Introduction to Data Science Tools & Techniques

## Group Assignment

# Implementation of Logistic Regression & Artificial Neural Network

### Group Members:

Ali Nauman – 18L1863  
Hammad Akram – 18L1808  
Ammar Hussain – 18L1834

## Introduction:

In this assignment we have implemented *Logistic Regression* and *Artificial Neural Network* on two separate data-sets. Both the implementations have been conducted in *R Programming Language*. Furthermore, several preprocessing techniques have been utilized to improve the accuracy results. We have implemented feature selection technique to select the appropriate features, which affect the analysis in the best possible way.

## Dataset Information

### Wisconsin Breast Cancer Database

The objective is to identify each of a number of benign or malignant classes. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Ten real-valued features are computed for each cell nucleus. All feature values are recoded with four significant digits.

### Johns Hopkins University Ionosphere database

This radar data was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The targets are free electrons in the ionosphere. *Good* radar returns are those showing evidence of some type of structure in the ionosphere. *Bad* returns are those that do not; their signals pass through the ionosphere. A data frame with 351 observations on 35 independent variables, some numerical and 2 nominal, and one last defining the class.

## Algorithms

### Logistic Regression

Logistic regression is a predictive modelling algorithm that is used when the Y variable is binary categorical. That is, it can take only two values like 1 or 0. The goal is to determine a mathematical equation that can be used to predict the probability of event 1. Once the equation is established, it can be used to predict the Y when only the X’s are known.

In linear regression the Y variable is always a continuous variable. If suppose, the Y variable was categorical, you cannot use linear regression model it.

So what would you do when the Y is a categorical variable with 2 classes?

Logistic regression can be used to model and solve such problems, also called as binary classification problems.

A key point to note here is that Y can have 2 classes only and not more than that. If Y has more than 2 classes, it would become a multi class classification and you can no longer use the vanilla logistic regression for that.

Yet, Logistic regression is a classic predictive modelling technique and still remains a popular choice for modelling binary categorical variables.

Another advantage of logistic regression is that it computes a prediction probability score of an event.

### Artificial Neural Network

A neural network is a computational paradigm inspired by the parallelism of the brain. There are certain things that humans do well (e.g. vision and language), while there are certain things which present day computers do well (e.g. numeric computations).

The idea of ANNs is based on the belief that working of human brain by making the right connections, can be imitated using silicon and wires as living neurons and dendrites.

ANNs are composed of multiple nodes, which imitate biological neurons of human brain. The neurons are connected by links and they interact with each other. The nodes can take input data and perform simple operations on the data. The result of these operations is passed to other neurons. The output at each node is called its activation or node value.

Each link is associated with weight. ANNs are capable of learning, which takes place by altering weight values.

There are multiple types of neural network, each of which come with their own specific use cases and levels of complexity. The most basic type of neural net is something called a feed-forward neural network, in which information travels in only one direction from input to output.

A more widely used type of network is the recurrent neural network, in which data can flow in multiple directions. These neural networks possess greater learning abilities and are widely employed for more complex tasks such as learning handwriting or language recognition.

Here comes a question, “what tasks can’t a neural network do?” From making cars drive autonomously on the roads, to generating shockingly realistic CGI faces, to machine translation, to fraud detection, to reading our minds, to recognizing when a cat is in the garden and turning on the sprinklers; neural nets are behind many of the biggest advances in A.I.

## Implementation

### Logistic Regression - Wisconsin Breast Cancer Database

The goal here is to model and predict if a given specimen (row in dataset) is benign or malignant, based on 9 other cell features. The dataset has 699 observations and 11 columns. The Class column is the response (dependent) variable and it tells if a given tissue is malignant or benign.

We have converted a factor to a numeric variable. All the columns are converted to numeric. The classes 'benign' and 'malignant' are split approximately in 1:2 ratio. Clearly there is a class imbalance. So, before building the *logit model*, we need to build the samples such that both the 1's and 0's are in approximately equal proportions. We can handle them either down-sampling or up-sampling.



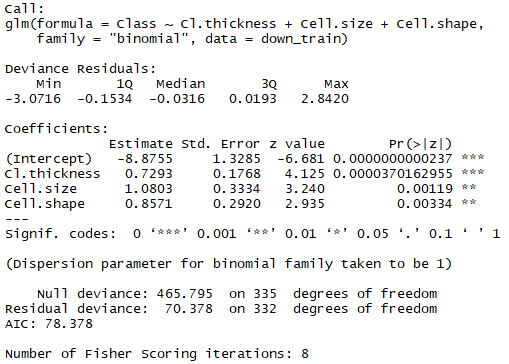
After Down-Sampling, the ratio is adjusted to the following,



After Up-Sampling, the ratio is adjusted to the following,



We have built the *logistic regression* function using the *glm* function.



### After implementing the logistic model on the full features of this dataset, we also extracted top features using stepwise variable selection.

### 

Using a full features logistic regression gives us an accuracy of *96.07%*. Whereas, the accuracy increased rapidly as we performed the logistic regression with just a few selected top features to *96.56%*.

### Logistic Regression - Johns Hopkins University Ionosphere database

The targets are free electrons in the ionosphere. Good radar returns are those showing evidence of some type of structure in the ionosphere. Bad returns are those that do not; their signals pass through the ionosphere.

We have converted a factor to a numeric variable. All the columns are converted to numeric.



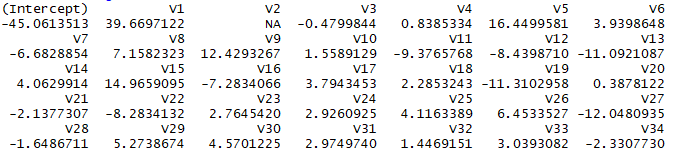
After Down-Sampling, the ratio is adjusted to the following,



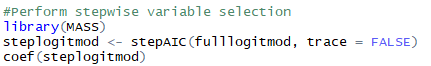
After Up-Sampling, the ratio is adjusted to the following,

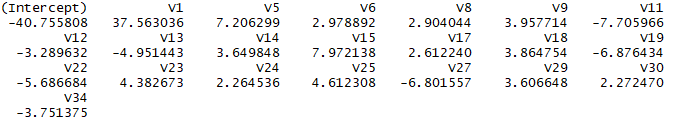


After implementing the logistic model on the full features of this dataset, we can see the following coefficients for this,



We also extracted top features using stepwise variable selection,





Using a full features logistic regression gives us an accuracy of *85.58%*. Whereas, the accuracy increased rapidly as we performed the logistic regression with just a few selected top features to *91.46%*.

### ANN - Wisconsin Breast Cancer Database

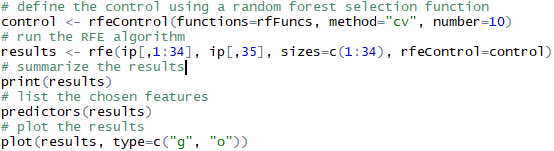
The goal here is to model and predict if a given specimen (row in dataset) is benign or malignant, based on 9 other cell features. The dataset has 699 observations and 11 columns. The Class column is the response (dependent) variable and it tells if a given tissue is malignant or benign.

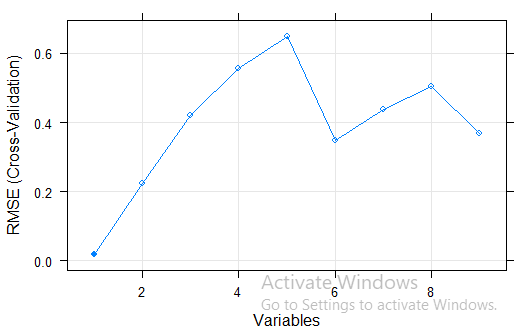
We have converted a factor to a numeric variable. All the columns are converted to numeric.

We have to find the best features available for creating an ANN. Automatic feature selection methods can be used to build many models with different subsets of a dataset and identify those attributes that are and are not required to build an accurate model.

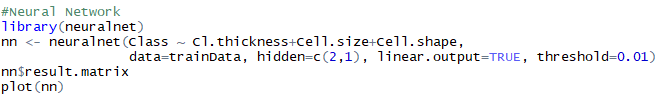
A popular automatic method for feature selection provided by the caret R package is called *Recursive Feature Elimination or RFE.* A Random Forest algorithm is used on each iteration to evaluate the model. The algorithm is configured to explore all possible subsets of the attributes.

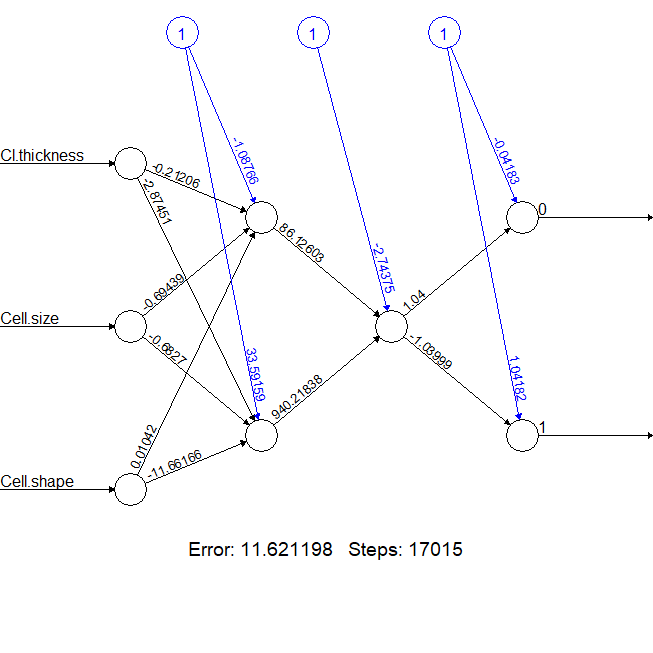
Below is the code for extracting the best features which in this case are, Cl.thickness, Cell.size, and Cell.shape.





The ANN is implemented using the following code and the three best features extracted earlier,





The confusion matrix as defined as following for the results,



### ANN - Johns Hopkins University Ionosphere database

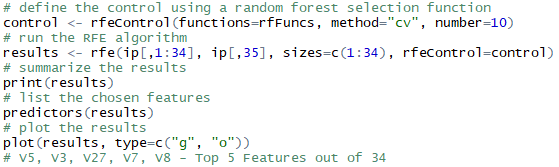
The targets are free electrons in the ionosphere. Good radar returns are those showing evidence of some type of structure in the ionosphere. Bad returns are those that do not; their signals pass through the ionosphere.

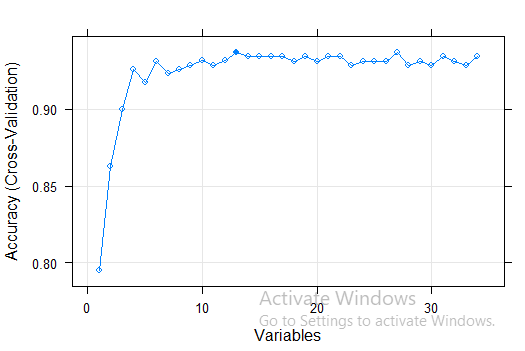
We have converted a factor to a numeric variable. All the columns are converted to numeric.

We have to find the best features available for creating an ANN. Automatic feature selection methods can be used to build many models with different subsets of a dataset and identify those attributes that are and are not required to build an accurate model.

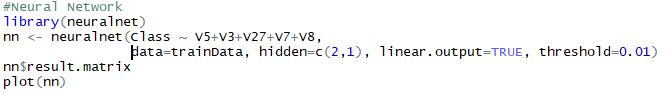
A popular automatic method for feature selection provided by the caret R package is called *Recursive Feature Elimination or RFE.* A Random Forest algorithm is used on each iteration to evaluate the model. The algorithm is configured to explore all possible subsets of the attributes

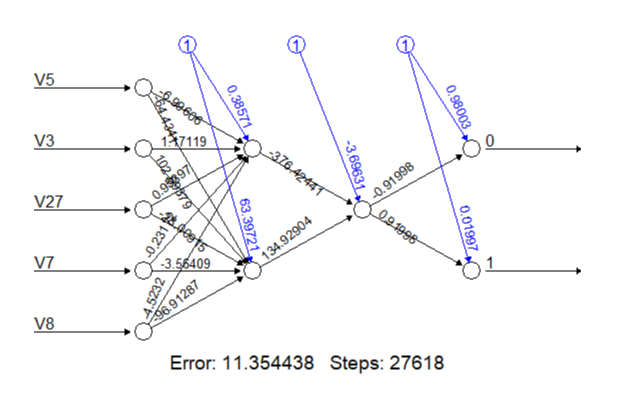
Below is the code for extracting the best features, V5, V3, V27, V7, V8 - Top 5 Features out of 34,





The ANN is implemented using the following code and the five best features extracted earlier,





The confusion matrix as defined as following for the results,



## Results

|  |  |  |
| --- | --- | --- |
|  | **Logistic Regression** | **Artificial Neural Network** |
| **Breast Cancer Dataset** | 96.56% | 94.11% |
| **Ionosphere Database** | 91.46% | 90.38% |

## Analysis

The results obtained by both the logistic regression and artificial neural networks for the selected datasets are quite similar but the logistic regression takes the center stage.

Though the idea and approach is completely different in both cases. Given that the logistic regression works on purpose only for a 2-class dataset, this restricted our scope of datasets to just 2-class datasets. Furthermore, as the number of variables have increased, using all of them would be a chaos. Rather, we have implemented Step-wise variable selection.

The stepwise regression (or stepwise selection) consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model that is a model that lowers prediction error.

This way we can see an increase in accuracy in both the datasets, as it reduces the prediction error and selects the best possible features. It has the following three strategies in place, *Forward selection, backward selection and Stepwise selection*.

In ANN, to extract the best possible features we have implemented a *Recursive Feature Elimination* algorithm. The algorithm is used on each iteration to evaluate the model. The algorithm is configured to explore all possible subsets of the attributes. The central premise when using a feature selection technique is that the data contains some features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information. Redundant and irrelevant are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated.

Overall, both the methods are extremely useful and the results obtained are quite strong in predicting the accuracy.