CMTH642 Data Analytics Advanced Method - Assignment 3

Mohammed Amir

December 16, 2016

Upload data

wine\_df <- 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\_df)

wine\_df[!complete.cases(wine\_df),]

## [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\_df)

## '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 ...

summary(wine\_df)

## fixed.acidity volatile.acidity citric.acid residual.sugar   
## Min. : 3.800 Min. :0.0800 Min. :0.0000 Min. : 0.600   
## 1st Qu.: 6.300 1st Qu.:0.2100 1st Qu.:0.2700 1st Qu.: 1.700   
## Median : 6.800 Median :0.2600 Median :0.3200 Median : 5.200   
## Mean : 6.855 Mean :0.2782 Mean :0.3342 Mean : 6.391   
## 3rd Qu.: 7.300 3rd Qu.:0.3200 3rd Qu.:0.3900 3rd Qu.: 9.900   
## Max. :14.200 Max. :1.1000 Max. :1.6600 Max. :65.800   
## chlorides free.sulfur.dioxide total.sulfur.dioxide  
## Min. :0.00900 Min. : 2.00 Min. : 9.0   
## 1st Qu.:0.03600 1st Qu.: 23.00 1st Qu.:108.0   
## Median :0.04300 Median : 34.00 Median :134.0   
## Mean :0.04577 Mean : 35.31 Mean :138.4   
## 3rd Qu.:0.05000 3rd Qu.: 46.00 3rd Qu.:167.0   
## Max. :0.34600 Max. :289.00 Max. :440.0   
## density pH sulphates alcohol   
## Min. :0.9871 Min. :2.720 Min. :0.2200 Min. : 8.00   
## 1st Qu.:0.9917 1st Qu.:3.090 1st Qu.:0.4100 1st Qu.: 9.50   
## Median :0.9937 Median :3.180 Median :0.4700 Median :10.40   
## Mean :0.9940 Mean :3.188 Mean :0.4898 Mean :10.51   
## 3rd Qu.:0.9961 3rd Qu.:3.280 3rd Qu.:0.5500 3rd Qu.:11.40   
## Max. :1.0390 Max. :3.820 Max. :1.0800 Max. :14.20   
## quality   
## Min. :3.000   
## 1st Qu.:5.000   
## Median :6.000   
## Mean :5.878   
## 3rd Qu.:6.000   
## Max. :9.000

head(wine\_df)

## 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  
## 1 6  
## 2 6  
## 3 6  
## 4 6  
## 5 6  
## 6 6

table(wine\_df$quality)

##   
## 3 4 5 6 7 8 9   
## 20 163 1457 2198 880 175 5

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

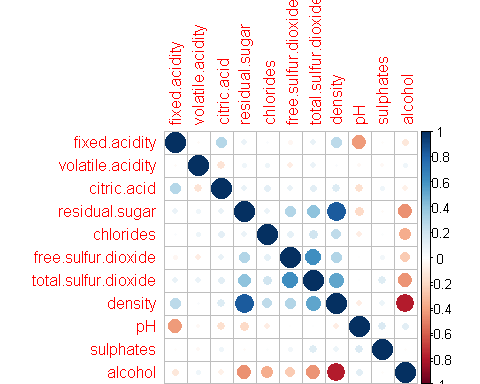
## remove the wine\_df quality attribute  
new\_wine\_df <- wine\_df[-12]  
head(new\_wine\_df)

## 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\_df attributes excluding quality  
wine\_df\_correlation <- cor(new\_wine\_df)  
wine\_df\_correlation

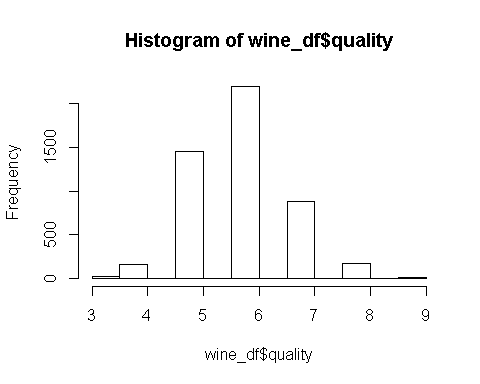
## fixed.acidity volatile.acidity citric.acid  
## fixed.acidity 1.00000000 -0.02269729 0.28918070  
## volatile.acidity -0.02269729 1.00000000 -0.14947181  
## citric.acid 0.28918070 -0.14947181 1.00000000  
## residual.sugar 0.08902070 0.06428606 0.09421162  
## chlorides 0.02308564 0.07051157 0.11436445  
## 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\_df attribute correlation  
corrplot(wine\_df\_correlation, method="circle")



# 3. Graph the distribution of wine quality.

hist(wine\_df$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\_df data frame  
wine\_df$rating <- mapply(rating,wine\_df$quality)

# Total Count for each rating  
 table(wine\_df$rating)

##   
## High Low Medium   
## 180 183 4535

# 5. Normalize the data set.

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

## 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

summary(winedf\_Norm)

## 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  
## Min. :0.00000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.08012 1st Qu.:0.07317 1st Qu.:0.2297   
## Median :0.10089 Median :0.11150 Median :0.2900   
## Mean :0.10912 Mean :0.11606 Mean :0.3001   
## 3rd Qu.:0.12166 3rd Qu.:0.15331 3rd Qu.:0.3666   
## Max. :1.00000 Max. :1.00000 Max. :1.0000   
## density pH sulphates alcohol   
## Min. :0.00000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.08892 1st Qu.:0.3364 1st Qu.:0.2209 1st Qu.:0.2419   
## Median :0.12782 Median :0.4182 Median :0.2907 Median :0.3871   
## Mean :0.13336 Mean :0.4257 Mean :0.3138 Mean :0.4055   
## 3rd Qu.:0.17332 3rd Qu.:0.5091 3rd Qu.:0.3837 3rd Qu.:0.5484   
## Max. :1.00000 Max. :1.0000 Max. :1.0000 Max. :1.0000

# 6 Divide the data to training and testing groups

## Randomly divide the datasett to two sections 67% to 33%   
ind <- sample(2, nrow(winedf\_Norm), replace=TRUE, prob=c(0.67, 0.33))  
winedf\_training <- winedf\_Norm[ind==1,] ## assign 0.67 to training  
winedf\_test <- winedf\_Norm[ind==2, ] ## assign 0.33 to test   
  
nrow(winedf\_training) # number of rows in training dataset

## [1] 3233

nrow(winedf\_test) # number of rows in testing dataset

## [1] 1665

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

winedf\_training\_Labels <- new\_wine\_df\_attribute[ind==1, 12] ## Create a train labels  
winedf\_test\_Labels <- new\_wine\_df\_attribute[ind==2, 12] ## Create a test labels  
  
#Apply a KNN function for K = 5  
wine\_prediction <- knn(train = winedf\_training, test = winedf\_test, cl = winedf\_training\_Labels, k=5)

# 8 Evaluate the model performance.

CrossTable(x=winedf\_test\_Labels, y = wine\_prediction, prop.chisq=FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1665   
##   
##   
## | wine\_prediction   
## winedf\_test\_Labels | High | Low | Medium | Row Total |   
## -------------------|-----------|-----------|-----------|-----------|  
## High | 4 | 0 | 58 | 62 |   
## | 0.065 | 0.000 | 0.935 | 0.037 |   
## | 0.286 | 0.000 | 0.035 | |   
## | 0.002 | 0.000 | 0.035 | |   
## -------------------|-----------|-----------|-----------|-----------|  
## Low | 0 | 4 | 60 | 64 |   
## | 0.000 | 0.062 | 0.938 | 0.038 |   
## | 0.000 | 0.800 | 0.036 | |   
## | 0.000 | 0.002 | 0.036 | |   
## -------------------|-----------|-----------|-----------|-----------|  
## Medium | 10 | 1 | 1528 | 1539 |   
## | 0.006 | 0.001 | 0.993 | 0.924 |   
## | 0.714 | 0.200 | 0.928 | |   
## | 0.006 | 0.001 | 0.918 | |   
## -------------------|-----------|-----------|-----------|-----------|  
## Column Total | 14 | 5 | 1646 | 1665 |   
## | 0.008 | 0.003 | 0.989 | |   
## -------------------|-----------|-----------|-----------|-----------|  
##   
##