Evaluation of Logistic Regression and Neural Network Model With Sensitivity Analysis on Medical Datasets

Raghavendra B.K. *raghavendra_bk@rediffmail.com*
Department of Computer Science & Engineering
Dr. M.G.R. Educational & Research Institute
Chennai, 600 095, India

S.K. Srivatsa *profsks@rediffmail.com*
Senior Professor
St. Joseph's College of Engineering
Chennai, 600 119, India

Abstract

Logistic Regression (LR) is a well known classification method in the field of statistical learning. It allows probabilistic classification and shows promising results on several benchmark problems. Logistic regression enables us to investigate the relationship between a categorical outcome and a set of explanatory variables. Artificial Neural Networks (ANNs) are popularly used as universal non-linear inference models and have gained extensive popularity in recent years. Research activities are considerable and literature is growing. The goal of this research work is to compare the performance of logistic regression and neural network models on publicly available medical datasets. The evaluation process of the model is as follows. The logistic regression and neural network methods with sensitivity analysis have been evaluated for the effectiveness of the classification. The classification accuracy is used to measure the performance of both the models. From the experimental results it is confirmed that the neural network model with sensitivity analysis model gives more efficient result.

Keywords: Artificial Neural Network, Classification Accuracy, Logistic Regression, Medical Dataset, Sensitivity Analysis.

1. INTRODUCTION

In the last few years, digital revolution has provided relatively inexpensive and available means to collect and store large amounts of patient data in databases, i.e., containing rich medical information and made available through the Internet for Health Services globally. Data mining techniques like logistic regression is applied on these databases to identify the patterns that are helpful in predicting or diagnosing the diseases and to take therapeutic measure of those diseases.

Nowadays statistical methods constitute a very powerful tool for supporting medical decisions. The size of medical data that any analysis or test of patients makes that doctors can be helped by statistical models to interpret correctly and to support their decisions. The models are a very powerful tool for doctors and these cannot substitute their viewpoint. On the other hand, the characteristics of medical data and the huge number of variables to be considered as fundamental point for the development of new technique as neural network for the analysis of the data [1].

Neural networks are considered as a field of artificial intelligence. The development of the models was inspired by the neural architecture of human brain. ANN have been applied in many disciplines, including biology, psychology, statistics, mathematics, medical science, and computer science. It has also been applied to a variety of business areas such as accounting and auditing, finance, management and decision making, marketing and production. Recently, artificial neural networks (ANNs) become a very popular model and have been applied to diagnose disease and predict the survival ratio of the patients. However, for medical analysis, ANNs have been shown to have some disadvantages as well as advantages. The most important advantages of ANNs are their discrimination power, detection of complex and nonlinear relationship between independent and dependent variables, and prediction of the case. The ANN model is developed empirically, they can be over-fitted for training data, and their usage is very difficult because of computational requirements. The performance of an ANN depends on the number of parameters, the network weights, the selection of an appropriate training algorithm, the type of transfer functions used, and the determination of the network size. Another disadvantage of using ANNs is that they require the initialization and adjustment of many individual parameters to optimize the classification performance. Many researchers have compared ANN versus LR. Some of them found that ANN and LR have similar classification performance. Compared to LR, neural network models are more flexible [2].

The rest of the paper is organized as follows: Section 2 reviews the prior literature, Logistic Regression technique is discussed in Section 3. Design of neural network is discussed in Section 4. Experimental validation using publicly available medical dataset is given in Section 5. Section 6 includes Experimental results and discussions followed by conclusion.

2. LITERATURE SURVEY

Logistic regression is used in power distribution fault diagnosis, while neural network, has been extensively used in power system reliability researches. Evaluation criteria of the goodness of the classifier includes: correct classification rate, true positive rate, true negative rate, and geometric mean [3].

The features of logistic regression and ANN have been compared and an experiment has been conducted on graft outcomes prediction using a kidney transplant dataset. The results shown reveal that ANN coupled with bagging is an effective data mining method for predicting kidney graft outcomes. This also confirms that different techniques can potentially be integrated to obtain a better prediction. Overall, the results reveal that in most cases, the ANN technique outperforms logistic regression [4].

The author’s presents an evaluation tool for the diagnosis of breast cancer of patients using clinical, pathological, and immunohistochemical data. The main aim was to compare the LR and NN models performances in classification. The neural network approach highlights different inputs from classical statistical model selection [5].

The research work was focuses on a machine learning approach to the classification of LR. The author’s applies a logistic regression based algorithm to three types of classification tasks: binary classification, multiple classifications and classification into a hierarchy. The results confirm that the logistic regression coupled with cross validation is an effective machine learning algorithm. It is not only a robust classification algorithm but also a very effective dimensionality reduction method. The author compared classification performance of logistic regression with several neural network algorithms: backpropagation, fuzzy artificial resonance ART, general regression, radial basis function, self-organizing map-kohonen. The best neural network: the fuzzy artificial resonance network, trained on 12 variables, and achieved 82.6% of correct predictions as compared to 90% for the logistic regression [6].

The research work from the authors was aims to identify the most and least significant factors for breast cancer survival analysis by means of feature evaluation indices derived from multilayer feedforward backpropagation neural networks (MLJFBPNN), fuzzy k-nearest neighbor classifier, and a logistic regression based backward stepwise method (LR). The results appear to suggest that SPF and NPIh appear to be the most and least important prognostic factors, respectively, for survival analysis in breast cancer patients, and should be investigated accordingly in future clinical studies in oncology. The results shown from each method identify a different set of factors as being the most important. It should therefore be suggested that it could be inappropriate to rely on one method's outcome for assessment of the factors, and thus it may be necessary to look at more than one method's outcome for a reliable prognostic factor assessment [7].

In another research work the author compares the performance of LR, NN, and CART decision tree methodologies and to identify important features for the small business credit scoring model on a Croatian bank dataset. The models obtained by all three methodologies were estimated and validated on the same hold-out sample, and their performance is compared. The results shows that the best NN model is better associated with data than LR and CART models [8].

The classification system was developed by the author and it was based on MLFFNN and LR to assess the risk of a family having HNPCC, purely on the basis of pedigree data. The proposed system can eliminate human errors associated with human fatigue and habits. Overall, MLFFNN outperformed to the LR in terms of the number of cases correctly classified and in terms of sensitivity, specificity and accuracy. Two out of 313 cases were misclassified by MLFFNN as opposed to 20 out of 313 by LR [9].

Artificial neural networks can be constructively used to improve the quality of linear models in medical statistics. ANNs are popularly used as universal non-linear inference models and they suffer from two major drawbacks. Their operation is not transparent because of the distributed nature of the representations they form, and this makes it different to interpret what they do. There is no clearly accepted model of generality, which makes it difficult to demonstrate reliability when applied to future data. In this paper neural networks generate hypotheses concerning interaction terms which are integrated into standard statistical models that are linear in the parameters, where the significance of the non-linear terms, and the generality of the model, can be assured using well established statistical tests [10].

The use of Artificial Neural Networks (ANN) is to construct distributions to carry out plausible reasoning in the field of medicine. It describes a comparison between Multivariate Logistic Regression (MLR) and the Entropy Maximization Network (EMN) in terms of explicit assessment of their predictive capabilities. The EMN and MLR have been used to determine the probability of harboring lymph node metastases at the time of initial surgery by assessment of tumor based parameters. Both predictors were trained on a set of 84 early breast cancer patient records and evaluated on a separate set of 92 patient records. Differences in performance were evaluated by comparing the areas under the receiver operating characteristic curve. The EMN model performed more accurately than the MLR model with $AZ = 0.839$, compared to the MLR model with AZ, = 0.809. The difference was statistically significant with two-tailed P value of less than 0.001. Accurate estimation of the prognosis would provide better stratification of patients for further treatment or investigation [11].

3. LOGISTIC REGRESSION

Regression is the analysis, or measure, of the association between a dependent variable and one or more independent variables. This association is usually formulated as an equation in which the independent variables have parametric coefficients that enable future values of the dependent variable to be predicted. Two of the main types of regression are: linear regression and logistic regression. In linear regression the dependent variable is continuous and in logistic it is either discrete or categorical. For logistic regression to be used, the discrete variable must be transformed into a continuous value that is a function of the probability of the event occurring. Regression is used for three main purposes: (1) description, (2) control and (3) prediction [12].

Logistic regression is also called as logistic model or logit model, is a type of predictive model which can be used, when the target variable is a categorical variable with two categories - for example active or inactive, healthy or unhealthy, win or loss, purchase product or doesn't purchase product etc. Logistic regression is used for the prediction of the probability of occurrence of an event by fitting the data into a logistic curve. Like many forms of regression analysis, it makes use of predictor variables; variables may be either numerical or categorical. For example, the probability that a person has a heart attack in a specified time that might be predicted from the knowledge of person's age, sex and body mass index. Logistic regression is used extensively in the medical and social sciences as well as in marketing applications such as prediction of customer's propensity to purchase a product or cease a subscription.

The response, Y, of a subject can take one of two possible values, denoted by 1 and 0 (for example, Y=1 if a disease is present; otherwise Y=0). Let X=(x_1, x_2, . . ., x_n) be the vector of explanatory variables. The logistic regression model is used to explain the effects of the explanatory variables in the form of binary response.

$$\text{Logit}\{\Pr(Y=1|x)\} = \log \{ \Pr(Y=1|x) / (1- \Pr(Y=1|x)) \} = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\ldots+\beta_kx_k \qquad (1)$$

Where β0 is called the intercept" and β_1, β_2, β_3, and so on are called the "regression coefficients" of x_1, x_2, x_3 respectively. Each of the regression coefficients describes the size of the contribution of the risk factor. A positive regression coefficient means that the risk factor increases the probability of outcome, where as a negative regression coefficient means that the risk factor decreases the probability of outcome, a large regression coefficient means that the risk factor strongly influences the probability of that outcome, a non-zero regression coefficient means that the risk factor has little influence on the probability of outcome [13].

The logistic function is given by

$$P=1/(1+e^{-\text{logit(p)}}) \qquad (2)$$

A graph of the function is shown in Figure 1. The logistic function is useful because it can take an input any value from negative infinity to positive infinity, whereas the output is confined to values between 0 and 1.

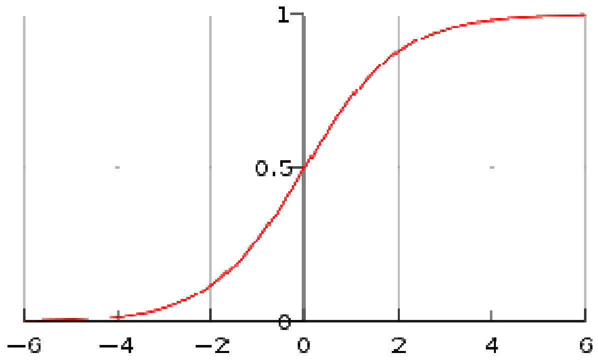

FIGURE 1: A graph of logistic regression function

4. DESIGN OF NEURAL NETWORK

A Neural network is a complex nonlinear modeling technique based on a model of a human neuron. A neural net is used to predict outputs (dependent variables) from a set of inputs (independent variables) by taking linear combinations of the inputs and then making nonlinear transformations of the linear combinations using activation function. It can be shown theoretically that such combinations and transformations can approximate virtually any type of response function. Thus, neural nets use large numbers of parameters to approximate any model. Neural nets are often applied to predict future outcome based on prior experience. For example, a neural net application could be used to predict who will respond to a direct mailing.

Neural networks are becoming very popular with data mining practitioners, particularly in medical research, finance and marketing. This is because they have proven their predictive power through comparison with other statistical techniques using real data sets. The example of a simple feed forward neural network with two layers is shown in Figure 2.

There are two main types of neural network models: supervised neural networks such as the multi-layer perceptron or radial basis functions, and unsupervised neural networks such as

Kohonen feature maps. A supervised neural network uses training and testing data to build a model. The data involves historical data sets containing input variables, or data fields, which correspond to an output. The training data is what the neural network uses to “learn” how to predict the known output, and the testing data is used for validation. The aim is for the neural networks to predict the output for any record given the input variables only [14].

One of the simplest feedforward neural networks (FFNN), such as the one in Figure 2, consists of two layers: an input layer, and output layer. In each layer there are one or more processing elements (PEs). PEs are meant to simulate the neurons in the brain and this is why they are often referred to as neurons or nodes. A PE receives inputs from either the outside world or the previous layer. There are connections between the PEs in each layer that have a weight (parameter) associated with them. This weight is adjusted during training. Information only travels in the forward direction through the network - there are no feedback loops.

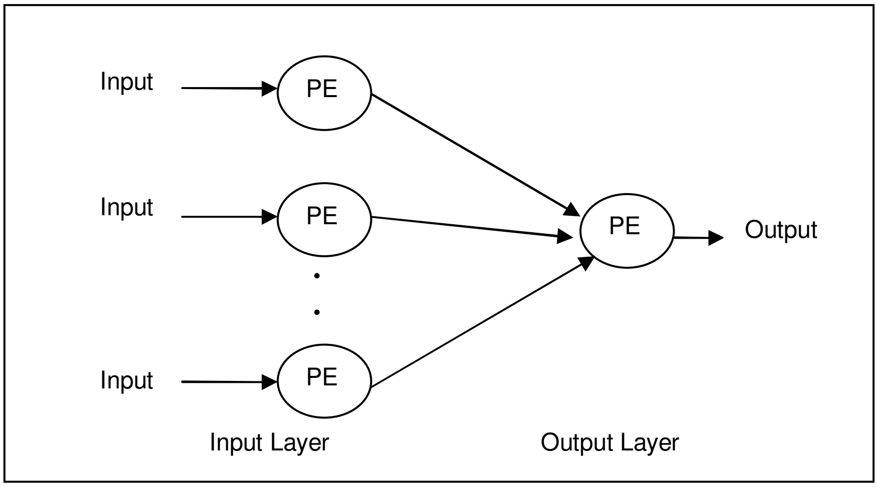

FIGURE 2: Example of a simple feed forward neural network with two layers

The simplified process for training a FFNN is as follows:

1. Input data is presented to the network and propagated through the network until it reaches the output layer. This forward process produces a predicted output.
2. The predicted output is subtracted from the actual output and an error value for the networks is calculated.
3. The neural network then uses supervised learning, which in most cases is backpropagation, to train the network. Backpropagation is a learning algorithm for adjusting the weights.
4. Once backpropagation has finished, the forward process starts again, and this cycle is continued until the error between predicted and actual outputs is minimized [13].

5. EXPERIMENTAL VALIDATION

The framework for neural network model with sensitivity analysis is shown in Figure 3. The process of evaluation is as follows. Sensitivity analysis has been done for the selected features from the dataset. The logistic function with steepness parameter (σ) is calculated using the following equation.

$$P=1/\ (1+e^{-\mathrm{logit}\ (p)\ *\ \sigma}) \qquad (3)$$

where σ=2, 3

The response Y is then calculated as follows by using threshold (τ). Then the probability is calculated to develop a predictive model for classification using neural network. A tenfold cross validation has been used for evaluation on all publicly available medical dataset [15].

$$Y = \begin{cases} 1 & \text{if } P \geq \tau \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

where τ=0.2, 0,4, . . .

In 10-fold cross validation, the original sample is partitioned into 10 sub samples, of the 10 sub samples, a single sub sample is retained as the validation data for testing the model, and the remaining 9 sub samples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 sub samples used exactly once as the validation data. The 10 results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub sampling is that all observations are used for both training and validation and each observation is used for validation exactly once.

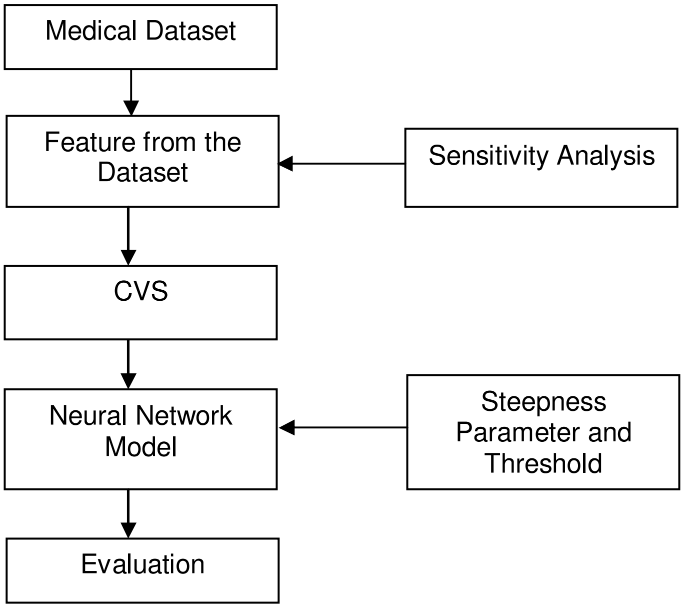

FIGURE 3: Neural network model with sensitivity analysis framework

6. RESULTS AND DISCUSSION

We have used publicly available medical datasets for our experiments whose technical specifications are as shown in Table 1. All the chosen datasets had at least one or more attributes that were continuous. The classification accuracy is used to measure the performance of logistic regression and neural network model on publicly available medical datasets. The results of the evaluation are given in Table 2. Figure 4 gives classification accuracy details after evaluation process. From the results it can be observed that the neural network model with sensitivity analysis gives more efficient result.

Sl. No	Medical Dataset	No of instances	Total no. of attributes	No of classes
1	Asthma	2464	5	2
2	Blood-transfusion	748	5	2
3	Flushot	159	4	2
4	Haberman	306	4	2
5	Liver-disorders	345	7	2
6	Spect test	187	23	2
7	Echocardiagram	132	11	2

TABLE 1: Specifications for the medical datasets

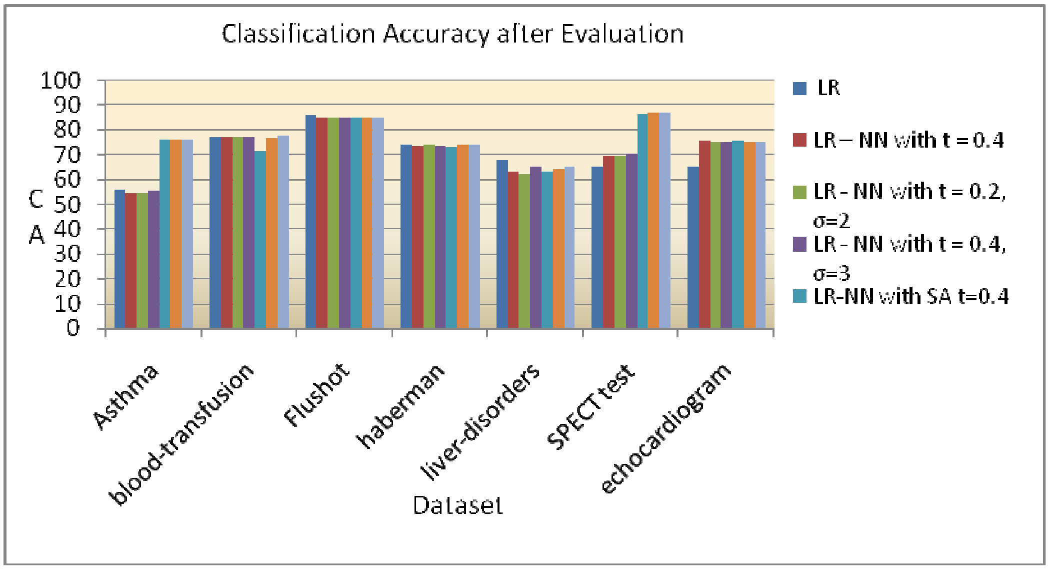

FIGURE 4: Classification accuracy after evaluation

Sl. No.	Name of the Dataset	LR	LR NN with t = 0.4	LR NN with t=0.2, σ=2	LR NN with t=0.4, σ=3	LR NN with SA, t=0.4	LR NN with SA, t=0.4, σ=2	LR NN with SA, t=0.4, σ=3
1	Asthma	56.16	54.58	54.62	55.19	76.29	76.29	76.33
2	blood-transfusion	77.13	77.4	77.27	77.27	71.65	76.6	77.94
3	Flushot	86.16	85	85	85	85	85	85
4	haberman	74.18	73.52	73.85	73.52	73.2	74.18	73.85
5	liver-disorders	68.11	63.18	62.31	65.21	63.18	64.34	65.21
6	SPECT test	65.24	69.51	69.51	70.58	86.63	87.16	87.16
7	echocardiogram	65.15	75.75	75	75	75.75	75	75

TABLE 2: Logistic regression and neural network specification with sensitivity analysis for the medical datasets

7. CONCLUSION

Finally, the conclusions of this work are based on the publicly available medical data sets. The results of this study are more promising. In this research work an attempt was made to evaluate logistic regression and neural network model with sensitivity analysis on publicly available medical data sets. The classification accuracy is used to measure the performance of both the models. From the experimental results it is confirmed that neural network model with sensitivity analysis gives more efficient result.

8. REFERENCES

[1]. Luis Mariano Esteban Escaño, Gerardo Sanz Saiz, Francisco Javier López Lorente, Ángel Borque Fernando and José Moría Vergara Ugarriza, “Logistic Regression Versus Neural Networks for Medical Data”, Monografías del Seminario Matemático García de Galdeano 33, 245-252, 2006.

[2]. Bahar Tasdelen, Sema Helvaci, Hakan Kaleagasi, Aynur Ozge, "Artificial Neural Network Analysis for Prediction of Headache Prognosis in Elderly Patients", Turk J Med Sci 2009; 39(1); 5-12.

[3]. LeXu, Mo-Yuen Chow, and Xiao-Zhi Gao, "Comparisons of Logistic Regression and Artificial Neural Network on Power Distribution Systems Fault Cause Identification", Proceedings of 2005 IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications (SMCia/05), Helsinki, Finland, June 28-30, 2005.

[4]. Fariba Shadabi and Dharmendra Sharma, "Comparison of Artificial Neural Networks with Logistic Regression in Prediction of Kidney Transplant Outcomes", Proceedings of the 2009 International Conference of Future Computer and Communication (ICFCC), 543-547, 2009.

[5]. V.S. Bourdes, S. Bonnevay, P.J.G. Lisbosa, M.S.H. Aung, S. Chabaud, T. Bachelot, D. Perol and S. Negrier, "Breast Cancer Predictions by Neural Networks Analysis: a Comparison with Logistic Regression", Proceedings of the 29th International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, August 23-26, 2007, 5424-7.

[6]. Jack R. Brzezinski George J. Knaft, "Logistic Regression Modeling for Context-Based Classification", DEXA Database and Expert Systems Applications Workshop, 1999.

[7]. Seker H., Odetayo M., Petrovic D., Naguib R.N.G., Bartoli C., Alasio L., Lakshmi M.S., Sherbet G.V. (2002), "An Artificial Neural Network Based Feature Evaluation Index for the Assessment of Clinical Factors in Breast Cancer Survival Analysis", Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering.

[8]. Marijana Zekic-Susac, Natasa Sarlija, Mirta Bensic, "Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network, and Decision Tree Models", 26th International Conference on Information Technology Interfaces (ITI 2004), Cavtat, Croatia, 265-270.

[9]. M Münevver Köküuer, Raouf N. G. Naguib, Peter Jančovič, H. Banfield Younghusband and Roger Green, "A Comparison of Multi-Layer Neural Network and Logistic Regression in Hereditary Non-Polyposis Colorectal Cancer Risk Assessment", Proceedings of the 2005 IEEE Engineering in Medicine and Biology , 27th Annual Conference, Shanghai, China, September 2005, 2417-2420.

[10]. Lisbosa P.J.G., and H. Wong (2001), "Are neural networks best used to help logistic regression? An example from breast cancer survival analysis", IEEE Transactions on Neural Networks, 2472-2477.

[11]. Poh Lian Choong, and Christopher J.S. DeSilva (1996), "A Comparison of Maximum Entropy Estimation and Multivariate Logistic Regression in the Prediction of Axillary Lymph Node Metastasis in Early Breast Cancer Patients", The 1996 IEEE International Conference on Neural Networks, 1468-1473.

[12]. Neter J., Kutner M.H., Nachtsheim C.J., Wasserman W., Applied Linear Regression Models, 3rd Ed. 1996, Irwin, USA (ISBN 0-256-08601-X).

[13]. http://en.wikipedia.org/wiki/Logistic_regression

[14]. Portia A. Cerny, 2001, Datamining and Neural Networks from a Commercial Perspective, Auckland, New Zealand Student of the Department of Mathematical Sciences, University of Technology, Sydney, Australia.

[15]. C.L. Blake, C.J. Merz, “UCI repository of machine learning databases”. [http://www.ics.uci.edu/~mlearn/ MLRepository.html], Department of Information and Computer Science, University of California, Irvine.