Tutorial 8

DSA1101
Introduction to Data Science

October 26, 2018

Exercise 1. Receiver Operating Characteristic (ROC) Curve in R

In last week's lecture, we introduced the ROC curve which is a common tool to evaluate classifiers in terms of trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) when the classification threshold varies. We have also studied area under the curve (AUC), which is calculated by measuring the area under the ROC curve. Higher AUC scores mean the classifier performs better. In this week's tutorial, we will illustrate how AUC can be computed with the R package 'ROCR'.

Consider the dataset 'bank-sample.csv' we discussed in the lectures. For this exercise, we will predict the binary outcome `subscribed` using the naïve Bayes classifier, and generate the ROC curve of the naïve Bayes classifier built on a training set of 2,000 instances and tested on a testing set of 100 instances. All the datasets for this exercise are posted under the folder 'Tutorial 8' in IVLE.

(a) Load the training and testing bank sample datasets.

```
# training set
banktrain <- read.table("bank-sample.csv",header=TRUE,sep=",")
# drop a few columns
drops <- c("balance", "day", "campaign", "pdays", "previous", "month"
    )
banktrain <- banktrain [,!(names(banktrain) %in% drops)]
# testing set
banktest <- read.table("bank-sample-test.csv",header=TRUE,sep=",")
banktest <- banktest [,!(names(banktest) %in% drops)]
```

(b) Build the naïve Bayes classifier based on the training dataset, and perform prediction for the test dataset.

```
library(e1071)

# build the naive Bayes classifier
nb_model <- naiveBayes(subscribed~.,
data=banktrain)
# perform on the testing set
nb_prediction <- predict(nb_model,
# remove column "subscribed"
banktest[,-ncol(banktest)],
type='raw')
```

(c) Plot the ROC curve for the naïve Bayes classifier.

```
library(ROCR)

score <- nb_prediction[, c("yes")]

actual_class <- banktest$subscribed =='yes'
pred <- prediction(score, actual_class)

perf <- performance(pred, "tpr", "fpr")
plot(perf, lwd=2, xlab="False Positive Rate (FPR)",
ylab="True Positive Rate (TPR)")
abline(a=0, b=1, col="gray50", lty=3)
```

(d) Compute AUC for the naïve Bayes classifier.

```
auc <- performance(pred, "auc")
auc <- unlist(slot(auc, "y.values"))
auc
```