MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Short Communication

Open Access

CrossMark

Neural networks for classification and regression

Abstract

Neural Networks are well known techniques for classification problems. They can also be applied to regression problems. For this, the R software packages neuralnet and RSNNS were utilized. Their application was tested with Fisher's iris dataset and a dataset from Draper and Smith and the results obtained from these models were studied.

Keywords: neural network, classification, regression, neuralnet, RSNNS, confusion matrix

Volume 2 Issue 6 - 2015

Krishna Devulapalli
Statistical Consultant, India

Correspondence: Krishna Devulapalli; Statistical Consultant, India

Received: August 06, 2015 | **Published:** September 21, 2015

Abbreviations: SNNS; stuttgart neural network simulator, MLP; multi layer perceptron, RMS; root mean square

Introduction

Neural networks are generally utilized for classification problems, in which we will train the network to classify observations into two or more classes. For eg: A network can be trained with observed data of Fisher's iris flowers data and later it can be utilized for classifying the data into one of the three classes viz, setosa, versicolor and virginica. In a regression problem, the dependent variable is a continuous variable and independent variables can be continuous or categorical variables. Neural networks can also be trained to regression problems, so that they can be utilized latter for prediction purpose. In this study, the application of Neural Networks for both classification and regression problems is studied and the results so obtained are discussed.

The R statistical software is well known open source statistical software and it is heavily utilized in many applications including data analytics, big data, data mining, text mining etc. In this study, for carrying out the classification analysis, the R package neuralnet is utilized. For complete details of utilizing this package, please refer to the CRAN reference document.[1] In classification problems, generally the Confusion Matrix is utilized to find out the percentage correct classification.

In order to do regression analysis, the RSNNS (Neural Networks in R using the Stuttgart Neural Network Simulator i.e. SNNS) package, which uses Stuttgart neural network simulator and is utilized.[2] In the case of regression problems, because the neural networks operate in terms of 0 to 1 or -1 to 1, we have to first transform the data into one of these scales. In this study, I have chosen the scale 0 to 1. The mlp (multi layer perceptron) function of RSNNS package is utilized for training the network. After the network is trained, the outputs will be available in the object fitted values. To get these values to original units, these values are converted to the original scale. One can compare the transformed fitted values with actual values, find the residuals and calculate the RMS (Root mean Square) error to find out the accuracy of the fit.

Results

Fisher's iris flower data is a multivariate dataset consisting of 50 samples for each of the species of iris viz., setosa, virginica and versicolor. For each of these samples four measurements were made viz., sepal length, sepal width, petal length and petal width. Utilizing this data, Fisher has developed Linear Discriminant analysis to classify these samples for the species.

For the classification analysis, the Fisher's iris data, which is available in R software was utilized. The iris data consists of 150 observations consisting of four variables viz., sepal length, sepal width, peal length and petal width for three species setosa, versicolor and virginica. The R software neuralnet package was utilized for training the network with the configuration of (4/3/3 i.e. 4 nodes in the input layer, 3 nodes in the hidden layer and 3 nodes in the output layer) as shown in Figure 1. In order to cross validate the classification results, cross validation is done by iterating the model fitting 10 times, each time with 75 randomly selected observations in the training dataset and the remaining 75 observations in the test dataset. The results of these ten iterations are presented in Table 1. In this table, column 1 represents iteration number, columns 2 and 3 represent the number of observations correctly classified in that iterations and the corresponding percentage correct classification. Similarly columns 4 and 5 of this table represent the corresponding figures for the remaining 75 observations of the test dataset. As expected, in general, the percentage correct classification is slightly better in the training dataset than in the test dataset for each iteration. Overall, it is observed that the average of ten iterations of the training dataset yielded 99.20% accurate classification results while the test dataset has yielded 95.73% correct classification as seen in the last row of this table.

For carrying out regression analysis, the RSNNS model of R software was utilized. The dataset dsa01a from the CRAN R package aprean3,[3] which contains all the datasets of "Applied Regression Analysis" by Draper & Smith[4] was utilized for the regression analysis. This dataset relates to the pounds of steam used monthly in a large industrial concern for 25 observations. The dataset consists of one dependent variable Y (available in the available in the variable x1 of the dataset), which is a pound of steam used monthly and 10 other independent variables. For the regression analysis, the dependent variable and two independent variables were considered viz., x6 – no of operating days per month and x8 – average monthly atmospheric pressure.

As per the regression equation fitted by Draper & Smith[4] (pages 154-159), the regression model is fitted by taking x1 as dependent variable and x6, x8 as independent variables. For the RSNNS regression, the data is initially transformed into a scale of 0 to 1 as mentioned above. Later to apply RSNNS Neural Network, the network was trained with mlp function using the parameters Std_Backpropagation, maximum 1000 iterations and the metric RSME. The network diagrams generated by the model along with various residual plots are provided in Figure 2. Further the Root Mean Square is calculated by utilizing the actual values and the fitted values from the network. The Root Mean Square error was found to be 0.1548, suggesting that neural network performance is very good for regression problem also.

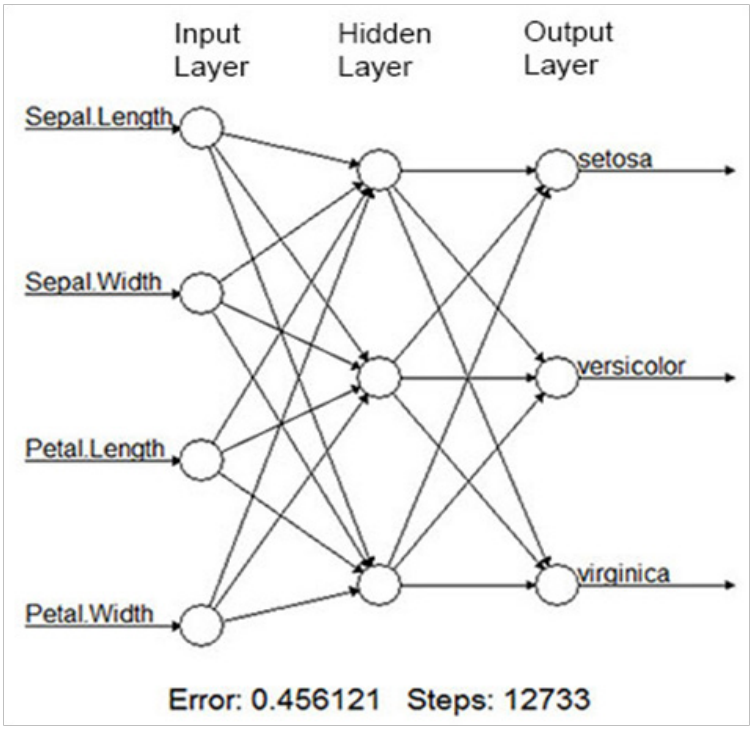

Figure 1 Neural Network diagram of the Fisher's iris data. Input Layer contains four nodes corresponding to the four input variables viz., sepal length, sepal width, petal length and petal width. Hidden layer contains 3 nodes and output layer contains 3 nodes corresponding to the outputs viz., setosa, versicolor and virginica.

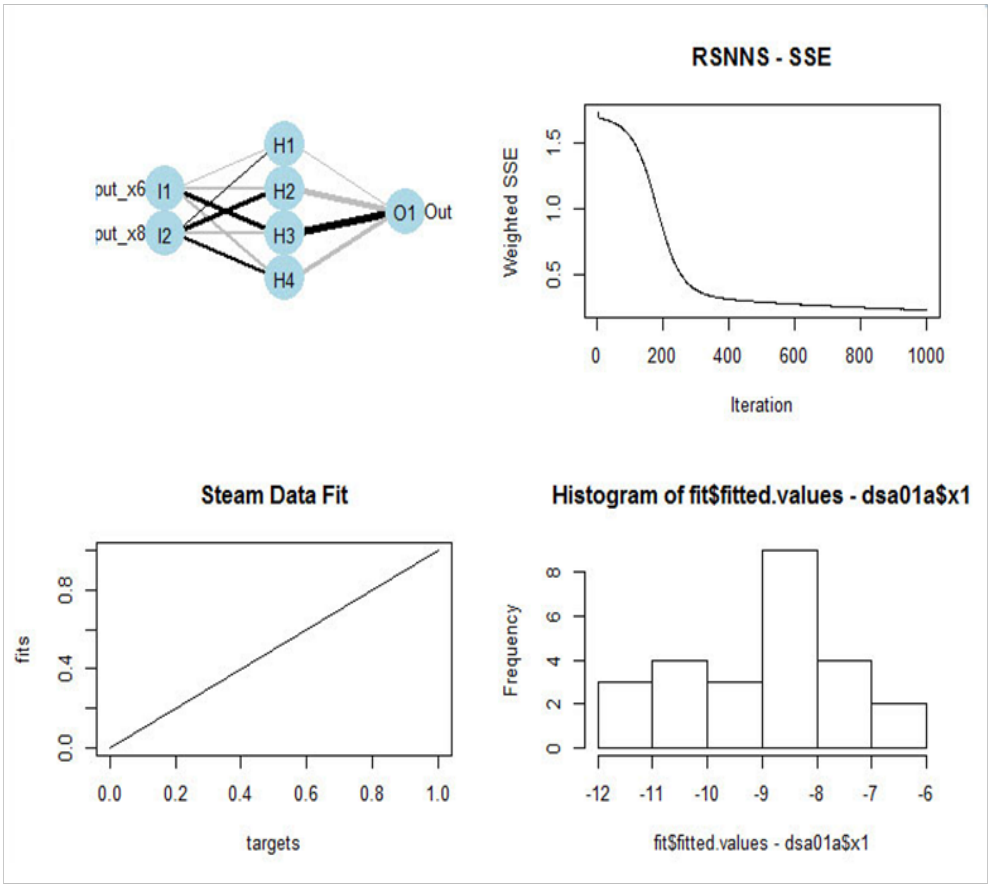

Figure 2 Neural network model results along with some residual plots for the dsa01a dataset.

Citation: Devulapalli K. Neural networks for classification and regression. *Biom Biostat Int J.* 2015;2(6):183–185.
DOI: 10.15406/bbij.2015.02.00046

Table 1 Cross validation Results of iris data for 10 iterations, each iteration consisting of 75 random observations in Training dataset and remaining 75 observations in Test dataset

	Training dataset		Test dataset	
Iteration	**No. of observations correctly classified (out of 75)**	**Percentage correct classification**	**No. of observations correctly classified (out of 75)**	**Percentage correct classification**
1	74	98.66	70	93.33
2	74	98.66	73	97.33
3	75	100.00	70	93.33
4	75	100.00	72	96.00
5	75	100.00	71	94.66
6	75	100.00	70	93.33
7	74	98.66	74	98.66
8	75	100.00	71	94.66
9	74	98.66	72	96.00
10	73	97.33	75	100.00
	Mean	99.20	Mean	95.73

Conclusion

An attempt has been made in this study to apply neural networks for both classification and regression problems, For carrying out the statistical analysis, the R Statistical package was utilized. For classification problem, the neuralnet package was used and for regression analysis, the RSNNS package was used. For the classification analysis, the Fisher's iris data was utilized and for regression analysis, the dataset of Draper & Smith[4] was used. In the case of both classification and regression analysis problems, it was observe that the neural networks have yielded good classification and regression results, suggesting their utility for these applications. However, there is a need to study them in comparison with other classical models and also with much bigger datasets and assess their performance.

Acknowledgement

None.

Conflict of Interest

None.

References

1. Stefan F, Frauke G. *CRAN package "neuralnet"*. 2015;1–13 p.
2. Christoph B, Benítez JM. *CRAN package "RSNNS"*. 2015;1–71 p.
3. Braglia L. *CRAN package "aprean3"*. Datasets from Draper and Smith "Applied Regression Analysis". 2015.
4. Draper NR, Smith H. *Applied Regression Analysis, 3rd ed*. 1998;154–159 p.

Citation: Devulapalli K. Neural networks for classification and regression. *Biom Biostat Int J.* 2015;2(6):183–185.
DOI: 10.15406/bbij.2015.02.00046