Tutorial 7

DSA1101
Introduction to Data Science

October 19, 2018

Exercise 1. The Naïve Bayes Classifier

This week, we will look at the CSV dataset "Titanic.csv" which provides information on the fate of passengers on the fatal maiden voyage of the ocean liner *Titanic*, and includes the variables economic status (class), sex, age and survival. We will train a naïve Bayes classifier using this dataset, and predict survival.

(a) Load the dataset "Titanic.csv" which has been posted under the folder for Tutorial 7.

```r
Titanic_dataset= read.csv("Titanic.csv")
dim(Titanic_dataset)
head(Titanic_dataset)
```

(b) Compute the probabilities $P(Y = 1)$ (survived) and $P(Y = 0)$ (did not survive).

```r
tprior <- table(Titanic_dataset$Survived)
tprior
tprior <- tprior/sum(tprior)
tprior
```

(c) Compute the conditional probabilities $P(X_i = x_i|Y = 1)$ and $P(X_i = x_i|Y = 0)$, where $i = 1, 2, 3, 4$ for the feature variables $X = \{class, sex, age\}$.

```
classCounts <- table(Titanic_dataset[,c("Survived", "Class")])
classCounts <- classCounts/rowSums(classCounts)
classCounts

genderCounts <- table(Titanic_dataset[,c("Survived", "Sex")])
genderCounts <- genderCounts/rowSums(genderCounts)
genderCounts

ageCounts <- table(Titanic_dataset[,c("Survived", "Age")])
ageCounts <- ageCounts/rowSums(ageCounts)
ageCounts
```

(d) Predict survival for an adult female passenger in 2^{nd} class cabin.

```
prob_survived <-
classCounts["Yes","2nd"]*
genderCounts["Yes","Female"]*
ageCounts["Yes","Adult"]*
tprior["Yes"]

prob_not_survived <-
classCounts["No","2nd"]*
genderCounts["No","Female"]*
ageCounts["No","Adult"]*
tprior["No"]

prob_survived
prob_not_survived
```

(e) Compare your prediction in (d) with the one performed by the `naiveBayes` function in package 'e1071'

```
library(e1071)

model <- naiveBayes(Survived ~.,
Titanic_dataset)

test <- data.frame(Class="2nd", Sex="Female",
Age="Adult")

results <- predict(model,test)
results
results <- predict(model,test, "raw")
results

#ratio of probability scores
prob_survived/prob_not_survived
#ratio of actual probabilities
results[1,"Yes"]/results[1,"No"]
```