Motrenko A., Strijov V., Weber G.-W. Bayesian sample size estimation for logistic regression // Journal of Computational and Applied Mathematics, 2014, 255 : 743-752.

Sample Size Bayesian Estimation 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*

Abstract

The problem of sample size estimation is important in the medical applications, especially in the cases of expensive measurements of immune biomarkers. The papers describes the problem of logistic regression analysis including model feature selection and includes the sample size determination algorithms, namely 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.

Keywords: logistic regression, sample size, feature selection, Bayesian inference, Kullback-Leibler divergence

1. Introduction

The paper is devoted to the 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 as biomarkers and is classified into two classes. Since the patient measurement is expensive the problem is to reduce number of measured features in order to increase sample size.

The responsive variable is assumed to follow a Bernoulli distribution. Also, parameters of the regression function are evaluated [2, 3].

With given set of features, the model is excessively complex. The problem is to select a set of features of a smaller size, that will classify patients effectively. In logistic regression, features are usually selected by stepwise regression [4, 5]. In the computational experiment, exhaustive search is implemented. This makes the experts sure that all possible combinations of the features were considered. The authors use the area under ROC curve [6] as the optimum criterion in the feature selection procedure.

The problem of classification is associated with minimum sample size determination. In the paper, the following methods are discussed:

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

Email address: `strijov@ccas.ru` (Vadim Strijov)

1. Method of confidence intervals: a method of univariate statistics.
2. Method of sample size evaluation in logistic regression [7, 8]: unlike the previous one, this method considers the distribution of the responsive variable according to the logistic regression model.
3. Cross-validation: a method which evaluates sample size by observing potential overfitting [9, 10].
4. Comparing different subsets of the same sample by computing Kullback-Leibler [11] divergence between probability density functions of model parameters, evaluated at these subsets.

The data, used while conducting computational experiment can be found here [12].

2. Classification problem

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

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

We use the maximim likelihood method, write the error function for (1) as

$$E(\mathbf{w}) = -\ln p(\mathbf{y}|\mathbf{w}) = -\sum_{i=1}^{m} y_i \ln \theta_i + (1 - y_i) \ln (1 - \theta_i). \tag{2}$$

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

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^n} E(\mathbf{w}). \tag{3}$$

Let us define the probability of a case as

$$f(\mathbf{x}_i^T \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{x}_i^T \mathbf{w})} = \theta_i. \tag{4}$$

To solve the problem (3), using

$$\frac{df(\xi)}{d\xi} = f(1 - f),$$

we compute gradient of the error function $E(\mathbf{w})$:

$$\nabla E(\mathbf{w}) = -\sum_{i=1}^{m} \bigl(y_i (1 - \theta_i) - (1 - y_i) \theta_i\bigr) \mathbf{x}_i = \sum_{i=1}^{m} (\theta_i - y_i) \mathbf{x}_i = \mathbf{X}^T (\boldsymbol{\theta} - \mathbf{y}),$$

in which $\boldsymbol{\theta} = [\theta_1, \dots, \theta_m]^T$ and the matrix $\mathbf{X} = \left[\mathbf{x}_1^T, \dots, \mathbf{x}_m^T\right]^T$ represents features sets.

Parameters are evaluated by Newton-Raphson method. Denote by $\mathbf{\Sigma}$ a diagonal matrix with diagonal elements $\Sigma_{ii} = \theta_i(1-\theta_i)$ $(i = 1,\dots,m)$. Set the initial value $\mathbf{w} = [w_1,\dots,w_n]^T$ of $\hat{\mathbf{w}}$

$$w_j = \sum_{i=1}^{m} y_i(1-y_i) \quad (j = 1,\dots,n),$$

Then the $(k+1)$-th iteration of evaluation of $\hat{\mathbf{w}}$ is

$$\begin{aligned}\mathbf{w}_{k+1} = \mathbf{w}_k - (\mathbf{X}^T\mathbf{\Sigma}\mathbf{X})^{-1}\mathbf{X}^T(\boldsymbol{\theta} - \mathbf{y}) = \\ (\mathbf{X}^T\mathbf{\Sigma}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}(\mathbf{X}\mathbf{w}_k - \mathbf{\Sigma}^{-1}(\boldsymbol{\theta} - \mathbf{y})).\end{aligned} \tag{5}$$

The process is repeated until the Euclidean distance $\| \mathbf{w}_{k+1} - \mathbf{w}_k \|$ is sufficiently small.

Thus, the classification algorithm is defined as:

$$a(\mathbf{x}, c_0) = \operatorname{sign}\big(f(\mathbf{x}, \mathbf{w}) - c_0\big), \tag{6}$$

where c_0 is a cut-off value of regression function (4), defined by (7).

Quality of classification. Let us use an additional to (1) quality functional AUC, or the area under the ROC-curve. 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

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

Here, the following denotation is used:

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

Thus, the bigger AUC value is, the better is 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 figure 1, the most distant from segment [(0,0);(1,1)] point of the ROC-curve corresponds to the c_0 value used in (6):

$$\hat{c}_0 = \arg\max_{\xi\in[0,1]} \| \big(\mathrm{TPR}(\xi), \mathrm{FPR}(\xi)\big) - (\xi, \xi) \| = \arg\max_{\xi\in[0,1]} \sqrt{(TPR(\xi) - \xi)^2 - (FPR(\xi) - \xi)^2}. \tag{7}$$

Defining $\hat{c}_0$ includes computing AUC value and, therefore, computation of (6) and iterative estimation of parameters $\mathbf{w}$ according to (5).

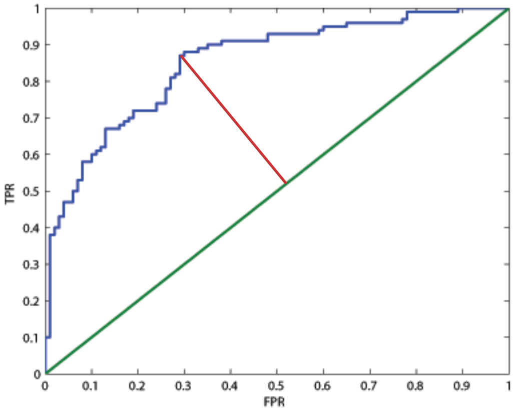


Figure 1: Sample size m^*, estimated by confidence interval method and method for logistic regression.

3. Feature selection problem

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

$$\hat{\mathcal{A}} = \arg\max_{\mathcal{A} \subseteq \mathcal{I}} \mathrm{AUC}(\mathcal{A}), \text{ subject to } |\mathcal{A}| = \text{const}. \tag{8}$$

The value of $\mathrm{AUC}(\mathcal{A}) \equiv \mathrm{AUC}(\mathbf{X}_{\mathcal{A}}, \hat{\mathbf{w}}_{\mathcal{A}}, \hat{c}_0, \mathbf{y})$ is computed for a set $\mathcal{A}$ of indexes and the parameters $\hat{\mathbf{w}}_{\mathcal{A}}$ and c_0 are defined by (3) and (7), respectively.

The maximization problem (8) 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 $\mathbf{w}$ are estimated at $D_{\mathcal{L}}$, while the classification quality is computed at $D_{\mathcal{T}}$. The maximum cardinality of $\mathcal{A}$ is limited by experts: $|\mathcal{A}|$ shall not exceed four elements. We refer to the feature sets, obtained by solving (8), as *optimal sets*, and name the features included into optimal sets as the *most informative features*.

4. Sample size determination

Investigated data describes patients of two classes: those who have already experienced a heart attack and patients that might experience it in future. Concentrations of proteins in blood cells are used as features. There are 31 patients in first class and 14 in the second. Having this few observations we must estimate the minimum sample size m^* required to obtain adequate results of classification. In this chapter four methods of sample size determination are presented. The results of implementing this methods are described and analyzed in Section 5 on computational experiment.

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

$$Z = \frac{\bar{x} - \mu}{\sigma} \sqrt{m} = \frac{\Delta}{\sigma} \sqrt{m} \sim \mathcal{N}(0, 1). \tag{9}$$

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

$$m^* = \left(\frac{z_{\alpha/2} \sigma}{\Delta} \right)^2, \tag{10}$$

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

When $m \geq 30$, the variable Z can be regarded as normally distributed even if the distribution of x_i is different from normal or if σ in (9) is replaced with

$$s = \sqrt{\frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})^2}.$$

Otherwise, it is essential that the variables x_i are normally distributed; moreover the variance σ should be known.

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 (10) 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:

$$x_i \sim \begin{cases} \mathcal{N}(\mu_1, \sigma_1^2), & \text{with probability } \theta_i, \\ \mathcal{N}(\mu_2, \sigma_2^2), & \text{with probability } 1 - \theta_i, \end{cases} \tag{11}$$

where θ_i is defined by (4).

4.2. Method of sample size evaluation in logistic regression.

Fixate a set $\mathcal{A}$ of indexes. 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. Consider the hypothesis

$$H_0 : w_j = 0, \; j \notin \mathcal{A},$$

where w_j being the jth element of the vector $\mathbf{w}$ of logistic regression parameters. In this way, we assume that the jth feature is not included into model. Having estimated the vector of parameters under H_0, we obtain the vector $\mathbf{w}_{\mathcal{A}}$, and under alternative $H_1 : w_j \neq 0$ we get $\mathbf{w}_{\mathcal{A}^*}$, 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

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

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

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

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^*

$$m^* = \frac{p_0 c_0 \left(Z_{1-\alpha/2} + Z_{1-\beta}\sqrt{\frac{p_1 c_1}{p_0 c_0}}\right)^2}{(p_1 - p_0)^2}. \tag{12}$$

Note, that m^*, given by (**??**) 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 $\mathcal{I} = \mathcal{L} \bigsqcup \mathcal{T}$. Fixate a set $\mathcal{A}$ of indexes of model features. Denote by $\text{AUC}(\mathcal{A}, \mathcal{D})$ the quality functional value computed based on the data set $\mathcal{D}$. A decrease of the quality functional $\text{AUC}(\mathcal{A}, D_{\mathcal{T}})$ value computed on the basis of the training set and compared to $\text{AUC}(\mathcal{A}, D_{\mathcal{L}})$ might indicate overfitting. We define overfitting as the following ratio:

$$\text{RS}(m) = \frac{\text{AUC}(\mathcal{A}, D_{\mathcal{T}(m)})}{\text{AUC}(\mathcal{A}, D_{\mathcal{L}(m)})}. \tag{13}$$

In this case, the model f approximates the learning set, but it can not be used to describe the test set. Overfitting might occur when the sample size m is too small. To estimate m^*, we consequentially increase 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 increase of m, the $\text{RS}(m)$ approaches to one. We find the sample size m^* adequate, if for every $m \geq m^*$ the $\text{RS}(m)$ 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 model parameters. Consider two "similar" sets of indexes of objects $\mathcal{B}_1 \in \mathcal{J}$ and $\mathcal{B}_2 \in \mathcal{J}$. Index sets $\mathcal{B}_1$ and $\mathcal{B}_2$ are regarded as "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 also differ. Figure 2 shows how the separating hyperplane given by

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

changes when two elements are added to the sample. If the sample $D_{\mathcal{B}_1}$ is large enough,

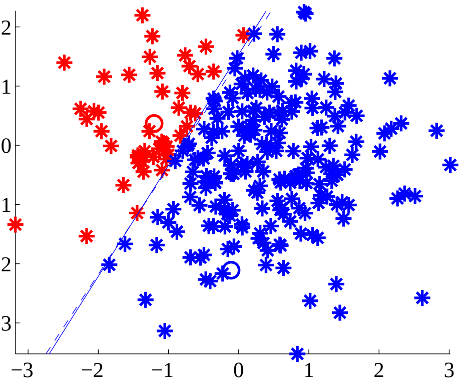


Figure 2: Two classes are separeted by hyperplane. Doted line represents the hyperplane position after the two objects (in circles) were added.

the parameters $\mathbf{w}_1$ evaluated based on $D_{\mathcal{B}_1}$ should not be significantly different from $\mathbf{w}_2$ obtained with a "similar" sample $D_{\mathcal{B}_2}$. The simplest way to compare them is to compute Euclidean distance between $\mathbf{w}_1$ and $\mathbf{w}_2$:

$$\parallel \mathbf{w}_1 - \mathbf{w}_2 \parallel = \sqrt{\sum_{i=1}^{|\mathcal{A}|} (w_i^1 - w_i^2)^2}.$$

In this paper probability density functions of parameters at $D_{\mathcal{B}_1}$ and $D_{\mathcal{B}_2}$ are compared by computing Kullback-Leibler divergence between them. Consider model function (4) and the assumption about the random variable y_i distribution (1). Having fixated the data set D and model $f_{\mathcal{A}} = f(X_{\mathcal{A}}^T\mathbf{w})$, we rewrite (1) as

$$p(\mathbf{y}|X, \mathbf{w}, f_{\mathcal{A}}) \equiv p(D|\mathbf{w}, f_{\mathcal{A}}) = \prod_{i=1}^{m} \theta_i^{y_i}(1-\theta_i)^{1-y_i}. \tag{14}$$

Suppose as well, that the vector of regression parameters $\mathbf{w}$ follows a normal distribution $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \sigma^2 I_{|\mathcal{A}|})$ with the density function

$$p(\mathbf{w}|f_{\mathcal{A}}, \alpha) = \left(\frac{\alpha}{2\pi}\right)^{\frac{|\mathcal{A}|}{2}} \exp\left(-\frac{\alpha}{2} \parallel \mathbf{w} - \mathbf{w}_0) \parallel^2\right), \tag{15}$$

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

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

$$p(\mathbf{w}|D, \alpha, f_{\mathcal{A}}) = \frac{p(D|\mathbf{w}, f_{\mathcal{A}})p(\mathbf{w}|\alpha, f_{\mathcal{A}})}{p(D|\alpha, f_{\mathcal{A}})}, \tag{16}$$

where $p(D|\mathbf{w}, f_{\mathcal{A}})$ is the data likelihood, $p(\mathbf{w}|\alpha, f_{\mathcal{A}})$ given a priori probability density function. In (16), the normalization factor $p(D|\alpha, f_{\mathcal{A}})$ is defined by

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

Substituting (14) and (15) into (16) and denoting $Z(\alpha) = p(D|\alpha, f_{\mathcal{A}})$, we obtain

$$p(\mathbf{w}|D, f_{\mathcal{A}}) = \frac{p(y|\mathbf{x}, \mathbf{w}, f_{\mathcal{A}})p(\mathbf{w}|f_{\mathcal{A}}, \alpha)}{Z(\alpha)} =$$

$$= \frac{\alpha^{\frac{|\mathcal{A}|}{2}}}{(2\pi)^{\frac{|\mathcal{A}|}{2}} Z(\alpha)} \exp(-\frac{\alpha}{2} \parallel \mathbf{w} - \mathbf{w}_0) \parallel^2) \prod_{i=1}^{m} \theta_i^{y_i}(1-\theta_i)^{1-y_i},$$

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

Consider two "similar" samples $D_{\mathcal{B}_1}$ and $D_{\mathcal{B}_2}$. Denote the posterior distributions $p_1(\mathbf{w}) \equiv p(\mathbf{w}|D_{\mathcal{B}_1}, \alpha, f_{\mathcal{A}})$ and $p_2(\mathbf{w}) \equiv p(\mathbf{w}|D_{\mathcal{B}_2}, \alpha, f_{\mathcal{A}})$, respectively. "Similarity" of these distribution can be computed as

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

To estimate the minimum sample size m^*, we randomly delete objects from data set one by one, consequently reducing the sample size m, and computing the posterior distribution of vector $\mathbf{w}$ by (15). Then the Kullback-Leibler divergence (17) between the probability density functions of parameters evaluated at “similar” data sets is computed. This process is repeated N times and then the results are everaged. The sample size m^* is considered adequate if Kullback-Leibler divergence (17) changes less than in ε_2 for $m \geq m^*$.

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. In Table 2 all features, or biomarkers, are listed.

Table 1: 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

Table 1 presents optimal sets of features, corresponding to maximum AUC values and the exact AUC values. Here, $K = 5$ optimal sets were selected for investigation.

Table 2: 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

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

Minimum sample size determination. In the histogram of Figure 3 the sample size values m^*, are computed for separate features by (10) and (12), are represented. The sample size m^* was only computed for those features included in model, the rest of them are not informative and should not be considered.

We note that sample size estimations, obtained by (10) and (12), have a similar dependence on a feature’s index. The reason is that in both methods sample size estimation

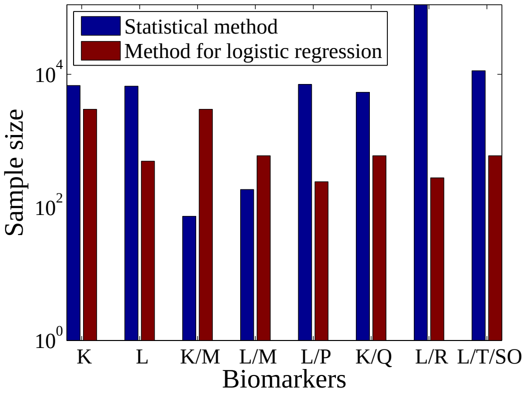


Figure 3: Sample size estimations computed by method of confidence intervals and method for logistic regression for the most informative features.

of jth feature depends on how informative the feature is. In the logistic regression, informative features have a significant value of the corresponding element w_j of parameters vector. In (**??**), $(p_0 - p_1)^2$ is placed in the denominator. The nearer w_j tend to zero, the less $(p_0 - p_1)^2$ the value 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 least w_j values.

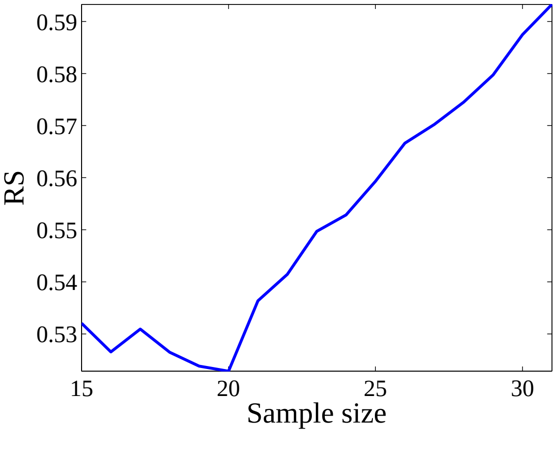
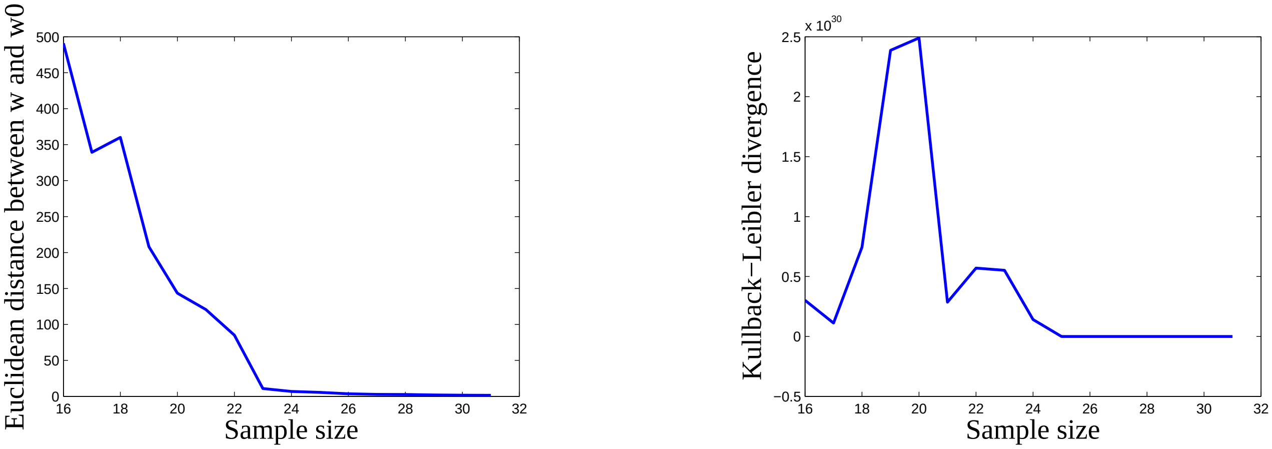


Figure 4: RS(m) ratio.

The dependence of the RS(m), defined by (13) on the sample size m is plotted in Figure 4. Provided with data set, described in Subsection 5.1 the RS(m) ratio is unable to reach an asymptote, and the following form of the dependence RS(m) can not be analyzed, so the estimation given by this method is $m^* \geq 30$.

Figure 5.1 demonstrates the dependence of averaged by $N = 100$ trials Kullback-Leibler (17) divergence on the sample size m is depicted. It is seen, that having more than

27 elements in the data set leads to changing of the Kullback-Leibler divergence relatively slowly: when the sample size $m > 27$ 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 $m^* \simeq 30$.



a

b

Figure 5: a. Averaged Euclidean divergence $||\mathbf{w}_m - \mathbf{w}_{m+1}||$ b.Kullback-Leibler divergence between probability density functions of model parameters.

To compare the results obtained by different methods, we represent them in the Table 3. The amount of observations in investigated data is quite small, so cross-validation method and the method involving Kullback-Leibler divergence computation only provide us with a lower bound of m^*. These methods are more suited for large data sets. 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.

Table 3: Sample size estimations.

confidence intervals	logistic	cross-validation	Kullback-Leibler
$10^2 - 10^4$	∼100	≥ 30	$\simeq 30$

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 Figure 6, that classes are easily distinguished.

Furthermore, it is seen in Figure 7, that for sample size $m \geq m^* = 100$ 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 (10) and (**??**), are illustrated by 8.

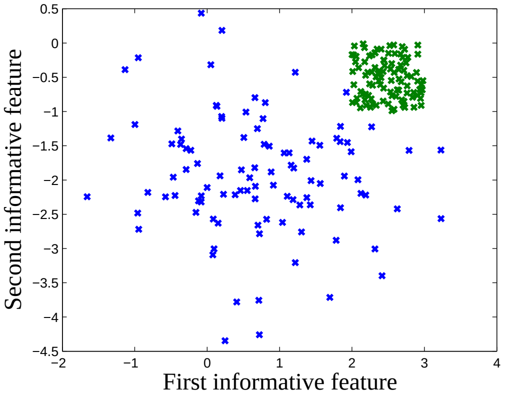


Figure 6: Data set represented by two informative features.

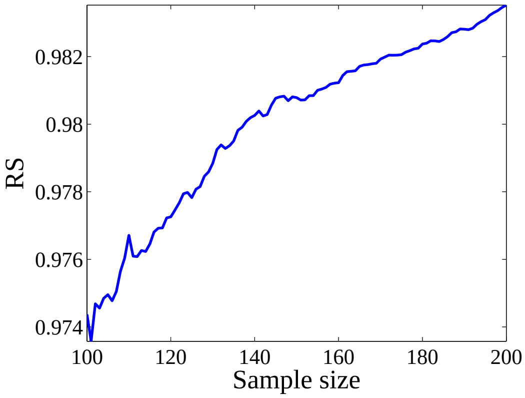


Figure 7: Dependence of RS ratio on m, obtained with cross-validation 3:1.

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 Subsection 5.1 is assumed to follow a mixture of normal distributions (11). To approximate real data, consider a data set with just one independent variable, distributed according to (11). Dependence of sample size estimations on the $|\mu_1 - \mu_2|$ difference is observed. It is seen in Figure 9, that in this case (10) gives overrated results, while estimations of m^*, obtained by (**??**) are more adequate.

6. Conclusion

The paper presents an algorithm that classifies patients with cardio-vascular decease. To select the regression model the exhaustive search algorithm is used. The paper proposes a new method of sample size determination. It is based on cross-validation technique and uses the Kullback-Leibler divergence between two distribution of model parameters,

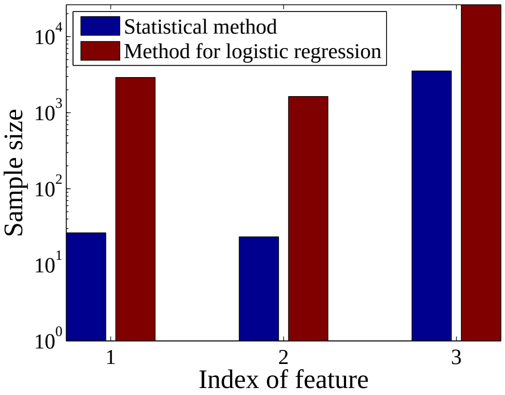


Figure 8: Sample size m^*, estimated for each model feature by confidence interval method and method of logistic regression.

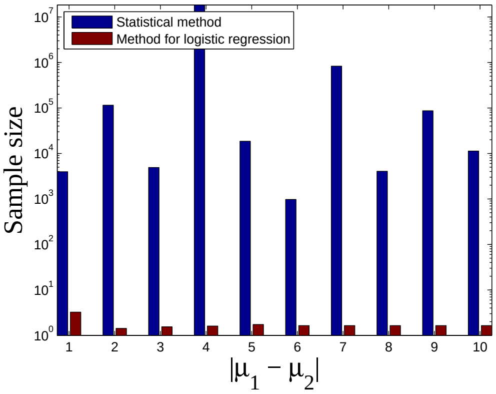


Figure 9: Sample size m^*, estimated by confidence interval method and method of logistic regression.

evaluated on similar data subsets. Four various algorithms os sample size determination are compared.

[1] Hosmer D., Lemeshow S. *Applied logistic regression.* N. Y.: Wiley, 2000. 375 p.

[2] Bishop C. M. *Pattern recognition and machine learning.* Springer, 2006. 738 p.

[3] MacKay D. J. C. *Information theory, inference, and learning algorithms.* Cambridge University Press, 2003. 628 p.

[4] Friedman J., Hastie, Tibshirani R. *Additve logistic regression: a statistical way of boosting* // The Annals of Statistics. 2000. V. 28, No 2. P. 337–407.

[5] Efron B. [et al.] *Discussion of least square regression.* Least Angle Regression // The Annals of Statistics. 2004. V. 32, No 2. P. 465–469.

[6] Fawcet T. *ROC graphs: notes and practical considerations for researchers* // HP Laboratories, 2004. 38 p.

[7] Demidenko E. *Sample size determination for logistic regression revisited* // Statist. Med. 2007; 26:33853397.

[8] Rosner B. *Fundamentals of biostatistics.* Duxbury Press, 1999. 816 p.

[9] Bos S. *How to partition examples between cross-validation set and fraining set?* / Saitama, Japan: Laboratory for information representation RIKEN. 1995. 4 p.

[10] Amari S., Murata N., Muller K.-R., Finke M., Yang H.H. *Asymptotic statistical theory of overtraining and cross-validation.* // IEEE Transactions on Neural Networks, 1997. V. 8, No. 5. P. 985–996.

[11] Perez-Cruz F. *Kullback-Leibler divergence estimation of continuous distributions* // IEEE International Symposium on Information Theory, 2008.

[12] Standart flow cytometry analysis of nondental patients. Paris: ImmunoClin laboratory. 2007. 1 p.