---

title: "Assignment3"

author: "Tim"

date: "October 24, 2019"

output:

pdf_document: default

html_document: default

```{r}
library(caret)
library(ISLR)
library(e1071)
library(gmodels)
library(dummies)
library(dplyr)
```

```{r}
FlightDelays <- read.csv("FlightDelays.csv")
FlightDelays$CRS_DEP_TIME<-as.factor(FlightDelays$CRS_DEP_TIME)
FlightDelays$DAY_WEEK<-as.factor(FlightDelays$DAY_WEEK)
FlightDelays <- FlightDelays[,c(10, 8, 1, 2, 4, 13)]
```

```{r}
**#1**
Index_Train<-createDataPartition(FlightDelays$Flight.Status, p=0.6, list=FALSE)
#Use 60% of data for training and the rest for validation
Train.data <-FlightDelays[Index_Train,]

Valid.data <-FlightDelays[-Index_Train,]
```

 # Building a naïve Bayes classifier

2 >

```{r}
nb_model <- naiveBayes(Flight.Status~CRS_DEP_TIME+CARRIER+DEST+ORIGIN+DAY_WEEK,data = Train.data)

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                 600          630          640          645          700
730          735
  delayed 0.0077821012 0.0116731518 0.0272373541 0.0000000000 0.0389105058
0.0077821012 0.0038910506
  ontime  0.0159774436 0.0347744361 0.0075187970 0.0150375940 0.0460526316
0.0131578947 0.0084586466
         CRS_DEP_TIME
Y                 759          800          830          840          845
850          900
  delayed 0.0000000000 0.0311284047 0.0038910506 0.0155642023 0.0000000000
0.0077821012 0.0233463035
  ontime  0.0018796992 0.0187969925 0.0056390977 0.0300751880 0.0018796992
0.0140977444 0.0385338346
         CRS_DEP_TIME
Y                 925          930         1000         1030         1039
1040         1100
  delayed 0.0000000000 0.0038910506 0.0000000000 0.0155642023 0.0000000000
0.0000000000 0.0116731518
  ontime  0.0018796992 0.0140977444 0.0093984962 0.0234962406 0.0028195489
0.0028195489 0.0225563910
         CRS_DEP_TIME
Y                1130         1200         1230         1240         1245
1300         1315
  delayed 0.0038910506 0.0000000000 0.0000000000 0.0116731518 0.0389105058
0.0389105058 0.0077821012
  ontime  0.0112781955 0.0159774436 0.0150375940 0.0140977444 0.0263157895
0.0516917293 0.0009398496
         CRS_DEP_TIME
Y                1330         1359         1400         1430         1455
1500         1515
  delayed 0.0000000000 0.0116731518 0.0194552529 0.0155642023 0.1167315175
0.0428015564 0.0077821012
  ontime  0.0131578947 0.0112781955 0.0169172932 0.0225563910 0.0563909774
0.0338345865 0.0009398496
```

```
         CRS_DEP_TIME
Y                1520         1525         1530         1600         1605
1610         1630
  delayed 0.0000000000 0.0194552529 0.0311284047 0.0311284047 0.0000000000
0.0116731518 0.0077821012
  ontime  0.0009398496 0.0037593985 0.0263157895 0.0197368421 0.0000000000
0.0140977444 0.0234962406
         CRS_DEP_TIME
Y                1640         1645         1700         1710         1715
1720         1725
  delayed 0.0155642023 0.0000000000 0.0272373541 0.0155642023 0.0428015564
0.0272373541 0.0000000000
  ontime  0.0112781955 0.0178571429 0.0338345865 0.0150375940 0.0206766917
0.0093984962 0.0009398496
         CRS_DEP_TIME
Y                1730         1800         1830         1900         1930
2000         2030
  delayed 0.0350194553 0.0000000000 0.0155642023 0.0700389105 0.0077821012
0.0155642023 0.0116731518
  ontime  0.0225563910 0.0169172932 0.0244360902 0.0347744361 0.0084586466
0.0065789474 0.0150375940
         CRS_DEP_TIME
Y                2100         2120         2130
  delayed 0.0194552529 0.0700389105 0.0000000000
  ontime  0.0234962406 0.0310150376 0.0000000000

         CARRIER
Y                CO           DH           DL           MQ           OH
RU           UA
  delayed 0.054474708 0.311284047 0.101167315 0.190661479 0.007782101
0.229571984 0.007782101
  ontime  0.041353383 0.233082707 0.185150376 0.131578947 0.014097744
0.186090226 0.014097744
         CARRIER
Y                US
  delayed 0.097276265
  ontime  0.194548872

         DEST
Y               EWR        JFK        LGA
  delayed 0.3852140 0.1984436 0.4163424
  ontime  0.3007519 0.1813910 0.5178571

         ORIGIN
Y                BWI        DCA        IAD
  delayed 0.08560311 0.52918288 0.38521401
  ontime  0.06672932 0.64191729 0.29135338

         DAY_WEEK
Y                1          2          3          4          5          6
7
  delayed 0.17120623 0.15564202 0.15175097 0.12840467 0.17509728 0.05447471
0.16342412
  ontime  0.12687970 0.15413534 0.13533835 0.17293233 0.18796992 0.12593985
0.09680451
```

3 >

```{r}

CrossTable(x=Train.data$Flight.Status,y=Train.data$DEST, prop.chisq = FALSE)
prop.table(table(Train.data$DEST, Train.data$Flight.Status))
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  1321


                           | Train.data$DEST
Train.data$Flight.Status   |       EWR |       JFK |       LGA | Row Total |
---------------------------|-----------|-----------|-----------|-----------|
                  delayed  |        99 |        51 |       107 |       257 |
                           |     0.385 |     0.198 |     0.416 |     0.195 |
                           |     0.236 |     0.209 |     0.163 |           |
                           |     0.075 |     0.039 |     0.081 |           |
---------------------------|-----------|-----------|-----------|-----------|
                   ontime  |       320 |       193 |       551 |      1064 |
                           |     0.301 |     0.181 |     0.518 |     0.805 |
                           |     0.764 |     0.791 |     0.837 |           |
                           |     0.242 |     0.146 |     0.417 |           |
---------------------------|-----------|-----------|-----------|-----------|
             Column Total  |       419 |       244 |       658 |      1321 |
                           |     0.317 |     0.185 |     0.498 |           |
---------------------------|-----------|-----------|-----------|-----------|



        delayed     ontime
  EWR 0.07494322 0.24224073
  JFK 0.03860712 0.14610144
  LGA 0.08099924 0.41710825
```

4 >

```{r}
library(pROC)

Predicted_Valid_labels <-predict(nb_model,Valid.data)
CrossTable(x=Valid.data$Flight.Status,y=Predicted_Valid_labels, prop.chisq = FALSE)
confusionMatrix(Predicted_Valid_labels, Valid.data$Flight.Status)

Pred_Valid1_labels <-predict(nb_model,Valid.data, type = "raw")
head(Pred_Valid1_labels)
#Passing the second column of the predicted probabilities
#That column contains the probability associate to 'ontime'
roc(Valid.data$Flight.Status, Pred_Valid1_labels[,2])
plot.roc(Valid.data$Flight.Status,Pred_Valid1_labels[,2])
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  880


                         | Predicted_Valid_labels
Valid.data$Flight.Status |   delayed |    ontime | Row Total |
-------------------------|-----------|-----------|-----------|
                 delayed |        32 |       139 |       171 |
                         |     0.187 |     0.813 |     0.194 |
                         |     0.386 |     0.174 |           |
                         |     0.036 |     0.158 |           |
-------------------------|-----------|-----------|-----------|
                  ontime |        51 |       658 |       709 |
                         |     0.072 |     0.928 |     0.806 |
                         |     0.614 |     0.826 |           |
                         |     0.058 |     0.748 |           |
-------------------------|-----------|-----------|-----------|
            Column Total |        83 |       797 |       880 |
                         |     0.094 |     0.906 |           |
-------------------------|-----------|-----------|-----------|


Confusion Matrix and Statistics

          Reference
Prediction delayed ontime
   delayed      32     51
   ontime      139    658

               Accuracy : 0.7841
                 95% CI : (0.7554, 0.8109)
    No Information Rate : 0.8057
    P-Value [Acc > NIR] : 0.9502
```

```
                 Kappa : 0.1432

 Mcnemar's Test P-Value : 2.761e-10

            Sensitivity : 0.18713
            Specificity : 0.92807
         Pos Pred Value : 0.38554
         Neg Pred Value : 0.82560
             Prevalence : 0.19432
         Detection Rate : 0.03636
   Detection Prevalence : 0.09432
      Balanced Accuracy : 0.55760

       'Positive' Class : delayed

        delayed    ontime
[1,] 0.15203475 0.8479652
[2,] 0.41764799 0.5823520
[3,] 0.15465752 0.8453425
[4,] 0.01761083 0.9823892
[5,] 0.09449460 0.9055054
[6,] 0.16533292 0.8346671
Setting levels: control = delayed, case = ontime
Setting direction: controls < cases

Call:
roc.default(response = Valid.data$Flight.Status, predictor =
Pred_Valid1_labels[,     2])

Data: Pred_Valid1_labels[, 2] in 171 controls (Valid.data$Flight.Status
delayed) < 709 cases (Valid.data$Flight.Status ontime).
Area under the curve: 0.6859
Setting levels: control = delayed, case = ontime
Setting direction: controls < cases
```

R Console