\mathcal{B}_1}$ and $D_{\mathcal{B}_2}$ are compared by computing the *Kullback–Leibler divergence* values between them. Consider model function (2) and the assumption about the random variable y_i distribution (1). Having fixated the data set D and the model $f_{\mathcal{A}} = f(X_{\mathcal{A}}^T \boldsymbol{\beta})$, we rewrite Eq. (1) as

$$p(\mathbf{y}|X, \boldsymbol{\beta}, f_{\mathcal{A}}) \equiv p(D|\boldsymbol{\beta}, f_{\mathcal{A}}) = \prod_{i=1}^{m} \theta_i^{y_i}(1 - \theta_i)^{1-y_i}. \tag{13}$$

We suppose as well, that the vector of regression parameters $\boldsymbol{\beta}$ follows a normal distribution $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma^2 \mathbf{I}_{|\mathcal{A}|})$ with the density function

$$p(\boldsymbol{\beta}|f_{\mathcal{A}}, \alpha) = \left(\frac{\alpha}{2\pi}\right)^{\frac{|\mathcal{A}|}{2}} \exp\left(-\frac{\alpha}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2\right), \tag{14}$$

in which $\alpha^{-1} = \sigma^2$, $\mathbf{I}_{|\mathcal{A}|}$ being the unit matrix of format $|\mathcal{A}| \times |\mathcal{A}|$.

To find the probability density function $p(\boldsymbol{\beta}|D, \alpha, f_{\mathcal{A}})$ of the regression parameters, we use *Bayes' Theorem*:

$$p(\boldsymbol{\beta}|D, \alpha, f_{\mathcal{A}}) = \frac{p(D|\boldsymbol{\beta}, f_{\mathcal{A}})p(\boldsymbol{\beta}|\alpha, f_{\mathcal{A}})}{p(D|\alpha, f_{\mathcal{A}})}, \tag{15}$$

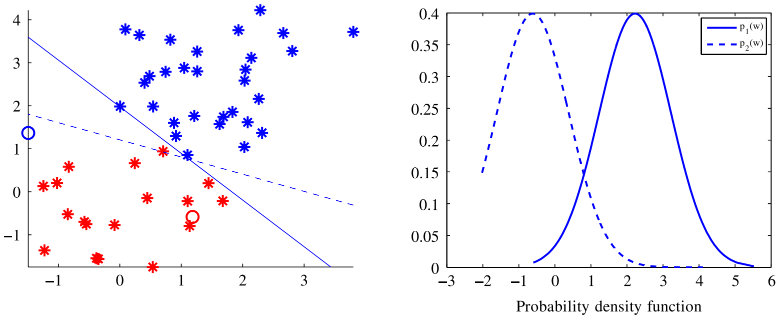


Fig. 2. Two classes are separated by a hyperplane. A dotted line represents the hyperplane position after the two random objects (in circles) were added.

where $p(D|\boldsymbol{\beta}, f_{\mathcal{A}})$ is the data likelihood, $p(\boldsymbol{\beta}|\alpha, f_{\mathcal{A}})$ given a priori probability density function. In (15), the normalization factor $p(D|\alpha, f_{\mathcal{A}})$ is defined by

$$p(D|\alpha, f_{\mathcal{A}}) = \int p(D|\boldsymbol{\beta}, f_{\mathcal{A}})p(\boldsymbol{\beta}|\alpha, f_{\mathcal{A}})d\boldsymbol{\beta}.$$

Substituting Eqs. (13) and (14) into Eq. (15) and denoting $Z(\alpha) = p(D|\alpha, f_{\mathcal{A}})$, we obtain

$$\begin{aligned} p(\boldsymbol{\beta}|D, f_{\mathcal{A}}) &= \frac{p(y|\mathbf{x}, \boldsymbol{\beta}, f_{\mathcal{A}})p(\boldsymbol{\beta}|f_{\mathcal{A}}, \alpha)}{Z(\alpha)} \\ &= \frac{\alpha^{\frac{|\mathcal{A}|}{2}}}{(2\pi)^{\frac{|\mathcal{A}|}{2}} Z(\alpha)} \exp\left(-\frac{\alpha}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2\right) \prod_{i=1}^{m} \theta_i^{y_i}(1-\theta_i)^{1-y_i}, \end{aligned} \tag{16}$$

where $Z(\alpha) = p(D|\alpha, f_{\mathcal{A}})$ is a normalization factor.

Consider two “similar” samples $D_{\mathcal{B}_1}$ and $D_{\mathcal{B}_2}$. We denote the posterior distributions $p_1(\boldsymbol{\beta}) \equiv p(\boldsymbol{\beta}|D_{\mathcal{B}_1}, \alpha, f_{\mathcal{A}})$ and $p_2(\boldsymbol{\beta}) \equiv p(\boldsymbol{\beta}|D_{\mathcal{B}_2}, \alpha, f_{\mathcal{A}})$, respectively. “Similarity” of these distribution can be computed as

$$D_{\mathrm{KL}}(p_1, p_2) = \int_{\boldsymbol{\beta}\in\mathcal{W}} p_1(\boldsymbol{\beta}) \ln \frac{p_1(\boldsymbol{\beta})}{p_2(\boldsymbol{\beta})} d\boldsymbol{\beta}. \tag{17}$$

To estimate the minimum sample size m^*, we randomly delete objects from our data set one by one, consequently reducing the sample size m, and computing the posterior distribution of the vector $\boldsymbol{\beta}$ by Eq. (14). Then, the Kullback–Leibler divergence (17) between the probability density functions of parameters evaluated at “similar” data sets is estimated. This process is repeated N times and then the results are averaged. The sample size m^* is considered adequate if the Kullback–Leibler divergence (17) changes less than by some given ε_2 for $m \geq m^*$.

We approximate the integral (17) with a sum. Here, we take the logistic regression parameters vector $\hat{\boldsymbol{\beta}}$, obtained by solving problem (4), as a mean vector, denoted $\boldsymbol{\beta}_0$ in Eq. (16). Then a sample of 500 vectors $\boldsymbol{\beta}$ is generated according to normal distribution $\mathcal{N}(\hat{\boldsymbol{\beta}}, \mathbf{I}_{\mathcal{A}})$. For each of them, we compute $p_1(\boldsymbol{\beta})$, $p_2(\boldsymbol{\beta})$ and from now on will treat a sum

$$\sum_{\boldsymbol{\beta}} p_1(\boldsymbol{\beta}) \ln \frac{p_1(\boldsymbol{\beta})}{p_2(\boldsymbol{\beta})}$$

as the Kullback–Leibler divergence between p_1 and p_2. The paper [13] describes a method of estimating Kullback–Leibler divergence between continuous densities that does not require estimation of probability density functions.

5. Computation experiment

5.1. Experiment on real data

The data set contains observations of concentrations of 20 proteins in blood cells for patients of two classes, containing 31 and 14 objects, respectively. All features, or biomarkers, are listed in the first and third rows of Table 1.

Table 2 presents optimal sets of features, corresponding to maximum AUC values and the exact AUC values. Here, $K = 5$ optimal sets were selected for investigation.

Due to high costs of medical investigation of one patient, it is essential to reduce the number of measured biomarkers. It is suggested to measure only the most informative features. Having united indexes of all the features from Table 2, we obtain a set of indices of the most informative features $\mathcal{S} = \bigcup_{i=1}^{K}\{\mathcal{A}_i\}$. For every feature the number of times that it was involved in $\mathcal{S}$ is computed. Table 1 shows this number for every feature.

Table 1
Number of entries into K optimal sets for each feature.

K	L	K/M	L/M	K/N	K/O	L/O	K/P	L/P	K/Q
5	4	3	1	0	0	0	0	2	1
K/R	L/R	$L/R/SA$	$L/T/SA$	$L/T/SO$	U/V	U/W	U/X	U/Y	U/Z
0	1	0	0	1	0	0	0	0	0

Table 2
The results of feature selection.

$\mathcal{A}$	$S(\mathcal{A})$
$K, L, L/P$	0.9750
$K, L, K/M, K/Q$	0.9671
$K, L, L/M, L/T/SO$	0.9933
$K, L, K/M, L/R$	0.9867
$K, K/M, L/P,$	0.9742

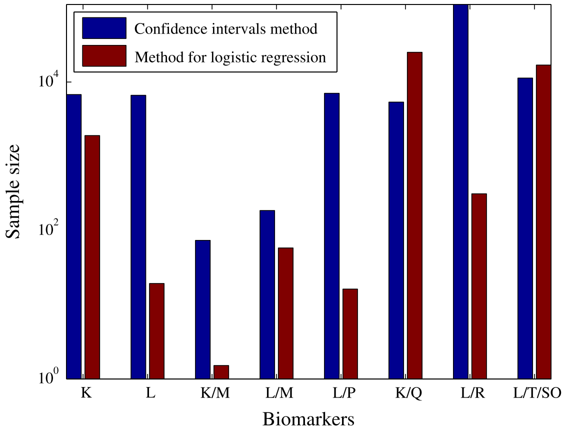


Fig. 3. Sample size estimations computed by method of confidence intervals and method for logistic regression for the most informative features.

Minimum sample size determination. In the histogram of Fig. 3, sample size values m^*, computed for separate features by Eqs. (9) and (11), are represented. The sample size m^* was only computed for those features included in the model, the rest of them are not informative and should not be considered.

We note that the sample size estimations, obtained by Eqs. (9) and (11), have a similar dependence on a feature's index. The reason is that in both methods, the sample size estimation of the jth feature depends on how informative the feature is. In logistic regression, informative features have a significant value of the corresponding element β_j of parameters vector. In Eq. (11), $(p_0 - p_1)^2$ is placed in the denominator. The nearer the parameter β_j tends to zero, the less the value $(p_0 - p_1)^2$ is, and, therefore, the larger m^* is. In this way, minimum values of m^* correspond to the most informative features, whereas abnormally large values ($\sim 10^4$ or more) answer to those features, that are not included in model—they have the smallest β_j values.

The dependence of the value of $\mathrm{RS}(m)$, defined by Eq. (12), on the sample size m is plotted in Fig. 4. Provided with a data set described in Section 5.1, the $\mathrm{RS}(m)$ ratio is unable to reach an asymptote, and the form of the dependence of $\mathrm{RS}(m)$ when $m > 30$ cannot be analyzed, so the estimation given by this method is $m^* \geq 30$.

Fig. 5b depicts the dependence of Kullback–Leibler divergence (17), averaged by $N = 100$ trials, on the sample size m. It is seen, that having more than 25 elements in the data set leads to changing of the Kullback–Leibler divergence relatively slowly: when the sample size $m > 25$ is reduced by one element, the graph shows almost no change of Kullback–Leibler divergence, compared to the area of smaller m. Thus, we obtain a minimum sample size estimation $m^* \simeq 25$.

To compare the results obtained by different methods, we represent them in Table 3. The amount of observations in investigated data is quite small, so the cross-validation method and the method involving Kullback–Leibler divergence computation only provides us with a lower bound of m^*. These methods are more suited for large data sets. The confidence interval method and method of logistic regression show numerically different results, as the confidence interval method is quite rough. However, the dependence of m^* on the feature index is practically the same for these methods, both of them give estimations which depend on how informative the feature is.

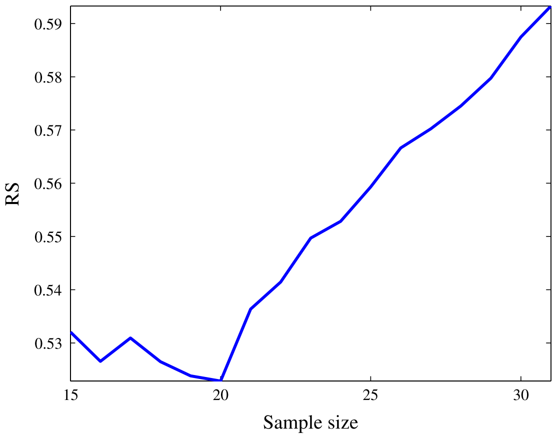


Fig. 4. RS(m) ratio.

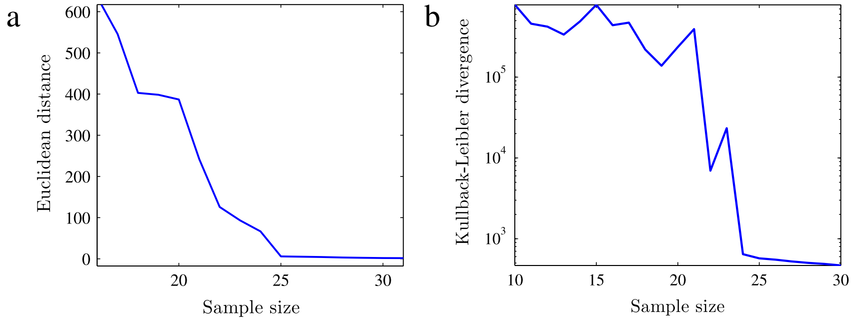


Fig. 5. a. Averaged Euclidean divergence $\|\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m+1}\|$. b. Kullback–Leibler divergence between probability density functions of model parameters.

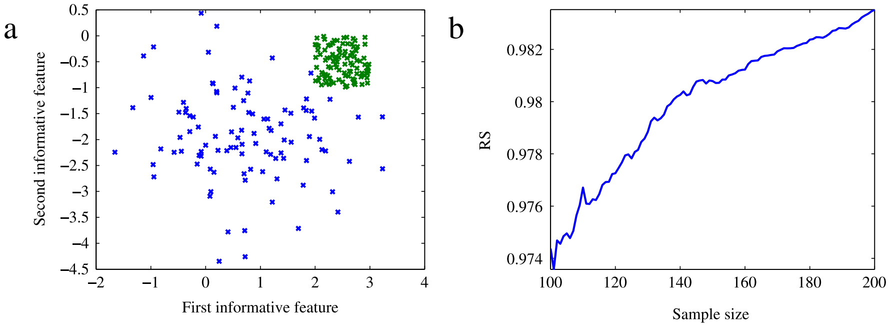


Fig. 6. a. Data set represented by two informative features. b. Dependence of RS ratio on m, obtained with cross-validation 3:1.

Table 3
Sample size estimations.

Confidence intervals	Logistic	Cross-validation	Kullback–Leibler
10^2–10^4	∼100	≥30	≃25

5.2. Experiment on synthetical data

The experiment was also carried out on synthetical data. Each class contains one noisy feature and two informative features (distributed normally and uniformly), and it contains 100 objects. It is seen in Fig. 6a, that classes are easily distinguished.

Furthermore, it is seen in Fig. 6b, that for sample size $m \geq m^* = 100$ the change of RS(m) ratio is not more than 0.01, so we conclude that $m^* \leq 100$.

The results of sample size estimation m^*, obtained by Eqs. (9) and (11), are illustrated by Fig. 7a.

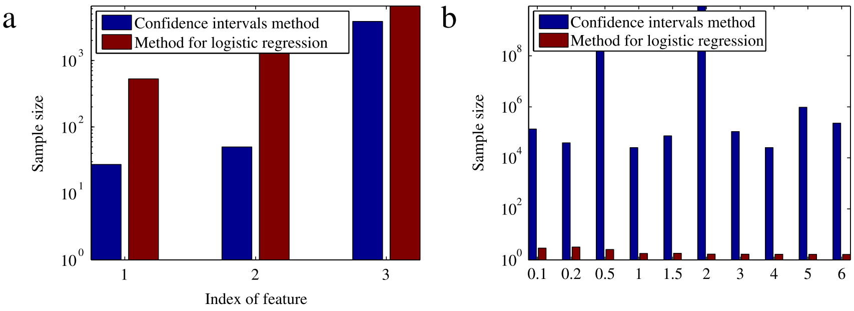
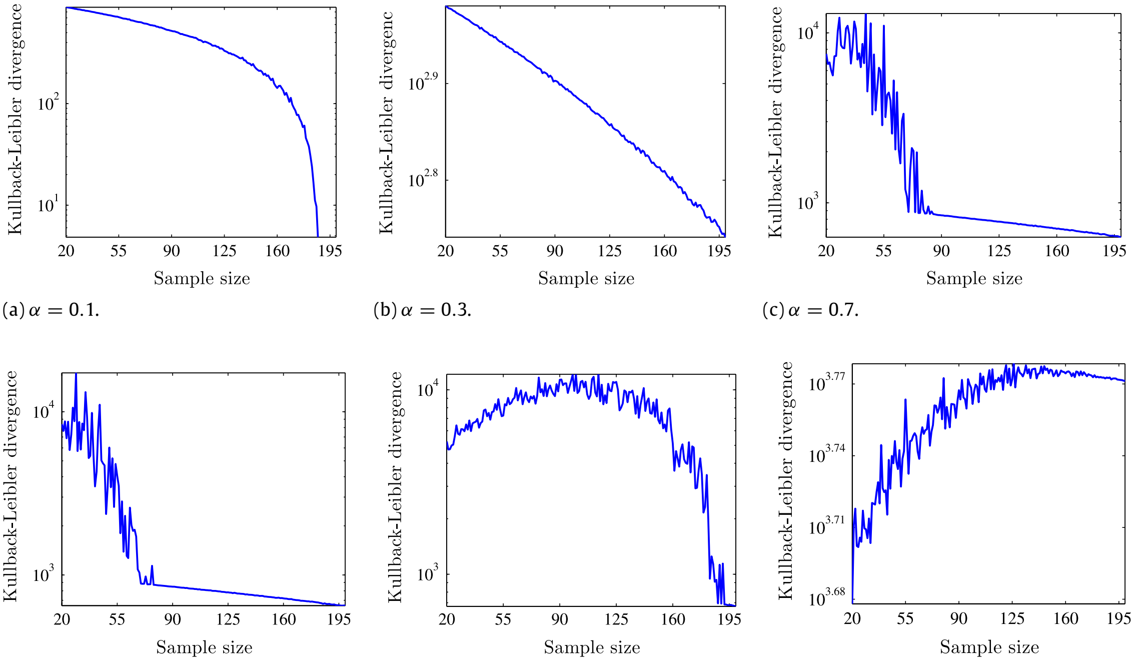


Fig. 7. Sample size m^*, estimated for each model feature by confidence interval method and method of logistic regression. a. Sample size estimation for each of three features describing the data set. b. The data set is described by a single feature x. The variable x follows a mixture of normal distributions with the difference $|\mu_1 - \mu_2|$ between mean values of the components.



(a) $\alpha = 0.1$.

(b) $\alpha = 0.3$.

(c) $\alpha = 0.7$.

(d) $\alpha = 1$.

(e) $\alpha = 1.2$.

(f) $\alpha = 2$.

Fig. 8. The dependence of the Kullback–Leibler divergence on the sample size. The data sample is synthetical, with the same means, but different variance matrices.

In this case, the estimations of m^* given by the confidence interval method are more precise (closer to those obtained by cross-validation). This might happen because the example is too simple. The real data, investigated in Section 5.1, is assumed to follow a mixture of normal distributions (10). To approximate real data, we consider a data set with just one independent variable, distributed according to relation (10). Dependence of sample size estimations on the $|\mu_1 - \mu_2|$ difference is observed. In Fig. 7b, it is seen that in this case Eq. (9) gives overrated results, while estimations of m^*, obtained by Eq. (11) are more adequate.

Let us consider another example. The data consists of two classes, each one is described by two features. The features are distributed normally with diagonal covariance matrix, $\Sigma = \alpha^{-1}I$. We fixate the mean vectors and vary α, computing the dependence of the Kullback–Leibler divergence on the sample size m for different values of α. Fig. 8 presents such dependences for $\alpha = 0.1, 0.3, 0.7, 1, 1.2, 2$. It is seen that when the variance is large ($\alpha = 0.1, 0.3$) the $D_{\mathrm{KL}}(p_1, p_2)$ dependence on sample size m does not have a "footboard", which can be seen on Fig. 5b at $m \approx 25$. Note that the values of Kullback–Leibler divergence are smaller for $\alpha = 0.1, 0.3$ than those in the other experiments. In the case of a high variance, it is hard to classify objects and the AUC value is small, no matter how many samples we have. Fig. 9 shows that the classifier for $\alpha = 0.1$ is close to a random guess, regardless of the sample size, that is why the regression parameters and their probability density function do not change significantly with the sample size m. When the variance is smaller ($\alpha = 0.7, 1$), the values of the Kullback–Leibler divergence get higher, but we are able to estimate the minimum sample size $m \approx 80$. If we keep increasing α, we see that the Kullback–Leibler divergence does also increase with the sample size for $m \leq 125$. This happens because

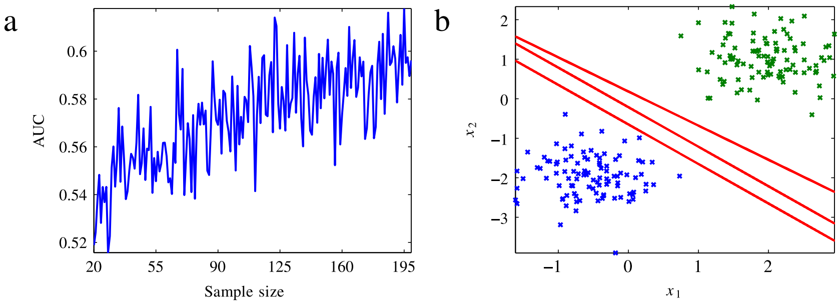


Fig. 9. a. The dependence of the AUC on the sample size, $\alpha = 0.1$. b. Examples of hyperplanes that can separate the data size without misclassified objects, $\alpha = 2$.

when with $\alpha = 1.2, 2$ the classes are so distant from each other and can be distinguished by different hyperplanes. So the position of the hyperplane is rather random for small values of m. Later, when the m value gets higher, the Kullback–Leibler divergence starts to decrease with the sample size m. These examples show that the sample size estimation should be based on the shape of the dependence, rather than on the exact values of the Kullback–Leibler divergence.

6. Conclusion

This paper presents an algorithm that classifies patients with cardio-vascular decease. To select the regression model, an exhaustive search algorithm is used. The authors proposes a new method of sample size determination. It is based on the cross-validation technique and uses the Kullback–Leibler divergence between two distributions of model parameters, evaluated on similar data subsets. Four different algorithms of sample size determination are compared.

By this paper the authors further introduced analytic and probabilistic methods into data mining and statistic learning. In the future, we shall go on with this insertion of modern applied mathematics towards improved prediction in all fields of science, engineering and real life.

References

[1] D. Hosmer, S. Lemeshow, Applied Logistic Regression, Wiley, NY, 2000.
[2] B. Rosner, Fundamentals of Biostatistics, Duxbury Press, 1999.
[3] E. Demidenko, Sample size determination for logistic regression revisited, Statistics in Medicine 26 (2007) 3385–3397.
[4] C.M Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[5] D.J.C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2003.
[6] J. Friedman, T. Hastie, R. Tibshirani, Additve logistic regression: a statistical way of boosting, The Annals of Statistics 28 (2) (2000) 337–407.
[7] B. Efron, et al., Discussion of least square regression, least angle regression, The Annals of Statistics 32 (2) (2004) 465–469.
[8] T. Fawcet, ROC Graphs: Notes and Practical Considerations for Researchers, Kluwer Academic Publishers, 2004.
[9] E. Kurum, K. Yildirak, G.-W. Weber, A classification problem of credit risk rating investigated and solved by optimisation of the ROC curve, Central European Journal of Operations Research (CEJOR) (2012) (special issue at the occasion of EURO XXIV 2010 in Lisbon).
[10] http://www.medicalbiostatistics.com.
[11] J.O. Berger, L.R. Pericchi, Training samples in objective Bayesian model selection, The Annals of Statistics 32 (3) (2004) 841–869.
[12] S. Amari, N. Murata, K.-R. Muller, M. Finke, H.H. Yang, Asymptotic statistical theory of overtraining and cross-validation, IEEE Transactions on Neural Networks 8 (5) (1997) 985–996.
[13] F. Perez-Cruz, Kullback–Leibler divergence estimation of continuous distributions, in: IEEE International Symposium on Information Theory, 2008.