Journal of Data Science 12(2014),563-574

Softmax Model as Generalization upon Logistic Discrimination Suffers from Overfitting

F. Mohammadi Basatini[1] and Rahim Chinipardaz[2*]

[1] *Deartment of Statistics, Shoushtar Branch, Islamic Azad University.*
[2] *Deartment of Statistics, Shahid Chamran Universityn.*

Abstract: The motivation behind this paper is to investigate the use of Softmax model for classification. We show that Softmax model is a nonlinear generalization for the logistic discrimination, that can approximate the posterior probabilities of classes where other Artificial neural network (ANN) models don't have this ability. We show that Softmax model has more flexibility than logistic discrimination in terms of correct classification. To show the performance of Softmax model a medical data set on thyroid gland state is used. The result is that Softmax model may suffer from overfitting.

Key words: Neural network model, Softmax model, logistic discrimination, overfitting.

1. Introduction

Discrimination and classification analysis are two multivariate techniques, which separate distinct observation sets and allocate a new observation to preidentified set of classes. In classification and discrimination, there are some explanatory or independent variables with a dependent variable, which is a categorical variable showing the class of observations. The purpose is to investigate a suitable technique for assigning new observations to one of the classes. Many classification methods have been developed and have been used, such as K-nearest neighbor, logistic discrimination, feed forward neural networks, support vector machine and learning vector quantization. Nevertheless, some of these techniques have disadvantages (Al-Daoud, 2009).

Logistic discrimination is one of the most popular methods for classification based on likelihood function of classes. This method was generalized by Anderson (1972) and then he obtained parameters to this method by different forms of sampling. Anderson (1975) also introduced quadratic logistic discrimination. Anderson and Richardson (1979) introduced an effective method for bias correction to obtain parameters to this method. Later Albert and Anderson (1984) studied existence or not existence of estimating parameters to this method. Cox and Ferry (1991) and Pearce (1996) introduced a powerful logistic discrimination. In

* Corresponding author.

logistic discrimination bayes rule is used to obtain posterior probabilities of the classes. In this procedure, each observation is allocated to the class which has higher posterior probability. This allocation is optimum (Webb, 2002).

Nowadays, statistical methods have constituted a very powerful tool to support medical decisions. Data mining techniques like logistic discrimination are applied to medical data to identify the patterns that are helpful in predicting or diagnosing the diseases and taking therapeutic measure of those diseases. Medical data and their statistical analysis are very powerful tools for doctors in interpreting property and supporting their decision. As in medical data we involve with the huge numbers of variables to be considered, the development of new techniques in the statistical analysis, as neural networks, are required (Esteban, et al., 2006). Neural networks is considered as a field of artificial intelligence. The development of the models was inspired by the neural architecture of the human brain. ANN models have been applied for many disciplines, including biology, statistics, mathematics, medical science, computer science, finance, management, and marketing. ANN models are well-known for capturing the complex non-linear relations present in data. ANN can be constructively used to improve the quality of linear models in medical data set. Raghavendra and Srivatsa (2011) reviewed the literature in the field of using logistic discrimination and artificial neural network model in medical databases. Logistic discrimination model has poor performance in many cases since it uses a hyperplane to separate classes. ANN models become very popular in recent years for classification and because of high flexibility of these methods, they have good results in classification. In this paper, we show that Softmax model, as a special case of ANN models, can be considered as a generalization of logistic discrimination, and so we set a statistical support for Softmax neural network model; and we also show that Softmax model has better results than the logistic discrimination, although this model may be suffered from over fitting.

The rest of the paper is organized as follows: Section 2 is dedicated to logistic discrimination. Artificial neural network models are discussed in Section 3. Section 4 provide the investigation of Softmax model. In Section 5, we analyze results on medical data set, and the conclusions of the paper are given in the last Section.

2. Logistic discrimination

Logistic discrimination is a predictive model with a categorical target variable which can be used as the prediction of the posterior probability of the classes. Suppose there exists J class $G_1,\ldots, G_J$ and the observation $\mathbf{x} = (x_1, \ldots, x_p)'$, has to be classified (the elements of $\mathbf{x}$ are explanatory variables) to one of the these classes. In logistic discrimination, one of the classes is considered as basis class and the ratio of other classes are modeled toward this basis class. Without loss of generality, we select class J as basis class then the essential assumption of logistic discrimination for class k can be written as:

$$\ln\left[\frac{l(x|G_k}{l(x|G_J}\right] = \omega_{0k}^{*} + \sum_{i=1}^{p} \omega_{ik} x_i, \quad k \in \{1,2,\ldots,J-1\} \tag{1}$$

where $l(\mathbf{x}|\ G_k\ )$ is the likelihood function to the class k, $k \in \{\ 1,2,\dots,J-1\ \}$and $\omega_{0k}^*, \omega_{1k}, \dots, \omega_{pk}, k \in \{1,\dots,J-1\}$ are the parameters to be estimated from training set of data. It could be seen that in the equation (1) the ratio of likelihood functions is modeled by a linear function of the observation which is a hyperplane. Therefore, although the logistic discrimination doesn't impose any assumption on the likelihood functions, but the ratio of them has been considered as parametric function of observation. Anderson(1972, 1975) proved that the equation (1) can be employed for different families of statistical distributions such as multivariate normal distribution with common covariance matrix and multivariate discrete distributions that follow loglinear model with same interaction terms. From equation (1) we have:

$$\ln\left[\frac{l(x|G_k}{l(x|G_J}\right] = \exp(\omega_{0k}^* + \sum_{i=1}^{p} \omega_{ik}x_i),\ \ k \in \{1,2,\dots,J-1\}. \qquad (2)$$

Using the Bayesian methodology, let π_k, $k \in \{1,\dots,J\}$ be the prior probability of G_k. Then $p(G_k|\mathbf{x}) = \frac{l(x|G_k)\pi_k}{p(x)}$, $k \in \{1,\dots,J\}$ $p(G_k|\mathbf{x})$ is the posterior probability for the class k conditioned on observation $\mathbf{x}$ and $p(\mathbf{x}) = \sum_{j=}^{J} l(\mathbf{x}|G_j)\pi_j$ is the marginal density function of $\mathbf{x}$. So from (2) we obtain:

$$\frac{p(G_k|\mathrm{x})}{p(G_j|\mathrm{x})} = \exp(\omega_{0k}^* + \sum_{i=1}^{p} \omega_{ik}\mathrm{x}_i)$$

and $\omega_{0k} = \omega_{0k}^* + \ln(\pi_k/\pi_j)$. If the classes cover all the observations space, then we have to write $\sum_{k=1}^{J} P(G_k|\mathrm{x}) = 1$ and from above equation will get:

$$P(G_k|\mathrm{x})\ = \frac{\exp\left(\omega_{0\mathrm{k}} + \sum_{i=1}^{p} \omega_{ik}x_i\right)}{1 + \sum_{\mathrm{j}=1}^{\mathrm{J}-1} \exp\left(\omega_{0k} + \sum_{i=1}^{p} \omega_{ik}\boldsymbol{x}_i\right)}, \qquad k \in \{1,2,\dots,J-1\} \qquad (3)$$

$$P(G_j|x)\ = \frac{1}{1 + \sum_{j=1}^{J-1} \exp(\omega_{0k} + \sum_{i=1}^{p} \omega_{ik}\boldsymbol{x}_i)}. \qquad (4)$$

Equations (3) and (4) show logistic function: the logistic function is useful because it can take an input of any value from negative infinity to positive infinity, whereas the output is confined to values between 0 and 1 (Raghavendra and Srivatsa, 2011). In logistic discrimination after obtaining posterior probabilities, the bayes optimum discrimination rule is used for classification. Any observation is allocated to the class with highest posterior probability. The boundaries between the classes, decision boundaries, are hyperplanes and can be obtained from equation $p(G_k|\mathrm{x}) = P(G_j|\mathrm{x})$ for $k, j \in \{1,\dots,J\}$ as bellow:

1) For classes'k and J where k

$$\omega_{0k} + \sum_{i=1}^{p} \omega_{ik} \mathrm{x}_i = 0 \quad (5)$$

2) For classes *m* and *n* where $m \neq n < J$ the decision boundary to consist of:

$$(\omega_{0m} - \omega_{0n}) + \sum_{i=1}^{p} (\omega_{im}\omega_{in})\mathbf{x}_i = 0 \quad (6)$$

As can be seen in logistic discrimination the decision boundaries between all classes are linear.

3. Artificial Neural Network models

Artificial neural network models are computing systems made up of the large number of simple, highly interconnected processing units (neurons) that abstractly emulate the structure and operation of biological nervous system(Subasi and Ercelebi, 2005). Every model has a set of units, which arranged in input, hidden and output layers. An artificial neural network model is a complex nonlinear modeling is used to predict output layers (dependent variables) from a set of input layers (independent variables) by taking linear combination of inputs and then making nonlinear transformations of the linear combinations using activation function. It can be shown that such combinations and transformations can approximate any type of response function. These methods are particularly valuable when ANN models use input variables in the first layer and Network outputs is a solution to a problem; in the classification problems, network output shows the observation class, several hidden layers can be placed between input and output layer. Neural networks can be broadly classified into three categories, namely, feedforward neural networks, feedback neural networks and the combination of both feedforward and feedback neural networks (Rao, 2011).

The multilayer perceptron (MLP) model is a kind of feedforward neural network that can be used for classification and function approximation tasks. The architecture of MLP may contain two or more layers. A simple two-layer ANN consists only of an input layer containing the input variables for the problem and output layer containing the solution for the problem. This type of network is a satisfactory approximation for linear problems. However, for approximating nonlinear systems, additional intermediate (hidden) processing layers are employed to handle the problem's nonlinearity and complexity, (Subasi and Ercelebi, 2005). In MLP model units connected in successive layers by one way forward connections. Figure (1) shows an MLP model with a hidden layer. In classification, the number of input layer units is equal in the number of explanatory variables, and the number of output layer units is equal to the number of classes. The number of hidden layer units is a problem which is, to some extent, difficult to solve and usually specified by trial and error such that minimum misclassification will be obtained. In general for MLP model, the weights are adjusted to realize the global minimum of the total error in the training data on the weight space. Irrespective of topology of the MLP, minimization of the training error leads to the optimization of the performance of their respective tasks. The designed tasks may be either classification or the function

approximation. The backpropagation learning algorithm is used for adjusting the weights within the network to minimize the mean squared error in the output (Rao, 2011).

Although it depends upon the complexity of the function or process being modeled, one hidden layer may be sufficient to map an arbitrary function to any degree of accuracy, (Subasi and Ercelebi, 2005). Hence three-layer architecture MLP model have been adopted for the present study. Equation (7) shows the multilayer perceptron model with one hidden layer and identity function in output layer units

$$g_k(x) = \omega_{0k}^o + \sum_{r=1}^{l} \omega_{rk}^o \, \varphi\left(\omega_{0r}^h + \sum_{i=1}^{p} \omega_{ir}^h x_i\right) \qquad k \in \{1,2,\dots,J\} \tag{7}$$

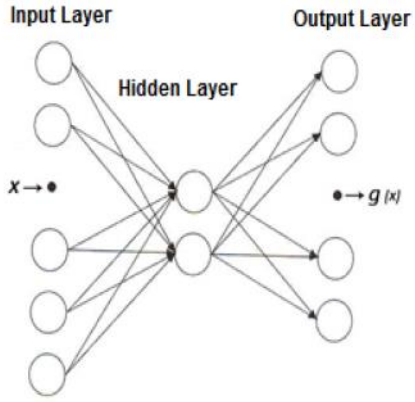

Figure 1

Figure 1 about here

where *p*, *l*, *J* are the units of input layer, hidden layer and output layer, respectively. The symbols *h* and *o* in order refer to hidden and output layers $\omega_{1r}^h, \dots, \omega_{pr}^h$ for $r \in \{1, \dots, l\}$ are the hidden layer weights and ω_{0r}^h for $r \in \{1, \dots, l\}$ are biases of this layer. $\omega_{1k}^o, \dots, \omega_{lk}^o$, for $k \in \{1,\dots,J\}$ are output layer weights and ω_{0k}^o for $k \in \{1, \dots, J\}$ biases of output layer. $\varphi(.)$ is a nonlinear transformation in hidden units and usually is a sigmoidal function such as logistic sigmoid function $\left(\varphi(x) = \frac{1}{1+\exp(-x)}\right)$. There are some reasons to use logistic function as $\varphi(.)$ (see Schalkoff, 1997 for more detail).

4. Softmax model as generalization of the logistic discrimination

As it was mentioned before the optimal bayes classification based on allocating with higher posterior probability can be used for logistic discrimination. To increase the ability of MLP models in classification the use of Softmax function in output units rather than identity function

was suggested. The main idea behind this model is to approximate the posterior probability of classes (Hastie et al., 2001, Lindemann et al., 2003). In the Softmax neural network model, the outputs of network are posterior probabilities of classes and have the form.

$$P(G_k|\mathrm{x}) = \frac{\exp\big(g_k(\mathrm{x})\big)}{\sum_{i=1}^{J}\exp\big(g_k(\mathrm{x})\big)} \qquad k \in \{1,\dots,J\} \tag{8}$$

where $g_k(\mathrm{x}),\ k \in \{1,\dots,J\}$ is defined in equation (7). The main idea of this model is to approximate the probability density function for the dependent variable. Notice that in the Softmax model, the outputs are positive and sum to 1. So using Softmax function in the units of output layer, we can approximate the posterior probabilities of classes in the neural network classification then select the most probable. The weights of the model are estimated alike other MLP model based on backpropagation algorithm which is explained in section 3. We can rewrite the (8) as

$$P(G_k|\mathrm{x}) = \frac{\exp\big(g_k(\mathrm{x})\big)}{1+\sum_{i=1}^{J-1}\exp\big(g_k(\mathrm{x})\big)}, \qquad k \in \{1,\dots,J-1\} \tag{9}$$

$$P\big(G_J|\mathrm{x}\big) = \frac{1}{1+\sum_{i=1}^{J-1}\exp\big(g_i(\mathrm{x})\big)}. \tag{10}$$

and we referred to these as Softmax model. Replacing (7) in (9) and (10) posterior probabilities obtain as bellow:

$$P(G_k|\mathrm{x}) = \frac{\exp[\omega_{0k}^{o}+\sum_{r=1}^{i}\omega_{rk}^{o}\varphi(\omega_{0r}^{h}+\sum_{i=1}^{p}\omega_{ir}^{h}\mathrm{x_i})]}{1+\sum_{j=1}^{J-1}\exp[\omega_{rj}^{o}\varphi(\omega_{0r}^{h}+\sum_{i=1}^{p}\omega_{ir}^{h}\mathrm{x_i})]} \tag{11}$$

$$P(G_k|\mathrm{x}) = \frac{1}{1+\sum_{j=1}^{J-1}\exp\big[\omega_{0k}^{o}+\sum_{r=1}^{l}\omega_{rj}^{o}\,\varphi\big(\omega_{0r}^{h}+\sum_{i=1}^{p}\omega_{ir}^{h}\mathrm{x_i}\big)\big]} \tag{12}$$

Probabilities in (11) and (12) can be interpreted as the generalization of logistic discriminations because they are generalization upon logistic discrimination with posterior probabilities, which are obtained in the equations (3) and (4). As that can be seen, the difference between equations (11) and (12) with equations (3) and (4) is in the power of exponential function. In following, we show that classification based on equation (7) is coincided on the classification with Softmax model because each class with the largest value in equation (7) has the higher posterior probability in the Softmax model. If it's supposed that $g(o_k)$ and $g(o_l)$ are network's outputs in equation (7), and if

$$g(o_k) > g(o_l), \quad k,l \in \{1,\dots,J\}.$$

then with monotonic property of exponential function we have:

$$exp\big(g(o_k)\big) \geq exp\big(g(o_l)\big)$$

and

$$\frac{\exp(g(o_k))}{1+\sum_{j=1}^{J-1}\exp(g(o_j))} \geq \frac{\exp\big(g(o_l)\big)}{1+\sum_{j=1}^{J-1}\exp\Big(g\big(o_j\big)\Big)} \Rightarrow P(G_k|\mathrm{x}) \geq P(G_l|\mathrm{x})$$

Notice that the essential assumption in logistic discrimination is:

$$\ln\left[\frac{p(\mathrm{x}|G_k)}{p\big(\mathrm{x}\big|G_J\big)}\right] = \omega_{0k} + \sum_{i=1}^{p}\omega_{ik}\mathrm{x}_i\,, \qquad k \in \{1,\dots,J-1\}$$

and $\omega_{0k} + \sum_{i=1}^{p}\omega_{ik}\mathrm{x}_i = 0$ is the separated hyperplane for two classes, but in many cases, it may be necessary that initially do some transformations on the observation because in such situation two classes will separate better with a hyperplane. For example, consider the two classes' case in figure (2); the linear discriminate function didn't separate two classes, even if two classes were separable.

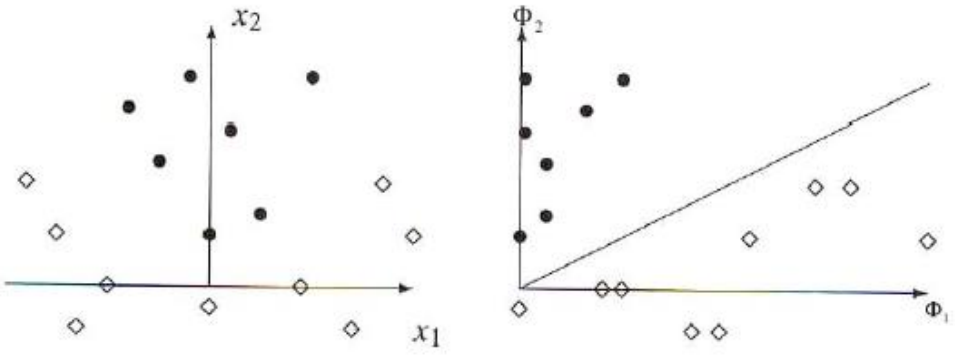

Figure 2

Figure 2 about here

But using the transformations

$$\varphi_1(x) = x_1^2,\ \ \varphi_2(x) = x_2$$

two classes leads to be separated in the φ space with a straight line (Webb, 2002). Notice that rewritten the logistic discrimination as:

$$\ln\left[\frac{p(\mathrm{x}|G_k)}{p\big(\mathrm{x}\big|G_J\big)}\right] = \omega_{0k}^{o} + \sum_{r=1}^{l}\omega_{rk}^{o}\varphi(\omega_{0r}^{h} + \sum_{i=1}^{p}\omega_{ir}^{h}\mathbf{x}_i), \qquad k \in \{1,\dots,J-1\} \tag{13}$$

then the right-hand side in the above equation shows that the observations are transferred to φ coordinates system initially then in this system the hyperplane $\omega_{0k}^{o} + \sum_{r=1}^{l}\omega_{rk}^{o}\varphi\big(\omega_{0r}^{h} +$

$\sum_{i=1}^{p} \omega_{ir}^{h}\mathbf{x}_i\big) = 0$ is used as separated boundary; and the right-hand side in the equation (13) is an output of *MLP* model with one hidden layer, *l* hidden units and identity function in output layer units. If we obtain the posterior probabilities from equation (13) the equations (9) and (10) will be obtained; so it can be said that Softmax model is a generalization of logistic discrimination and Softmax model without any hidden layer is the same of logistic discrimination. If the decision boundaries obtain from equation (9) and (10) so:

1) For class *k* and *J* where *k*

$$P(G_k|x) = P\big(G_J\big|x\big)$$

$$\Longrightarrow \frac{\exp[\omega_{0k}^{o} + \sum_{r=1}^{l} \omega_{rk}^{o}\varphi(\omega_{0r}^{h} + \sum_{i=1}^{p} \omega_{ir}^{h}\mathbf{x}_i)]}{1 + \sum_{j=1}^{J-1} \exp[\omega_{0k}^{o} + \sum_{r=1}^{l} \omega_{rk}^{o}\varphi(\omega_{0r}^{h} + \sum_{i=1}^{p} \omega_{ir}^{h}\mathbf{x}_i)]}$$

$$= \frac{1}{1 + \sum_{j=1}^{J-1} \exp[\omega_{0k}^{o} + \sum_{r=1}^{l} \omega_{rk}^{o}\varphi(\omega_{0r}^{h} + \sum_{i=1}^{p} \omega_{ir}^{h}\mathbf{x}_i)]}$$

$$\Longrightarrow \omega_{0k}^{o} + \sum_{r=1}^{l} \varphi(\omega_{0r}^{h} + \sum_{i=1}^{p} \omega_{ir}^{h}\mathbf{x}_i)] = 0 \tag{14}$$

2) For class *m* and *n* where $m \neq n < J$ have:

$$P(G_m|x) = P(G_n|x)$$

$$\Longrightarrow \frac{\exp[\omega_{0m}^{o} + \sum_{r=1}^{l} \omega_{rm}^{o}\varphi(\omega_{0r}^{h} + \sum_{i=1}^{p} \omega_{ir}^{h}\mathbf{x}_i)]}{1 + \sum_{j=1}^{J-1} \exp[\omega_{0j}^{o} + \sum_{r=1}^{l} \omega_{rj}^{o}\varphi(\omega_{0r}^{h} + \sum_{i=1}^{p} \omega_{ir}^{h}\mathbf{x}_i)]}$$

$$= \frac{1 + \exp[\omega_{0n}^{o} + \sum_{r=1}^{l} \omega_{rn}^{o}\varphi(\omega_{0r}^{h} + \sum_{i=1}^{p} \omega_{ir}^{h}\mathbf{x}_i)]}{1 + \sum_{j=1}^{J-1} \exp[\omega_{0j}^{o} + \sum_{r=1}^{l} \omega_{rj}^{o}\varphi(\omega_{0r}^{h} + \sum_{i=1}^{p} \omega_{ir}^{h}\mathbf{x}_i)]}$$

$$\Longrightarrow (\omega_{0m}^{o} - \omega_{0n}^{o}) + \sum_{r=1}^{l} (\omega_{rm}^{o} - \omega_{rn}^{o})\varphi(\omega_{0r}^{h} + \sum_{i=1}^{p} \omega_{ir}^{h}\mathbf{x}_i) = 0 \tag{15}$$

If the equations (14) and (15) are compared with equations (5) and (6) it can be seen that the decision boundaries obtained in the above equations are generalized linear or in the other words nonlinear; because the observations are transferred to the new coordinate system φ initially, and then a hyperplane is used to separate every two classes in this system. It is obvious that the boundaries obtained in equation (14) and (15) have more flexibility with regard to a linear boundary in equations (5) and (6). Furthermore, Softmax model has some advantages. The first property is that Softmax model has the same discrimination power as ANN model. The second one is that the Softmax model can detect complex and nonlinear relations between dependent and independent variables alike ANN models. The third one leads the Softmax model and ANN models to have the same prediction. The latter property which is not the case for ANN models is that the Softmax model can calculate the posterior probabilities of the classes as logistic discrimination. This property allows the Softmax model to use Bayesian

optimum rule for classification. However, there are some disadvantages in using the Softmax model. First, in appose of logistic discrimination, because of the final version of Softmax model is a very complicated function of independent variables; it works similar to a black box model and therefore, the coefficients of independent variables cannot be easily interpretated. Secondly, because high complexity to the Softmax model, the model suffers from overfitting. For explanation, the Softmax model has many parameters so it may follow the noise in the training data set due to overparameterization which leading to over fitting and so poor generalization for untrained data (Subasi and Ercelebi, 2005). Generally, ANN models have too many parameters and will overfit the data at a global minimum. There are two main strategies to prevent overfitting. In some developments of ANN models, an early stopping rule was used to avoid overfitting. It means that the model is trained only for a while, and stop before approaching the global minimum. However, this has the effect of shrinking the final model toward a linear model (see Hastie et al., 2001). A more explicit method to prevent overfitting is weight decay, which adds a penalty to the error function then the optimization algorithm is used. The penalty term takes care of the weight size in a way that it prefers smaller weights over bigger weights (Hastie et al., 2001, Lindeman et al., 2003). We show how adding a penalty to error function not solved the overfitting in Softmax model.

5. Determination the state of Thyroid Gland using discrimination analysis

This section aims to compare the traditional method of logistic discrimination to the more advanced Softmax technique as the statistical tool for developing classifiers for the diagnosis of thyroid gland state. The data show the state of the thyroid gland; generally, the secretions of the thyroid gland have three states, normal, low (hypothyroid), and up (hyperthyroid). The abnormal secretion of the thyroid gland (low or up) is the cause of many illnesses. This example has three independent variables:

x_1: Three Iodo Thyronin

x_2: Thyroxine

x_3: Thyrotropin

In this research, the data are collected from the 225 cases of Ahvaz University Jahad laboratory and the three factor x_1, x_2 and x_3 together with the secretion of the thyroid gland is measured; it has been discovered that 105 cases have normal thyroid, 72 cases have hyperthyroid and 48 cases have hypothyroid. The original sample is partitioned into two subsamples. One of

Table 1: The misclassification rate for the two models

Method	Training	test
Logistic discrimination	0.026	0.026
Softmax model	0.020	0.026

them is used as training data (150 cases) and another subsample is used as a test sample for testing the models (75 cases). We represented three different class with G_1, G_2 and G_3 in order for hypothyroid, normal thyroid and hyperthyroid, G_1 is used as basis class and Likelihood function of other classes is modeled toward G_1. Logistic discrimination and Softmax models were developed using 150 cases, and the test set was used for model validation; for Softmax model in optimal situations, three units are determined for hidden layer by trial and error. The maximum likelihood estimation (MLE) method is used to estimate the parameters in the logistic discrimination, and the Softmax model was trained based on a back propagation algorithm for optimization error function with weight decay. Readers can refer to the Hastie et al. (2001) and Webb, (2002) for more details. We followed Ripley (2004) who recommended if input data is in the rang of [0,1] an appropriate value of weight decay can be used in [10^{-4}, 10^{-1}] using the trial-and-error method we obtained 10^{-3} is an suitable value with minimum error. The misclassification rate is calculated in the training and test samples. The following table shows the misclassification rate for the two models: It can be seen that the Softmax model has better results in training sample but in the test sample the two models have same performance. It is obvious that Softmax model suffered from overfitting in this example. So according to a lot of parameters and high computational requirements in training Softmax model in comparison to the logistic discrimination it is obvious that using of Softmax model is not profitable in this example.

6. Conclusion

In this paper, we have shown that Softmax model can be considered as the generalization of logistic discrimination. It has nonlinear boundary decisions with regard to linear boundary in logistic discrimination. The Softmax model have some advantages and disadvantages witch are mentioned in section 4.

The main advantage of the Softmax model is that it can approximate posterior probabilities of classes, but the main drawback of Softmax model is that it suffers seriously from the curse of overfitting, because in overfitting situation the model has good performance on training data but it has poor performance on untrained data. Using the weight decay method to prevent overfitting is not effective for Softmax model in our data. Therefore, strategies preventing overfitting in the Softmax models should be investigated in order to use the advantages of Softmax model in the classification.

References

[1] Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**(1), 1-10.

[2] Al-Doud, E. (2009). A comparison between three neural network models for classification problems. *Journal of Artificial Intelligence* **2**(2), 56-64

[3] Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika*, **59**(1), 9-35.

[4] Anderson, J.A. (1975). Diagnostic by logistic discriminant function: further practical problems and results. *Applied Statistics*, **23**(3), 397-404.

[5] Anderson, J.A. and Richardson, S.C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, **21**(1), 71-78.

[6] Cox, D. R. (1966). Some procedures associated whit the logistic qualitative response curve in research papers in statistic: Festschrift for J. Neyman, Ed. David, F. N. New York: Wiley, 55-70.

[7] Cox, T. F. and Ferry, G. (1991). Robust logistic discrimination. *Biometrika*, **78**(4), 841-849.

[8] Day, N.E. and Kerridge, D.F. (1967). A general maximum likelihood discriminant. *Biometrics*, **23**, 313-323

[9] Eteban M.E., Sanz, S.G., Lopez, F.G., Borque, A. and Vergara J.M. (2006). Logistic regression versus neural networks for medical data. Monografias del Seminario Matematico Garciade Galdeano **33**, 245-252.

[10] Hastie, T., Tibshirani, R. and Freidman, J. H. (2001). The elementary of statistical learning: prediction, Inference and Data-Mining. Springer-Verlag.

[11] Lindemann, A., Dunis, C. and Lisboa, P. (2003). Level estimation, classification and probability distribution architectures for trading the EUR/USD exchange rate. *Neural Computing and Applications*, **14**(3), 256-271.

[12] Pearce, K. F. (1996). *Robust logistic discrimination*. Ph.D thesis, University of Newcastle upon Tyne.

[13] Raghavendra B.K. and Srivatsa S.K. (2011). Evaluation of logistic regression and neural network model with sensitivity analysis on medical datasets. *International Journal of Computer Science and Security* (IJCSS), **5**(5), 503-511

[14] Rao, K.S. (2011). Role of neural network models for developing speech systems. *Sadhana* 36(5), 783-836.

[15] Razi, M.A. and Athappilly, K. (2005). A comparative predictive analysis of neural networks(NNs), nonlinear regression and classification and regression tree(CART) models. *Expert Systems with Applications*, **29**, 65-74.

[16] Ripley, B.D. (2004), *Pattern Recognition and Neural networks*. 7th edition. Cambridge University press, Cambridge.

[17] Schalkoff, R.J. (1997), *Artificial Neural Networks*. New York: McGraw- Hill.

[18] Subasi, A. and Ercelebi, E. (2005). Classification of EEG signals using neural network and logistic regression. *Computer Methods and Programs in Biomedicine*, **78**, 87-99

[19] Webb, A.R. (2002). *Statistical Pattern Recognition*. Second edition. New York: Wiley.

Received March 3, 2013; accepted June 7, 2014.

F. Mohammadi Basatini
Department of Statistics
Shoushtar Branch, Islamic Azad University
Shoustar, Iran.
fe_mohamadi2011@yahoo.com

Rahim Chinipardaz
Department of Statistics
Shahid Chamran University
Ahvaz, Iran.
chinipardaz_r@scu.ac.ir