Motrenko A., Strijov V., Weber G.-W. Bayesian sample size estimation for logistic regression // International Conference on Applied and Computational Mathematics, 2012 : 1-5.

Bayesian sample size estimation for logistic regression

Anastasiya Motrenko[1], **Vadim Strijov**[2], **Gerhard-Wilhelm Weber**[3]

[1] *Moscow Institute of Physics and Technology*
anastasia.motrenko@gmail.com
[2] *Computing Center of the Russian Academy of Sciences* strijov@ccas.ru
[3] *Institute of Applied Mathematics, Middle East Technical University*
gweber@metu.edu.tr

Extended Abstract

1 Introduction.

The paper[1] 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].

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 [3]. 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 [4] 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.
2. Method of sample size evaluation in logistic regression [5]. 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 [6].

[1]Supported by the Russian Foundation for Basic Research, grant 10-07-00422.

4. Comparing different subsets of the same sample by computing Kullback-Leibler [7] divergence between probability density functions of model parameters, evaluated at these subsets.

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

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

Then the classification algorithm is defined as:

$$a(\mathbf{x}, c_0) = \text{sign}\big(f(\mathbf{x}, \mathbf{w}) - c_0\big), \tag{5}$$

where c_0 is a cut-off value of regression function (4).

3 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".

3.1 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

$$Z = \frac{\bar{x} - \mu}{\sigma} \sqrt{m} = \frac{\Delta}{\sigma} \sqrt{m} \sim \mathcal{N}(0, 1). \tag{6}$$

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

$$m^* = \left(\frac{z_{\alpha/2} \sigma}{\Delta} \right)^2, \tag{7}$$

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, the formula (7) can be used for each feature separately as components of $\mathbf{x}_i$ are assumed to be independent.

3.2 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:

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

In this case, the formula for m^* is

$$m^* = \frac{p_0 c_0 \left(Z_{1-\alpha/2} + Z_{1-\beta} \sqrt{\frac{p_1 c_1}{p_0 c_0}} \right)^2}{(p_1 - p_0)^2}. \tag{8}$$

Note that m^*, given by (8) depends on index j of feature appearing in H_0.

3.3 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.

If sample $D_{\mathcal{B}_1}$ is large enough, parameters $\mathbf{w}_1$ 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

$$p(\mathbf{y}|X, \mathbf{w}, f_{\mathcal{A}}) \equiv p(D|\mathbf{w}, f_{\mathcal{A}}) = \prod_{i=1}^{m} \theta_i^{y_i} (1 - \theta_i)^{1-y_i}. \tag{9}$$

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

$$p(\mathbf{w}|f_{\mathcal{A}}, \alpha) = \left(\frac{\alpha}{2\pi}\right)^{\frac{|\mathcal{A}|}{2}} \exp(-\frac{\alpha}{2}||\mathbf{w} - \mathbf{w}_0)||^2), \tag{10}$$

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

$$p(\mathbf{w}|D, \alpha, f_{\mathcal{A}}) = \frac{p(D|\mathbf{w}, f_{\mathcal{A}}) p(\mathbf{w}|\alpha, f_{\mathcal{A}})}{p(D|\alpha, f_{\mathcal{A}})}, \tag{11}$$

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 (11) 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 (9) and (10) into (11) and denoting $Z(\alpha) = p(D|\alpha, f_{\mathcal{A}})$, we obtain

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

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

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

$$D_{\mathrm{KL}}(p_1, p_2) = \int_{\mathbf{w}\in\mathcal{W}} p_1(\mathbf{w}) \ln \frac{p_1(\mathbf{w})}{p_2(\mathbf{w})} d\mathbf{w}. \tag{12}$$

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 (10). Then Kullback-Leibler divergence (12) 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 (12) changes less than in ε_2 for $m \geq m^*$.

4 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

[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] *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.

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

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

[6] *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.

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