ML Assignment 3 - Naive Bayes Classification

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Loading the Data file in R

```
FlightData<-read.csv("FlightDelays.csv")
str(FlightData)
```

```
## 'data.frame':    2201 obs. of  13 variables:
##  $ CRS_DEP_TIME : int  1455 1640 1245 1715 1039 840 1240 1645 1715 2120 ...
##  $ CARRIER      : chr  "OH" "DH" "DH" "DH" ...
##  $ DEP_TIME     : int  1455 1640 1245 1709 1035 839 1243 1644 1710 2129 ...
##  $ DEST         : chr  "JFK" "JFK" "LGA" "LGA" ...
##  $ DISTANCE     : int  184 213 229 229 229 228 228 228 228 228 ...
##  $ FL_DATE      : chr  "01/01/2004" "01/01/2004" "01/01/2004" "01/01/2004" ...
##  $ FL_NUM       : int  5935 6155 7208 7215 7792 7800 7806 7810 7812 7814 ...
##  $ ORIGIN       : chr  "BWI" "DCA" "IAD" "IAD" ...
##  $ Weather      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ DAY_WEEK     : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ DAY_OF_MONTH : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ TAIL_NUM     : chr  "N940CA" "N405FJ" "N695BR" "N662BR" ...
##  $ Flight.Status: chr  "ontime" "ontime" "ontime" "ontime" ...
```

```
head(FlightData)
```

```
##   CRS_DEP_TIME CARRIER DEP_TIME DEST DISTANCE    FL_DATE FL_NUM ORIGIN Weather
## 1         1455      OH     1455  JFK      184 01/01/2004   5935    BWI       0
## 2         1640      DH     1640  JFK      213 01/01/2004   6155    DCA       0
## 3         1245      DH     1245  LGA      229 01/01/2004   7208    IAD       0
## 4         1715      DH     1709  LGA      229 01/01/2004   7215    IAD       0
## 5         1039      DH     1035  LGA      229 01/01/2004   7792    IAD       0
## 6          840      DH      839  JFK      228 01/01/2004   7800    IAD       0
##   DAY_WEEK DAY_OF_MONTH TAIL_NUM Flight.Status
## 1        4            1   N940CA        ontime
## 2        4            1   N405FJ        ontime
## 3        4            1   N695BR        ontime
## 4        4            1   N662BR        ontime
## 5        4            1   N698BR        ontime
## 6        4            1   N687BR        ontime
```

```
#View(FlightData)
```

Library for Naive Bayes theorem

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(ISLR)
# install.packages("e1071") #install first
library(e1071)
```

Change the numerical variables to categorical

```
FlightData$DAY_WEEK<-factor(FlightData$DAY_WEEK)
FlightData$DEP_TIME<-factor(FlightData$DEP_TIME)

#creating hourly bins for the departure time
FlightData$CRS_DEP_TIME<-factor(round(FlightData$CRS_DEP_TIME/100))
```

Divide into training and test

```
set.seed(123)
Index_train<-createDataPartition(FlightData$Flight.Status, p=0.6, list = FALSE)
#Column data needed for our test
var.req<-c(1,2,4,8,10,13)
#Split the data
TrainData<-FlightData[Index_train,]
TestData<-FlightData[-Index_train,]
#Trimming of the unwanted columns from the dataframe.
Trg<-TrainData[,var.req]
Test<-TestData[,var.req]

#Data validations at the Training and Test data set
table(Trg$Flight.Status)
```

```
##
## delayed  ontime
##     257    1064
```

```
summary(Trg)
```

```
##   CRS_DEP_TIME   CARRIER              DEST              ORIGIN
##  15     :178   Length:1321        Length:1321        Length:1321
##  17     :139   Class :character   Class :character   Class :character
##  8      :104   Mode  :character   Mode  :character   Mode  :character
##  16     :103
##  21     : 85
##  12     : 84
##  (Other):628
##  DAY_WEEK Flight.Status
##  1:202    Length:1321
##  2:176    Class :character
##  3:172    Mode  :character
##  4:232
##  5:241
##  6:145
##  7:153
```

```
summary(Test)
```

```
##   CRS_DEP_TIME   CARRIER              DEST              ORIGIN
##  15     :114   Length:880         Length:880         Length:880
##  17     :102   Class :character   Class :character   Class :character
##  16     : 75   Mode  :character   Mode  :character   Mode  :character
##  8      : 60
##  12     : 58
##  6      : 56
##  (Other):415
##  DAY_WEEK Flight.Status
##  1:106    Length:880
##  2:131    Class :character
##  3:148    Mode  :character
##  4:140
##  5:150
##  6:105
##  7:100
```

Run Naive Bayes

```
nb_model<-naiveBayes(Trg$Flight.Status~., data=Trg)
nb_model
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##   delayed    ontime
## 0.1945496 0.8054504
##
## Conditional probabilities:
##          CRS_DEP_TIME
## Y                   6           7           8           9          10
##   delayed 0.011673152 0.062256809 0.042801556 0.019455253 0.031128405
##   ontime  0.062969925 0.061090226 0.087406015 0.060150376 0.054511278
##          CRS_DEP_TIME
## Y                  11          12          13          14          15
##   delayed 0.007782101 0.070038911 0.038910506 0.046692607 0.194552529
##   ontime  0.039473684 0.062030075 0.063909774 0.052631579 0.120300752
##          CRS_DEP_TIME
## Y                  16          17          18          19          20
##   delayed 0.081712062 0.143968872 0.042801556 0.097276265 0.019455253
##   ontime  0.077067669 0.095864662 0.037593985 0.041353383 0.025375940
##          CRS_DEP_TIME
## Y                  21
##   delayed 0.089494163
##   ontime  0.058270677
##
##          CARRIER
## Y                  CO          DH          DL          MQ          OH
```

```
##   delayed 0.066147860 0.322957198 0.112840467 0.178988327 0.007782101
##   ontime  0.037593985 0.240601504 0.186090226 0.124060150 0.013157895
##          CARRIER
## Y                 RU          UA          US
##   delayed 0.206225681 0.011673152 0.093385214
##   ontime  0.178571429 0.015037594 0.204887218
##
##          DEST
## Y               EWR       JFK       LGA
##   delayed 0.3891051 0.2217899 0.3891051
##   ontime  0.2819549 0.1823308 0.5357143
##
##          ORIGIN
## Y                BWI        DCA        IAD
##   delayed 0.07392996 0.51361868 0.41245136
##   ontime  0.06109023 0.64849624 0.29041353
##
##          DAY_WEEK
## Y                 1          2          3          4          5          6
##   delayed 0.18677043 0.15953307 0.11284047 0.15175097 0.17509728 0.05447471
##   ontime  0.14473684 0.12687970 0.13439850 0.18139098 0.18421053 0.12312030
##          DAY_WEEK
## Y                 7
##   delayed 0.15953307
##   ontime  0.10526316
```

Pivot table for Flight status by destination

```
pr<-prop.table(table(Trg$Flight.Status, Trg$DEST), margin = 1)
pr
```

```
##
##                 EWR       JFK       LGA
##   delayed 0.3891051 0.2217899 0.3891051
##   ontime  0.2819549 0.1823308 0.5357143
```

Using the model on Test set

```
# Predict probalities
PredProb <- predict(nb_model, newdata = Test, type = "raw")
#Predict class
PredClass <- predict(nb_model, newdata = Test)

#df <- data.frame(actual = Test$Flight.Status, predicted = PredClass, PredProb)
#df[Test$CARRIER == "DL" & Test$DAY_WEEK == 7 & Test$CRS_DEP_TIME == 10 &
#Test$DEST == "LGA" & Test$ORIGIN == "DCA",]
```

Confusion Matrix

```
#library(caret)
# training
#pred.class <- predict(nb_model, newdata = Trg)
#confusionMatrix(pred.class, Trg$Flight.Status)
# validation
#pred.class <- predict(nb_model, newdata = Test)
#confusionMatrix(pred.class, Test$Flight.Status)
```

Plot ROC curve

```
library(caret)
# training
#pred.class <- predict(delays.nb, newdata = train.df)
#confusionMatrix(pred.class, train.df$Flight.Status)
# validation
#pred.class <- predict(delays.nb, newdata = valid.df)
#confusionMatrix(pred.class, valid.df$Flight.Status)
```