Bayesian sample size estimation for logistic regression[☆]

Anastasiya Motrenko[a], Vadim Strijov[b]

[a] *Moscow Institute of Physics and Technology*
[b] *Computing Center of the Russian Academy of Sciences*

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 the multivariate variable, propose to estimate 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 [**?**], 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 [**?** **?**].

With given set of features, the model is excessively complex. The problem is to select a set of features of smaller size, that will classify patients effectively. In logistic regression features are usually selected by stepwise regression [**?** **?**]. 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 [**?**] 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:

1. Method of confidence intervals [**?**], a method of univariate statistics.

[☆]This project was supported by the Russian Foundation for Basic Research, grant 10-07-00422.
Email address: strijov@ccas.ru (Vadim Strijov)

2. Method of sample size evaluation in logistic regression [**?** **?**]. Unlike the previous one, this method considers the distribution of the responsive variable according to the logistic regression model.
3. Cross-validation, method which evaluates sample size by observing potential overfitting [**?** **?**].
4. Comparing different subsets of the same sample by computing Kullback-Leibler [**?**] divergence between probability density functions of model parameters, evaluated at these subsets.

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

2. Classification problem

Consider the sample set $D = \{(\mathbf{x}_i, y_i)\}, i = 1, \ldots, m$, 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 vector of responsive variables $\mathbf{y} = [y_1, \ldots, y_m]^T$ is a vector of bernullean random variables, $y_i \sim \mathcal{B}(\theta_i)$ with the probability density function

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

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

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

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

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

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

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

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

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

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

The process is repeated until the Euclidean distance $\|\mathbf{w}_{k+1} - \mathbf{w}_k\|_2$ 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 $\mathrm{TPR}(\xi)$, 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 $\mathrm{FPR}(\xi)$, 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 more AUC value is, the better classifier is.

Defining c_0 value. Every point of the ROC-curve corresponds to some c_0 value. As shown in 1, the most distant from segment [(0,0);(1,1)] point of the ROC-curve corresponds to c_0 value used in (6):

$$\hat{c}_0 = \arg\max_{\xi \in [0,1]} \big\| \big(\mathrm{TPR}(\xi), \mathrm{FPR}(\xi)\big) - (\xi, \xi)\big\|_1 . \tag{7}$$

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

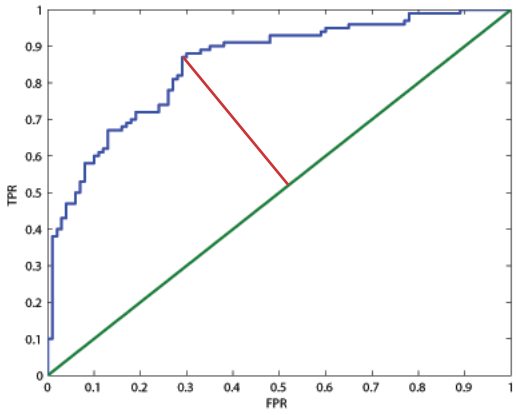


Figure 1: Sample size m^*, estimated by confidence interval method and method for logistic regression.

3. Feature selection problem

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

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

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

The maximization problem (8) is solved in the computational experiment by exhaustive search. This approach is possible due to relatively small amount of features and is required by experts.

As the cardinality of $\mathcal{A}$ is unknown, set of indexes of objects $\mathcal{I}$ is divided into two subsets $\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$, learning set and test set. Parameters $\mathbf{w}$ are estimated at $D_{\mathcal{L}}$, while the classification quality is computed at $D_{\mathcal{T}}$. Maximum cardinality of $\mathcal{A}$ is limited by experts: $|\mathcal{A}|$ shall not exceed four elements. 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 thirty one patients in first class and fourteen in the second. Having this few observations we must estimate 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 the section "Computational experiment".

4.1. Method of confidence intervals

Consider the data set $D = \{(x_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$ 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 x_i is 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, the formula (10) can be used for each feature separately as components of $\mathbf{x}_i$ are assumed to be independent.

This method only helps to obtain 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 hypothesis

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

where w_j — j-th element of vector $\mathbf{w}$ of logistic regression parameters. This way, we assume that j-th feature is not included into model. Having estimated vector of parameters under H_0, we obtain vector $\mathbf{w}_{\mathcal{A}}$, and under alternative $H_1 : w_j \neq 0$ we get $\mathbf{w}_{\mathcal{A}^*}$, where indexes 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 Bernullean distribution $\mathcal{B}(\theta)$ and rewritten as

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

Note that the exact values of θ_i in every 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 hypothesis H_0 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 likelyhood 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 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 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 (12) depends on index j of feature appearing in H_0.

4.3. Cross-validation.

This method provides minimum sample size estimation, based on observing overfitting. When using this approach, 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 $\mathrm{AUC}(\mathcal{A}, \mathcal{D})$ as thye quality functional value, computed at the data set $\mathcal{D}$.

Decrease of thew quality functional $\mathrm{AUC}(\mathcal{A}, D_{\mathcal{T}})$ value computed at training set compared to $\mathrm{AUC}(\mathcal{A}, D_{\mathcal{L}})$ might indicate overfitting. Define overfitting as the following ratio

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

In this case model f approximates learning set, but can't be used to discribe test set. Overfitting might occur when sample size m is too small. To estimate m^*, we consequentially increase sample size m while splitting data set into learning and test sets in a given ratio:

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

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

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 $\mathcal{B}_2 \in \mathcal{J}$. Indexes sets $\mathcal{B}_1$ and $\mathcal{B}_2$ are regarded as "similar" if

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

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

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

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

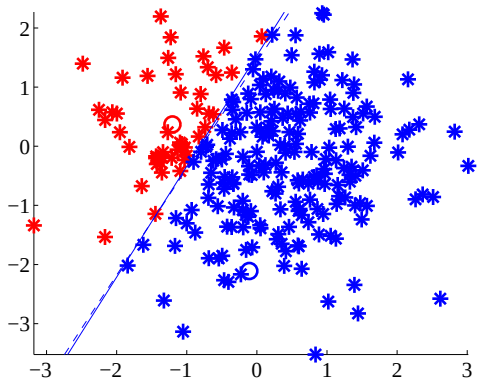


Figure 2: Two classes are separeted by hyperplane. Doted line represents the hyperplane position after the two objects (in circles) were added.

$\mathbf{w}_1$ evaluated at $D_{\mathcal{B}_1}$ should not be significantly different from $\mathbf{w}_2$ obtained at "similar"

sample $D_{\mathcal{B}_2}$. The simplest way to compare them is to compute Euclidean distance between $\mathbf{w}_1$ and $\mathbf{w}_2$:

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

In this paper probability density functions of parameters at $D_{\mathcal{B}_1}$ and $D_{\mathcal{B}_2}$ are compared by computing Kullback-Leibler divergence between them. Consider model function (4) and assumption about the random variable y_i distribution (1). Having fixated the data set D and model $f_{\mathcal{A}} = f(X_{\mathcal{A}}^T \mathbf{w})$, 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 normal distribution $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \sigma^2 I_{|\mathcal{A}|})$ with the density function

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

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

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

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

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 sample size m, and computing the posterior distribution of vector $\mathbf{w}$ by (15). Then Kullback-Leibler divergence (17) between the probability density functions of parameters evaluated at "similar" data sets. 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 the 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

The table 1 presents optimal sets of features, corresponding to maximum AUC values and the exact AUC values. $K = 5$ 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 the table 1, obtain a set of indexes of most informative features $\mathcal{S} = \bigcup_{i=1}^{K} \{\mathcal{A}_i\}$. For every feature from **??** number of times it was involved in $\mathcal{S}$ is computed. The table 2 show this number for every feature.

Minimum sample size determination. To evaluate quality of classification leave one out cross-validation was used. Every object of data set was once in a test set, and was classified by (6). Resutls of this procedure are in the table 3. For every class it's rate of correctly classified objects is presented.

Decrease of quality of classification with decrease of sample size signifies low sample size, that's why computational experiment also includes minimum sample size determination.

Table 3: Rates of correctly classified objects at LOO

A_1	A_3
??	??

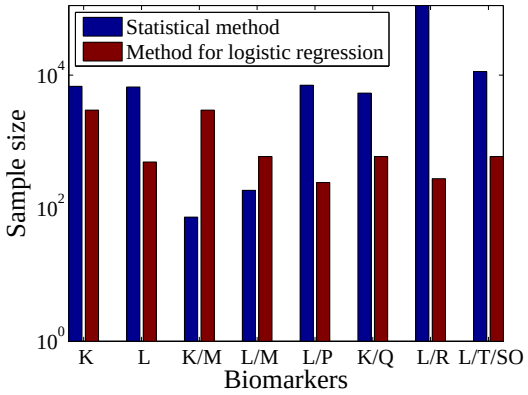


Figure 3: Sample size estimations computed by method of confidence intervals and method for logistic regression for the most informative features.

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

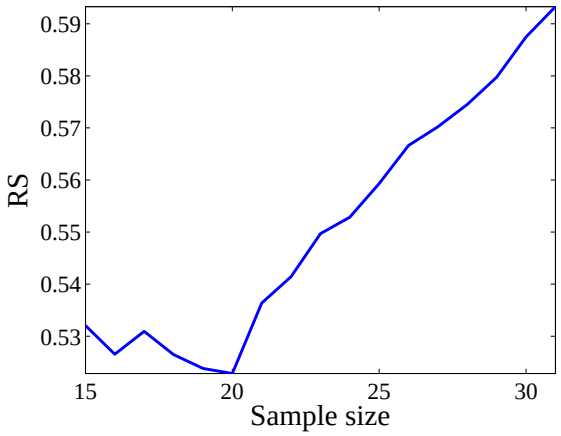


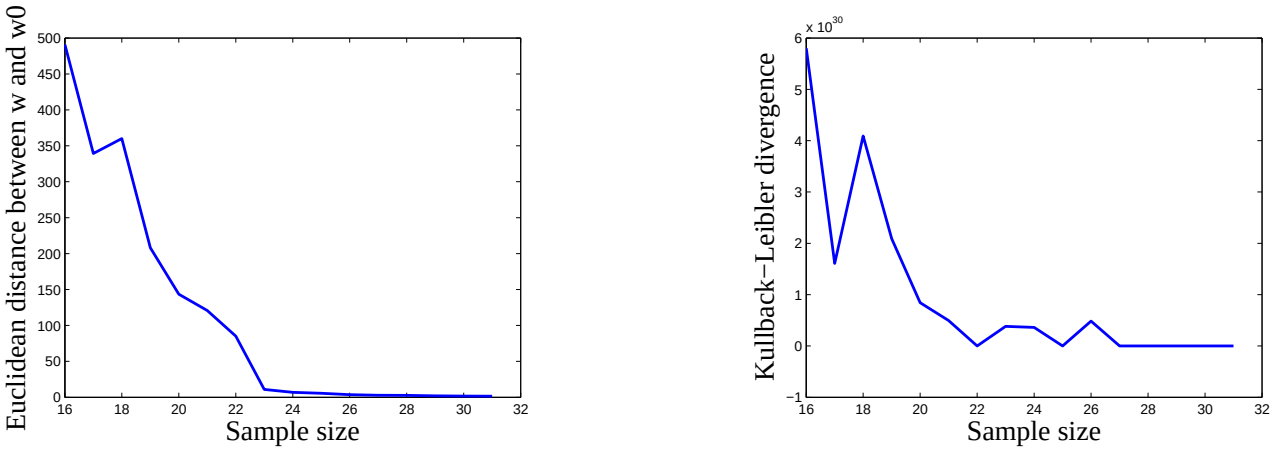
Figure 4: RS(m) ratio.

Note that sample size estimations, obtained by (10) and (**??**) have similar dependence on feature's index. The reason is that in both methods sample size estimation of j-th feature depends on how informative the feature is. In logistic regression informative features have significant value of corresponding element w_j of parameters vector. In (**??**) $(p_0 - p_1)^2$ is placed in denominator. The nearer w_j to zero, the less $(p_0 - p_1)^2$ value is, and therefore,

the larger m^* is. This way, minimum values of m^* correspond to the most informative features, 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.

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

Figure 5.1 shows the dependence of averaged by $N = 100$ trials Kullback-Leibler (17) divergence on sample size m is depicted. It is seen, that having more than 27 elements in data set leads to changing of 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 minimum sample size estimation $m^* \geq 30$.



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 4. The amount of observations in investigated data is quite small, so cross-validation method end method involving Kullback-Leibler divergence computation only provide us with lower bound of m^*.These methods are more suited for large data sets. Confidence interval method and method for logistic regression show numerically different result, as the confidence interval method is quite rough. However the dependence of m^* on feature index is practically the same for these methods, both of them give estimations which depend on how informative the feature is.

Table 4: Sample size estimations.

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

5.2. Experiment on synthetical data.

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

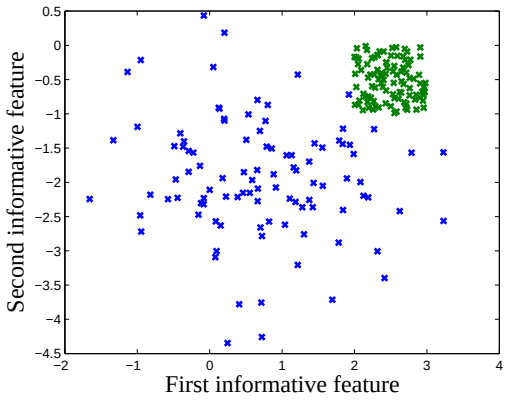


Figure 6: Data set represented by two informative features.

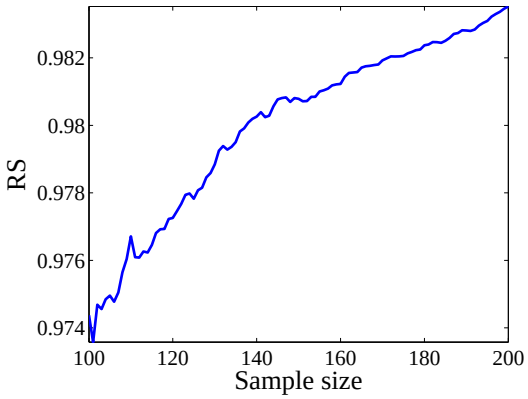


Figure 7: Dependence of RS ratio on m, obtained with cross-validation 3:1.

It is seen in 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.

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

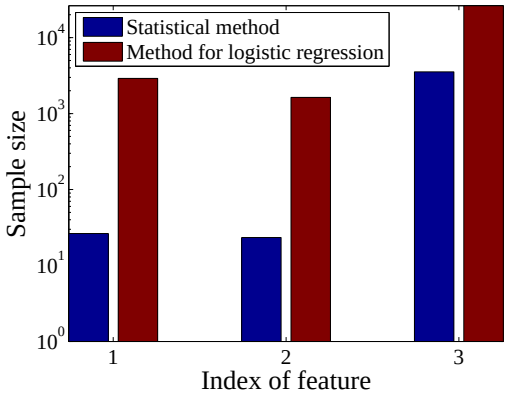


Figure 8: Sample size m^*, estimated for each model feature by confidence interval method and method for logistic regression.

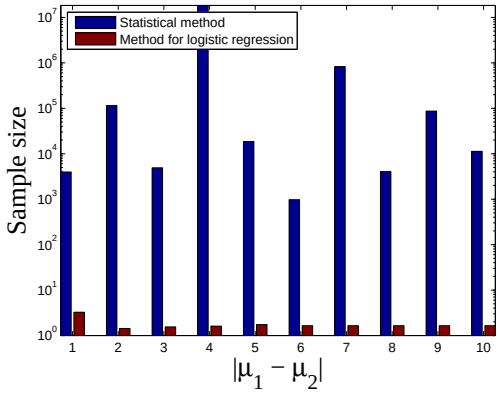


Figure 9: Sample size m^*, estimated by confidence interval method and method for logistic regression.

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

References

[] L. S. Hosmer, D., Applied logistic regression, N. Y.: Wiley, 2000.

[] C. M. Bishop, Pattern recognition and machine learning, Springer, 2006.

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

[] H. T. R. Friedman, J., Additve logistic regression: a statistical way of boosting, The Annals of Statistics 28 (2000) 337–407.

[] R. G. Madigan, D., Discussion of least square regression, The Annals of Statistics 32 (2004).

[] T. Fawcet, Roc graphs: notes and practical considerations for researchers (????).

[] E. Demidenko, Sample size determination for logistic regression revisited, Statist. Med. 26 (2007) 3385–3397.

[] B. Rosner, ????

[] S. Bos, How to partition examples between cross-validation set and fraining set? (????).

[] N. . M. K.-R. . F. M. . Y. H. Amari, S.; Murata, Asymptotic statistical theory of overtraining and cross-validation 8 (1997) 985 – 996.

[] F. Perez-Cruz, Kullback-leibler divergence estimation of continuous distributions, in: IEEE International Symposium on Information Theory.