Assignment 2-Data Mining and Machine Learning

Sai Krishna Lakshminarayanan (18230229)

MSc Computer Science (Data Analytics)

2 November 2018

Task -1

- In the first assignment, data has been taken and pre processing has been done along with 10 fold cross validation.
- Now, this data is taken and is split into two parts as train data and test data randomly in the parts of 2/3 and 1/3 respectively
- From the previous assignment, the packages of e1071 and RWeka are used again here to perform naive bayes classification and C4.5 respectively.
- The naive bayes classification model is generated by training the algorithm with the training data set values for Autoimmune Disease.
- It is then used to predict the values in test data which is 1/3rd of the data set and a confusion matrix is obtained.
- The same process is repeatead for the C4.5 algorithm by generating a model with the training data set and using it to predict the values that are to be in the test dataset.
- Now,inorder to generate ROC curve, the ROCR package is used. ROCR is used to generate 2 dimensional ROC curves by combining two performance measures. It is highly flexible and usually has three commands with default values provided for the optional ones.
- For Naive bayes, the ROC curve is obtained by taking the predicted scores and storing it in a variable predicted score and taking the actual scores of the autoimmune disease in the test data set and storing it in a variable actual score.In the curve, the x axis is false positive rate and for y axis is true positive rate
- Now, performance measure of this parameter is done using performance() and stored which is then plotted to give the desired ROC plot.
- Using this, the Area of the Curve is also determined and its value is given.
- The same procedure is followed for the C 4.5 and the ROC curve and the area is obtained for it also by taking the predicted scores and acutal scores and checking the performance measure using them.

Task -2

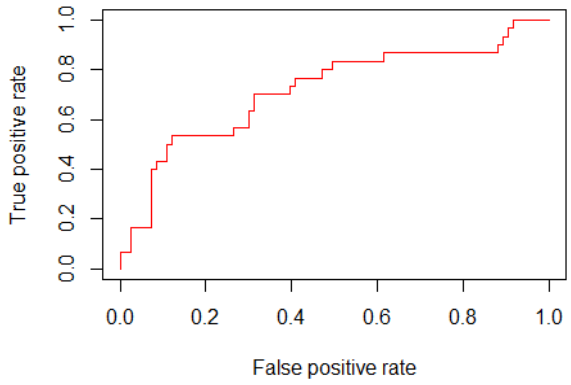
ROC Curve For Naive Bayes Algorithm

```
library(e1071)#package for naive bayes algorithm
Naive_Bayes_Model1=naiveBayes(Autoimmune_disease~., data=trainData)# generating a training model for the al
gorithm
NB_Predictions1<-predict(Naive_Bayes_Model1,testData)#predicting the values for the test data
print(table(NB_Predictions1,testData$Autoimmune_disease))#giving out the confusion matrix for the test case
```

```
##
## NB_Predictions1 negative positive
##       negative       68       14
##       positive       15       16

library(ROCR)# package for generating ROC curves

NB_Predictions1<-predict(Naive_Bayes_Model1,testData,type = 'raw')#converting to suitable type
predicted_score <- NB_Predictions1[, "positive"]#taking the predicted value forpositive ones
actual_score <- testData$Autoimmune_disease#taking the actual values from the test data
pred <- prediction(predicted_score, actual_score) # predicting the measures
naive_bayes_performance = performance(pred, "tpr", "fpr") #taking the performance measure for naive bayes
plot(naive_bayes_performance, col = "red")# generating the required ROC curve
```



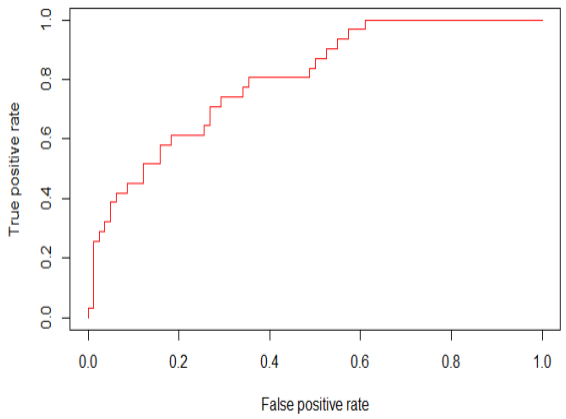

```
auc <- performance(pred,"auc") # obtaining the area value
print(auc@y.values[[1]]) #giving out the result

## [1] 0.7176707
```

ROC Curve for C 4.5 Algorithm

```
library(RWeka)#loading the package for performing C 4.5
C45 <- J48(Autoimmune_disease~., data=trainData) #training the model with the train data values
predictions2 <- predict(C45, testData) #using it to predict the test data value
print(table(predictions2, testData$Autoimmune_disease))#giving out the final result##
## predictions2 negative positive
##     negative       83       30
##     positive        0        0

a<- (predict(C45, testData,type='prob')) #converting it to a suitable type for execution
predicted_score1<-a[,"positive"] #obtaining the positive predicted values
actual_score1<-testData$Autoimmune_disease #obtaining the actual values in the test data
final<-prediction(predicted_score1,actual_score1)#checking the prediction that has been done by comparing the m
easures
C45_performance<-performance(final,"tpr","fpr")# evaluating the performance of the C 4.5 algorithm
plot(C45_performance, col = "red")#plotting the required ROC curve
```




```
auc <- performance(final,"auc") # Generating the Area of the curve
print(auc@y.values[[1]])#Giving out the result

## [1] 0.8013375
```

Task-3

- The ROC curve of C 4.5 algorithm is smoother than that of the Naïve Bayes one. This means that Naïve Bayes is effective only when a large data set is used. This can be proved with the evidence that the initial values in prediction in the naïve bayes is low and only grows gradually as the count of the data increases. This can be deemed as one of the disadvantages of Naïve Bayes classification. The curve becomes smoother as the dataset size is increased.
- The area of the ROC curve for C 4.5 is greater than Naïve Bayes classification. This is supported by the fact that the ROC curve of C 4.5 is present more on the left upper part than the Naïve Bayes. By law, we know that accuracy of an algorithm is higher as the area of the curve present in the left upper part increases. This also helps us to observe that the accuracy of C 4.5 is greater than Naïve Bayes for the given test data.
- It is observed that every point on the ROC curve is a cut off of the values between true positive, false positive and true negative ,false negative rates respectively. The steps in the curves are indicating the fact that continuity is missing from the confusion matrix that is generated to the original value. The curve has to be plotted by data pairs for predicted score and the actual score which are stored in the final which is the measure of prediction of them. Due to the nature of it being discrete, the steps are prevalent and hence indicating the need for a larger dataset to execute it efficiently. It can also be improved by changing the cut off points for the values which are to be considered as true positive or false positive thereby knowing the better tradeoff between them.

References

1.https://machinelearningmastery.com/non-linear-classification-in-r-with-decision-trees/

2.https://gist.github.com/duttashi/a51c71acb7388c535e30b57854598e77

3.https://cran.r-project.org/web/packages/RWeka/RWeka.pdf

4.https://hackerbits.com/data/c4-5-data-mining-algorithm/

5.http://blog.aylien.com/naive-bayes-for-dummies-a-simple-explanation/

6.https://stackoverflow.com/questions/30818188/roc-curve-in-r-using-rpart-package

7.http://information-gain.blogspot.com/

8.https://cran.r-project.org/web/packages/ROCR/ROCR.pdf

9. https://stats.stackexchange.com/questions/191805/r-plotting-a-roc-curve-for-a-naive-bayes-classifier-using-rocr-not-sure-if-i

10. https://www.quora.com/What-does-it-mean-when-an-ROC-curve-is-not-smooth

11. https://www.medcalc.org/manual/roc-curves.php

12. https://acutecaretesting.org/en/articles/roc-curves-what-are-they-and-how-are-they-used