Tutorial 10

DSA1101
Introduction to Data Science

November 9, 2018

Exercise 1. Logistic regression in R
In tutorial 7, we looked at the CSV dataset "Titanic.csv" which provides information on the fate of passengers on the fatal maiden voyage of the ocean liner Titanic, and includes the variables economic status (class), sex, age and survival. We trained a naïve Bayes classifier using this dataset, and predict survival. This week, we will use logistic regression to predict survival, and compare the performances of the two classifiers visually using an ROC curve.

(a) Load the dataset "Titanic.csv" which has been posted under the folder for Tutorial 7.

```
Titanic_dataset= read.csv("Titanic.csv")
dim(Titanic_dataset)
head(Titanic_dataset)
```

(b) Perform logistic regression of 'Survived' on all the feature variables.

```
Survival_logistic <- glm (Survived~.,
data=Titanic_dataset,
family=binomial(link="logit"))
```

(c) Perform naïve Bayes classification of 'Survived' based on all the feature variables.

```
library(e1071)
Survival_Nbayes <- naiveBayes(Survived~.,
data=Titanic_dataset)
```

(d) Observe and compare the ROC curves for the two classifiers.

```
library(ROCR)
pred = predict(Survival_logistic, type="response")
predObj = prediction(pred, Titanic_dataset$Survived)
rocObj = performance(predObj, measure="tpr", x.measure="fpr")
plot(rocObj)


nb_prediction <- predict(Survival_Nbayes, Titanic_dataset,type='raw')
score <- nb_prediction[, 2]
pred_nb <- prediction(score, Titanic_dataset$Survived)
roc_nb = performance(pred_nb, measure="tpr", x.measure="fpr")
plot(roc_nb, add = TRUE, col = 2)


legend("bottomright", c("logisic regression","naive Bayes"),col=c("
    black","red"), lty=1)
```