CIS 635 Data Mining Name \_Dhynasah Çakir\_\_\_

Homework 5

Hand in sheet

Write in or copy and paste the answers for the following:

### Part 1

virus = n virus = y

1. priors \_\_\_0.4\_\_\_\_\_ \_\_\_\_\_0.6\_\_\_\_
2. likelihoods
   1. gender = F \_\_\_\_\_0.4\_\_\_\_\_ \_\_\_\_\_\_\_0.6\_\_\_\_\_\_  
        
      gender = M \_\_\_0.8\_\_\_\_ \_\_\_\_.2\_\_\_\_\_
   2. ageGroup = 1 \_\_\_\_.33\_\_\_\_\_ \_\_\_\_\_.66\_\_\_\_\_\_  
        
      ageGroup = 2 \_\_\_0.5\_\_\_\_\_ \_\_\_\_\_\_\_0.5\_\_\_\_\_\_  
        
      ageGroup = 3 \_\_\_1\_\_\_\_\_ \_\_\_\_\_\_\_\_0\_\_\_\_\_
3. likelihoods (numeric)
   1. pulse mean \_\_\_\_79\_\_\_ \_\_\_\_\_68.25\_\_\_\_\_\_  
        
      pulse sd \_\_\_\_13.6\_\_\_\_ \_\_\_\_\_\_\_\_8.8\_\_\_\_\_  
        
      test mean \_\_\_66.2\_\_\_\_\_\_\_ \_\_\_\_\_45.8\_\_\_\_\_\_  
        
      test sd \_\_\_\_\_5.9\_\_\_\_\_ \_\_\_\_\_\_5.37\_\_\_\_\_\_\_

### Part 2

1. \_\_a patient who is male is more likely to get the virus, a patient in group group 3 were more likely to have the virus. the average heart rate of someone with the virus is 75. The average test of a patient with a virus is 52
2. \_\_\_\_more likely to be male\_\_\_\_
3. categorize (fill in the numbers for virus=n and virus=y)
   1. \_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_
   2. \_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_
   3. \_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_

### Part 3

1. attach the code for the function you wrote and answer part e for each dataset below  
     
   e) hw05a \_\_\_\_\_\_pulse\_\_\_\_\_ \_\_\_\_\_\_\_temp\_\_\_\_\_\_  
     
    hw05b \_\_\_\_\_pulse\_\_\_\_\_\_ \_\_\_\_\_\_\_temp\_\_\_\_\_\_  
     
    hw05c \_\_\_\_\_pulse\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_
2. general questions
   1. \_\_\_\_\_\_\_\_\_normal distribution\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   2. \_\_\_\_\_\_\_\_no I don’t think so. And the reason is using a naïve bayes model would be assuming a normal distribution even if its not, and this would mean the analysis is missing a significant factor in the data and could lead to a wrong or misleading conclusion.

#part 3

cData = read.table("hw05c.txt", header =TRUE)

# read data file

aData=read.table("hw05a.txt",header=TRUE)

bData=read.table("hw05b.txt",header=TRUE)

for(i in c(2,3,6)){aData[,i]=as.factor(aData[,i])}

for(i in c(2,3,6)){bData[,i]=as.factor(bData[,i])}

for(i in c(2,3,6)){cData[,i]=as.factor(cData[,i])}

ind = sample(nrow(aData)) %% 10

ind = sample(nrow(bData)) %% 10

ind = sample(nrow(cData)) %% 10

modA=naiveBayes(virus~.,aData[ind!=0,2:6])

modB = naiveBayes(virus~.,bData[ind!=0,2:6])

modC=naiveBayes(virus~.,cData[ind!=0,2:6])

modA

modB

modC