Data Science Project

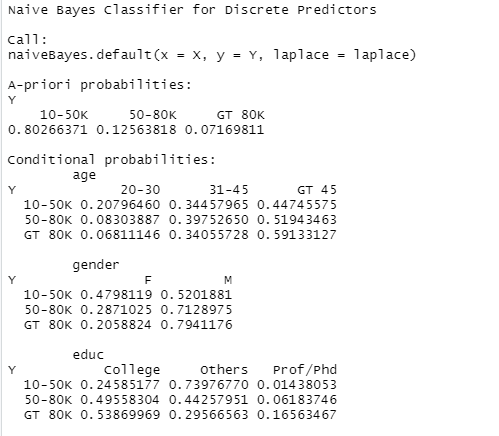
Naïve Bayes

SC Department

|  |  |
| --- | --- |
| id | Name |
| 2017170500 | هبه جمال صالح |
| 2017170308 | مارتينا جرجس عبدالنور |
| 2017170556 | مايكل منصور غالي |

Part 1: income

* Print the classifier output (the model’s priori and conditional probabilities) and explain it.



A-priori probabilities: it is categories the income values to 3 ranges and calculate the probability for each range.

P(10-50k) = 0.80266371

P(50-80k) = 0.12563818  
P (Greater than 80K) = 0.07169811  
conditional probabilities:

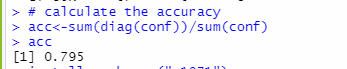
P (income| age) =

P (10-50k| 20-30) =

P(10-50k|20-30) = 0.20796460

And so on …

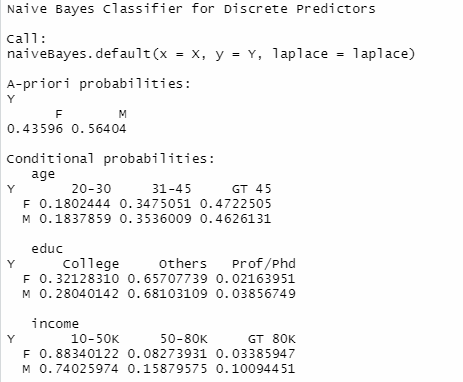
* Calculate your model accuracy. Write your observation about this model.



I see that the accuracy is good, so the columns (age, gender, education) sufficient to predict the income for the person.

Part 2: gender

* Write down the classifier output and explain it.



A-priori probabilities:

P(F)=0.43596

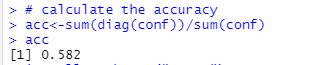
P(M)=0.56404

conditional probabilities:

P(Female|20-30) = 0.1802444

And so on …

* Calculate this model accuracy. Write your observation about this model and explain whether it is improved or not and why?



The accuracy goes down in compare of the part 1.

The columns (education, age, income) it is not sufficient to predict the gender   
because that there is not significant variation in values of conditional probability between both male and female given the (education, age, income) columns.

Part 3:

* How well does the model classify the testing data? Explain why.



It is less than the part two because we train with fewer data.

* Repeat step (f) 4 several times. What effect does the random selection of records have on the model’s performance?



The model performance has no change when random selection the records   
because that the probabilities did not change so the output of the bayes rule did not change so do the prediction of naïve bayes.

* What conclusions can one draw from this whole task?

1. (Education, gender, age) are sufficient to predict the income.
2. (Income, age, education) not sufficient to predict the gender.
3. The randomize does not affect the performance of the model, because the probabilities for each gender given that other variables do not change.

**Part 1 code:**

#set the working directory

setwd('H:\\4.2\\4.2\\Data Sceince\\Data Science Practical Task-20210526\\Dataset for practical task')

# install the package

install.packages("e1071")

# Loading package

library(e1071)

# read the csv file into table

sample <- read.table('nbtrain.csv',header = TRUE,sep = ',')

# split the data into train and test

traindata <- sample[1:9010,1:4]

testdata <- sample[9011:10010,1:4]

# call the naive bayes by income

model <- naiveBayes(income ~.,traindata)

model

# get the prediction for test data

y\_pred <- predict(model, newdata = testdata)

# create confusion matrix for the result

conf <- table(actual=testdata$income,predicted=y\_pred)

conf

# calculate the accuracy

acc<-sum(diag(conf))/sum(conf)

acc

**part 2 code:**

#set the working directory

setwd('H:\\4.2\\4.2\\Data Sceince\\Data Science Practical Task-20210526\\Dataset for practical task')

# install the package

install.packages("e1071")

# Loading package

library(e1071)

# read the csv file into table

sample <- read.table('nbtrain.csv',header = TRUE,sep = ',')

# split the data into train and test

traindata <- sample[1:9010,1:4]

testdata <- sample[9011:10010,1:4]

# call the naive bayes by gender

model <- naiveBayes(gender ~.,traindata)

model

# get the prediction for test data

y\_pred <- predict(model, newdata = testdata)

# create confusion matrix for the result

conf <- table(actual=testdata$gender,predicted=y\_pred)

conf

# calculate the accuracy

acc<-sum(diag(conf))/sum(conf)

acc

**part 3 code:**

#set the working directory

setwd('H:\\4.2\\4.2\\Data Sceince\\Data Science Practical Task-20210526\\Dataset for practical task')

# install the package

install.packages("e1071")

# Loading package

library(e1071)

# read the csv file into table

sample <- read.table('nbtrain.csv',header = TRUE,sep = ',')

# split the data into train and test

traindata <- sample[1:9010,1:4]

testdata <- sample[9011:10010,1:4]

# filter the training data with gender (male , female)

trainmale=traindata[traindata$gender == 'M',]

trainfemale=traindata[traindata$gender == 'F',]

# try 3 times to get random records

for (i in 1:4){

# get 3500 record randomly from each gender

a <- trainmale[sample(nrow(trainmale), 3500 ), ]

b <- trainfemale[sample(nrow(trainfemale), 3500 ), ]

# combine each 3500 record in one data frame

new\_traindata<-rbind(a,b)

# call naive bayes model on gender col

model <- naiveBayes(gender ~.,new\_traindata)

print(model)

# get the predictions

y\_pred <- predict(model, newdata = testdata)

# create the confusion metrics

conf <- table(actual=testdata$gender,predicted=y\_pred)

# calculate the accuracy

acc <-sum(diag(conf))/sum(conf)

print(acc)

}