KENTUCKY STATE UNIVERSITY, COMPUTER SCIENCE

Data Mining
HW6

Anil Varma B

December 2, 2019

1. Apply Naive Bayes algorithm to attached UniversalBank data . you need to partition the data into two parts: training and testing, compare the results.

Preliminary steps for the data set, remove columns ID and Zip code as they are not useful. Change the CreditCard column to categorical since we use naive bayes for prediction.

```
library(e1071)
ub <-read.csv(file.choose(), header = TRUE)
head(ub)
```

```
> head(ub)
  Age Experience Income Family CCAvg Education Mortgage Personal.Loan Securities.Account CD.Account
1  25          1     49      4   1.6         1        0             0                  1          0
2  45         19     34      3   1.5         1        0             0                  1          0
3  39         15     11      1   1.0         1        0             0                  0          0
4  35          9    100      1   2.7         2        0             0                  0          0
5  35          8     45      4   1.0         2        0             0                  0          0
6  37         13     29      4   0.4         2      155             0                  0          0
  Online CreditCard
1      0         no
2      0         no
3      0         no
4      0         no
5      0        yes
6      1         no
```

Figure 0.1: Reading the data set

Splitting the data set into training and test data set

```
trainingRowIndex <- sample(1:nrow(ub), 0.8*nrow(ub))
trainingData <- ub[trainingRowIndex, ]
testData <- ub[-trainingRowIndex, ]
```

Applying naive bayes function on the training dataset

```
a_naive <-naiveBayes(CreditCard ~., data = trainingData)
a_naive
```

```
> a_naive

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
     no     yes
0.70425 0.29575

Conditional probabilities:
     Age
Y          [,1]     [,2]
  no   45.33333 11.42997
  yes  45.43702 11.42574

     Experience
Y          [,1]     [,2]
  no   20.09336 11.43439
  yes  20.23753 11.40443

     Income
Y          [,1]     [,2]
  no   73.44728 46.04276
  yes  74.44379 47.00121

     Family
Y          [,1]     [,2]
  no   2.386936 1.158013
  yes  2.415046 1.137431

     CCAvg
Y          [,1]     [,2]
  no   1.917618 1.739082
  yes  1.926374 1.752867

     Education
Y          [,1]      [,2]
  no   1.905573 0.8412489
  yes  1.860524 0.8369565

     Mortgage
Y          [,1]      [,2]
  no   57.53426 103.01880
  yes  55.41251  99.33395

     Personal.Loan
Y           [,1]      [,2]
  no   0.09371672 0.2914859
  yes  0.10397295 0.3053545

     Securities.Account
Y           [,1]      [,2]
  no   0.10933617 0.3121159
```

Figure 0.2: Summary of naive bayes

Checking the prediction of Credit Card to be given.

```
a_predict <-predict(a_naive,trainingData )
table(a_predict, trainingData$CreditCard)
```

```
> a_predict <-predict(a_naive,trainingData )
> table(a_predict, trainingData$CreditCard)

a_predict   no  yes
      no  2767  981
      yes   50  202
```

Figure 0.3: Confusion matrix of prediction