CMTH 642 Data Analytics: Advanced Methods

Assignment 3 (10%)

Ajay Herod

D10 500764104

#### 1. Import to R the following fiel: <http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv> (The dataset is related to white Portuguese “Vinho Verde” wine. For more info: <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>) (3 points)

#install.packages('RCurl')  
library(RCurl)  
web <- getURL("http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv")  
wine <- read.csv2(text = web)

#### 2. Check the datatypes of the attributes. (3 points)

str(wine)

## 'data.frame': 4898 obs. of 12 variables:  
## $ fixed.acidity : chr "7" "6.3" "8.1" "7.2" ...  
## $ volatile.acidity : chr "0.27" "0.3" "0.28" "0.23" ...  
## $ citric.acid : chr "0.36" "0.34" "0.4" "0.32" ...  
## $ residual.sugar : chr "20.7" "1.6" "6.9" "8.5" ...  
## $ chlorides : chr "0.045" "0.049" "0.05" "0.058" ...  
## $ free.sulfur.dioxide : chr "45" "14" "30" "47" ...  
## $ total.sulfur.dioxide: chr "170" "132" "97" "186" ...  
## $ density : chr "1.001" "0.994" "0.9951" "0.9956" ...  
## $ pH : chr "3" "3.3" "3.26" "3.19" ...  
## $ sulphates : chr "0.45" "0.49" "0.44" "0.4" ...  
## $ alcohol : chr "8.8" "9.5" "10.1" "9.9" ...  
## $ quality : int 6 6 6 6 6 6 6 6 6 6 ...

#OR  
sapply(wine, class)

## fixed.acidity volatile.acidity citric.acid   
## "character" "character" "character"   
## residual.sugar chlorides free.sulfur.dioxide   
## "character" "character" "character"   
## total.sulfur.dioxide density pH   
## "character" "character" "character"   
## sulphates alcohol quality   
## "character" "character" "integer"

#### 3. Are there any missing values in the dataset? (4 points)

sum(is.na(wine))

## [1] 0

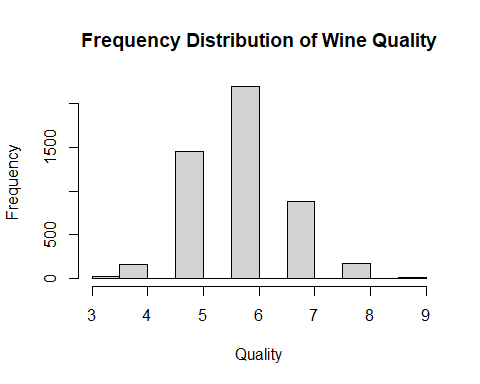
#### 4. What is the correlation between the attributes other than Quality? (10 points)

wine <- as.data.frame(apply(wine, 2, as.numeric))  
cor(wine[1:11])

## fixed.acidity volatile.acidity citric.acid residual.sugar  
## fixed.acidity 1.00000000 -0.02269729 0.28918070 0.08902070  
## volatile.acidity -0.02269729 1.00000000 -0.14947181 0.06428606  
## citric.acid 0.28918070 -0.14947181 1.00000000 0.09421162  
## residual.sugar 0.08902070 0.06428606 0.09421162 1.00000000  
## chlorides 0.02308564 0.07051157 0.11436445 0.08868454  
## free.sulfur.dioxide -0.04939586 -0.09701194 0.09407722 0.29909835  
## total.sulfur.dioxide 0.09106976 0.08926050 0.12113080 0.40143931  
## density 0.26533101 0.02711385 0.14950257 0.83896645  
## pH -0.42585829 -0.03191537 -0.16374821 -0.19413345  
## sulphates -0.01714299 -0.03572815 0.06233094 -0.02666437  
## alcohol -0.12088112 0.06771794 -0.07572873 -0.45063122  
## chlorides free.sulfur.dioxide total.sulfur.dioxide  
## fixed.acidity 0.02308564 -0.0493958591 0.091069756  
## volatile.acidity 0.07051157 -0.0970119393 0.089260504  
## citric.acid 0.11436445 0.0940772210 0.121130798  
## residual.sugar 0.08868454 0.2990983537 0.401439311  
## chlorides 1.00000000 0.1013923521 0.198910300  
## free.sulfur.dioxide 0.10139235 1.0000000000 0.615500965  
## total.sulfur.dioxide 0.19891030 0.6155009650 1.000000000  
## density 0.25721132 0.2942104109 0.529881324  
## pH -0.09043946 -0.0006177961 0.002320972  
## sulphates 0.01676288 0.0592172458 0.134562367  
## alcohol -0.36018871 -0.2501039415 -0.448892102  
## density pH sulphates alcohol  
## fixed.acidity 0.26533101 -0.4258582910 -0.01714299 -0.12088112  
## volatile.acidity 0.02711385 -0.0319153683 -0.03572815 0.06771794  
## citric.acid 0.14950257 -0.1637482114 0.06233094 -0.07572873  
## residual.sugar 0.83896645 -0.1941334540 -0.02666437 -0.45063122  
## chlorides 0.25721132 -0.0904394560 0.01676288 -0.36018871  
## free.sulfur.dioxide 0.29421041 -0.0006177961 0.05921725 -0.25010394  
## total.sulfur.dioxide 0.52988132 0.0023209718 0.13456237 -0.44889210  
## density 1.00000000 -0.0935914935 0.07449315 -0.78013762  
## pH -0.09359149 1.0000000000 0.15595150 0.12143210  
## sulphates 0.07449315 0.1559514973 1.00000000 -0.01743277  
## alcohol -0.78013762 0.1214320987 -0.01743277 1.00000000

#### 5. Graph the frequency distribution of wine quality by using Quality. (10 points)

hist(wine$quality, xlab="Quality", main="Frequency Distribution of Wine Quality")



#### 6. Reduce the levels of rating for quality to three levels as high, medium and low. Assign the levels of 3 and 4 to level 0; 5 and 6 to level 1; and 7,8 and 9 to level 2. (10 points)

for (i in 1:length(wine$quality)){  
 if(wine$quality[i] == 3 || wine$quality[i] == 4){  
 wine$quality[i] = 0  
 }  
 else if(wine$quality[i] == 5 || wine$quality[i] == 6){  
 wine$quality[i] = 1  
 }  
 else {  
 wine$quality[i] = 2  
 }  
}

#### 7. Normalize the data set by using the following function: (12 points)

normalize <- function(x){  
 return ((x - min(x)) / (max(x) - min(x)))  
}

wine <- as.data.frame(lapply(wine, normalize))  
summary(wine)

## fixed.acidity volatile.acidity citric.acid residual.sugar   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.2404 1st Qu.:0.1275 1st Qu.:0.1627 1st Qu.:0.01687   
## Median :0.2885 Median :0.1765 Median :0.1928 Median :0.07055   
## Mean :0.2937 Mean :0.1944 Mean :0.2013 Mean :0.08883   
## 3rd Qu.:0.3365 3rd Qu.:0.2353 3rd Qu.:0.2349 3rd Qu.:0.14264   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.00000   
## chlorides free.sulfur.dioxide total.sulfur.dioxide density   
## Min. :0.00000 Min. :0.00000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.08012 1st Qu.:0.07317 1st Qu.:0.2297 1st Qu.:0.08892   
## Median :0.10089 Median :0.11150 Median :0.2900 Median :0.12782   
## Mean :0.10912 Mean :0.11606 Mean :0.3001 Mean :0.13336   
## 3rd Qu.:0.12166 3rd Qu.:0.15331 3rd Qu.:0.3666 3rd Qu.:0.17332   
## Max. :1.00000 Max. :1.00000 Max. :1.0000 Max. :1.00000   
## pH sulphates alcohol quality   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.3364 1st Qu.:0.2209 1st Qu.:0.2419 1st Qu.:0.5000   
## Median :0.4182 Median :0.2907 Median :0.3871 Median :0.5000   
## Mean :0.4257 Mean :0.3138 Mean :0.4055 Mean :0.5895   
## 3rd Qu.:0.5091 3rd Qu.:0.3837 3rd Qu.:0.5484 3rd Qu.:0.5000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000

#### 8. Divide the dataset to training and test sets. (12 points)

set.seed(1)  
index <- sample(1:nrow(wine), 0.7\*nrow(wine))  
wine\_train <- wine[index,]  
wine\_test <- wine[-index,]

#### 9. Use the KNN algorithm to predict the quality of wine using its attributes. (12 points)

#install.packages("class")  
library(class)  
wine\_train\_labels <- wine\_train[,12]  
wine\_test\_labels <- wine\_test[,12]  
wine\_test\_pred <- knn(train = wine\_train[,1:11],  
 test = wine\_test[,1:11],  
 cl = wine\_train\_labels, k=10)

#### 10. Display the confusion matrix to evaluate the model performance. (12 points)

table(Actual=wine\_test\_labels, Predicted=wine\_test\_pred)

## Predicted  
## Actual 0 0.5 1  
## 0 1 41 1  
## 0.5 1 1019 84  
## 1 0 171 152

#### 11. Evaluate the model performance by computing Accuracy, Sensitivity and Specificity. (12 points)

#Accuracy = (TN+TP)/Total  
(1019+152+1)/(1+41+1+1+1019+84+171+152)

## [1] 0.7972789

#Sensitivity = TP/(TP+FN) For Actual 0  
1/(1)

## [1] 1

#Specificity = TN/(TN+FP) For Actual 0  
(1019+84+171+152)/(1019+84+171+152+41+1)

## [1] 0.9713896

#Sensitivity = TP/(TP+FN) For Actual 0.5  
1019/(1019+41+171)

## [1] 0.8277823

#Specificity = TN/(TN+FP) For Actual 0.5  
(1+1+0+152)/(1+1+0+152+1+84)

## [1] 0.6443515

#Sensitivity = TP/(TP+FN) For Actual 1  
152/(152+84+1)

## [1] 0.6413502

#Specificity = TN/(TN+FP) For Actual 1  
(1+41+1+1019)/(1+41+1+1019+0+171)

## [1] 0.8613139

This is the end of Assignment 3

Ceni Babaoglu, PhD