CMTH642 - Assignment 3

Mohammed Amir

December 16, 2016

Upload data

wine <- read.csv(file="D:/Big Data/CMTH642 - DATA ANALYTICS ADVANCED METHODS/ASSIGNMENT 3/winequality-white.csv",head=TRUE,sep=";")

# 1. Check data characteristics. Is there missing data?

#is.na(wine)  
wine[!complete.cases(wine),]

## [1] fixed.acidity volatile.acidity citric.acid   
## [4] residual.sugar chlorides free.sulfur.dioxide   
## [7] total.sulfur.dioxide density pH   
## [10] sulphates alcohol quality   
## <0 rows> (or 0-length row.names)

str(wine)

## 'data.frame': 4898 obs. of 12 variables:  
## $ fixed.acidity : num 7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...  
## $ volatile.acidity : num 0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...  
## $ citric.acid : num 0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...  
## $ residual.sugar : num 20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...  
## $ chlorides : num 0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...  
## $ free.sulfur.dioxide : num 45 14 30 47 47 30 30 45 14 28 ...  
## $ total.sulfur.dioxide: num 170 132 97 186 186 97 136 170 132 129 ...  
## $ density : num 1.001 0.994 0.995 0.996 0.996 ...  
## $ pH : num 3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...  
## $ sulphates : num 0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...  
## $ alcohol : num 8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...  
## $ quality : int 6 6 6 6 6 6 6 6 6 6 ...

# 2. What is the correlation between the attributes other than wine quality?

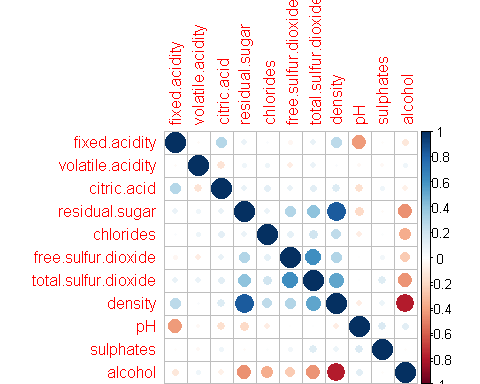
## remove the wine quality attribute  
newdf <- wine[-12]  
head(newdf)

## fixed.acidity volatile.acidity citric.acid residual.sugar chlorides  
## 1 7.0 0.27 0.36 20.7 0.045  
## 2 6.3 0.30 0.34 1.6 0.049  
## 3 8.1 0.28 0.40 6.9 0.050  
## 4 7.2 0.23 0.32 8.5 0.058  
## 5 7.2 0.23 0.32 8.5 0.058  
## 6 8.1 0.28 0.40 6.9 0.050  
## free.sulfur.dioxide total.sulfur.dioxide density pH sulphates alcohol  
## 1 45 170 1.0010 3.00 0.45 8.8  
## 2 14 132 0.9940 3.30 0.49 9.5  
## 3 30 97 0.9951 3.26 0.44 10.1  
## 4 47 186 0.9956 3.19 0.40 9.9  
## 5 47 186 0.9956 3.19 0.40 9.9  
## 6 30 97 0.9951 3.26 0.44 10.1

## Correlate wine attributes excluding quality  
wine\_correlation <- cor(newdf)  
wine\_correlation

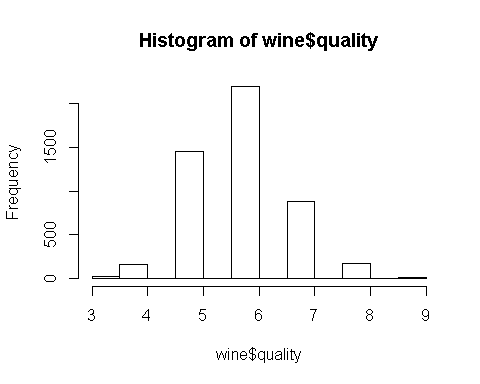
## fixed.acidity volatile.acidity citric.acid  
## fixed.acidity 1.00000000 -0.02269729 0.28918070  
## volatile.acidity -0.02269729 1.00000000 -0.14947181  
## citric.acid 0.28918070 -0.14947181 1.00000000  
## residual.sugar 0.08902070 0.06428606 0.09421162  
## chlorides 0.02308564 0.07051157 0.11436445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

## plot wine attribute correlation  
corrplot(wine\_correlation, method="circle")



# 3. Graph the distribution of wine quality.

hist(wine$quality)



# 4. Reduce the levels of rating for quality to three levels as high, medium and low

rating <- function(quality)  
{  
 if(quality <= 4)  
 {  
 print ("Low")  
 }  
 else  
 {  
 if(quality >= 5 && quality <= 7)  
 {  
 print ("Medium")   
 }  
 else  
 {  
 print ("High")  
 }  
 }  
}

## add rating attribute to the wine data frame  
wine$rating <- mapply(rating,wine$quality)

head(wine)

## fixed.acidity volatile.acidity citric.acid residual.sugar chlorides  
## 1 7.0 0.27 0.36 20.7 0.045  
## 2 6.3 0.30 0.34 1.6 0.049  
## 3 8.1 0.28 0.40 6.9 0.050  
## 4 7.2 0.23 0.32 8.5 0.058  
## 5 7.2 0.23 0.32 8.5 0.058  
## 6 8.1 0.28 0.40 6.9 0.050  
## free.sulfur.dioxide total.sulfur.dioxide density pH sulphates alcohol  
## 1 45 170 1.0010 3.00 0.45 8.8  
## 2 14 132 0.9940 3.30 0.49 9.5  
## 3 30 97 0.9951 3.26 0.44 10.1  
## 4 47 186 0.9956 3.19 0.40 9.9  
## 5 47 186 0.9956 3.19 0.40 9.9  
## 6 30 97 0.9951 3.26 0.44 10.1  
## quality rating  
## 1 6 Medium  
## 2 6 Medium  
## 3 6 Medium  
## 4 6 Medium  
## 5 6 Medium  
## 6 6 Medium

# 5. Normalize the data set.

# remove quality attribute from wine data frame before normalization as it is not numeric attribute  
new\_wine\_attribute <- wine[-12]  
  
# Normaization function  
normalize <- function(x)   
 {  
 return (   
 ( (x - min(x)) / (max(x) - min(x)))  
 )   
 }  
  
dfNorm <- as.data.frame(lapply(new\_wine\_attribute[1:11], normalize))  
head(dfNorm)

## fixed.acidity volatile.acidity citric.acid residual.sugar chlorides  
## 1 0.3076923 0.1862745 0.2168675 0.30828221 0.1068249  
## 2 0.2403846 0.2156863 0.2048193 0.01533742 0.1186944  
## 3 0.4134615 0.1960784 0.2409639 0.09662577 0.1216617  
## 4 0.3269231 0.1470588 0.1927711 0.12116564 0.1454006  
## 5 0.3269231 0.1470588 0.1927711 0.12116564 0.1454006  
## 6 0.4134615 0.1960784 0.2409639 0.09662577 0.1216617  
## free.sulfur.dioxide total.sulfur.dioxide density pH sulphates  
## 1 0.14982578 0.3735499 0.2677848 0.2545455 0.2674419  
## 2 0.04181185 0.2853828 0.1328321 0.5272727 0.3139535  
## 3 0.09756098 0.2041763 0.1540389 0.4909091 0.2558140  
## 4 0.15679443 0.4106729 0.1636784 0.4272727 0.2093023  
## 5 0.15679443 0.4106729 0.1636784 0.4272727 0.2093023  
## 6 0.09756098 0.2041763 0.1540389 0.4909091 0.2558140  
## alcohol  
## 1 0.1290323  
## 2 0.2419355  
## 3 0.3387097  
## 4 0.3064516  
## 5 0.3064516  
## 6 0.3387097

# 6 Divide the data to training and testing groups.

dfNorm\_train <- dfNorm[1:3265,] ## two third of data set for training  
dfNorm\_test <- dfNorm[3266:4898,] ## one third of data set for testing  
nrow(dfNorm\_train)

## [1] 3265

nrow(dfNorm\_test)

## [1] 1633

# 7 Use the KNN algorithm to predict the quality of wine using its attributes.

dfNorm\_train\_labels <- wine[1:3265, 12] ## predict the quality in column 12  
dfNorm\_test\_labels <- wine[3266:4898,12] ## predict the quality in column 12  
  
prediction <- knn(train = dfNorm\_train, test = dfNorm\_test,cl = dfNorm\_train\_labels, k=10)

# 8 Evaluate the model performance.

#install.packages("gmodels")  
  
CrossTable(x=dfNorm\_test\_labels, y = prediction, prop.chisq=FALSE)

## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1633   
##   
##   
## | prediction   
## dfNorm\_test\_labels | 4 | 5 | 6 | 7 | 8 | Row Total |   
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|  
## 3 | 0 | 2 | 2 | 1 | 0 | 5 |   
## | 0.000 | 0.400 | 0.400 | 0.200 | 0.000 | 0.003 |   
## | 0.000 | 0.006 | 0.003 | 0.002 | 0.000 | |   
## | 0.000 | 0.001 | 0.001 | 0.001 | 0.000 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|  
## 4 | 0 | 20 | 14 | 4 | 1 | 39 |   
## | 0.000 | 0.513 | 0.359 | 0.103 | 0.026 | 0.024 |   
## | 0.000 | 0.057 | 0.018 | 0.008 | 0.032 | |   
## | 0.000 | 0.012 | 0.009 | 0.002 | 0.001 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|  
## 5 | 1 | 186 | 191 | 42 | 2 | 422 |   
## | 0.002 | 0.441 | 0.453 | 0.100 | 0.005 | 0.258 |   
## | 0.333 | 0.531 | 0.247 | 0.088 | 0.065 | |   
## | 0.001 | 0.114 | 0.117 | 0.026 | 0.001 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|  
## 6 | 2 | 137 | 425 | 240 | 11 | 815 |   
## | 0.002 | 0.168 | 0.521 | 0.294 | 0.013 | 0.499 |   
## | 0.667 | 0.391 | 0.549 | 0.505 | 0.355 | |   
## | 0.001 | 0.084 | 0.260 | 0.147 | 0.007 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|  
## 7 | 0 | 5 | 125 | 156 | 16 | 302 |   
## | 0.000 | 0.017 | 0.414 | 0.517 | 0.053 | 0.185 |   
## | 0.000 | 0.014 | 0.161 | 0.328 | 0.516 | |   
## | 0.000 | 0.003 | 0.077 | 0.096 | 0.010 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|  
## 8 | 0 | 0 | 17 | 32 | 1 | 50 |   
## | 0.000 | 0.000 | 0.340 | 0.640 | 0.020 | 0.031 |   
## | 0.000 | 0.000 | 0.022 | 0.067 | 0.032 | |   
## | 0.000 | 0.000 | 0.010 | 0.020 | 0.001 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 3 | 350 | 774 | 475 | 31 | 1633 |   
## | 0.002 | 0.214 | 0.474 | 0.291 | 0.019 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##