Shetty\_S\_M7\_kNN

Sooraj Shetty

March 5, 2017

Load the necessary libraries as per requirement

require(ggplot2)

## Loading required package: ggplot2

require(class)

## Loading required package: class

require(useful)

## Loading required package: useful

require(cluster)

## Loading required package: cluster

require(amap)

## Loading required package: amap

require(energy)

## Loading required package: energy

require(gmodels)

## Loading required package: gmodels

Loading my data - Teaching Assistant Evaluation Data Set

data\_url <- 'https://archive.ics.uci.edu/ml/machine-learning-databases/tae/tae.data'  
TAE <- read.csv(url(data\_url),sep=",",header = FALSE)  
head(TAE)

## V1 V2 V3 V4 V5 V6  
## 1 1 23 3 1 19 3  
## 2 2 15 3 1 17 3  
## 3 1 23 3 2 49 3  
## 4 1 5 2 2 33 3  
## 5 2 7 11 2 55 3  
## 6 2 23 3 1 20 3

names(TAE)

## [1] "V1" "V2" "V3" "V4" "V5" "V6"

table(TAE$V6)

##   
## 1 2 3   
## 49 50 52

TAE$V6

## [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1  
## [36] 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1  
## [71] 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2  
## [106] 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 1 1  
## [141] 1 1 1 1 1 1 1 1 1 1 1

length(TAE$V6)

## [1] 151

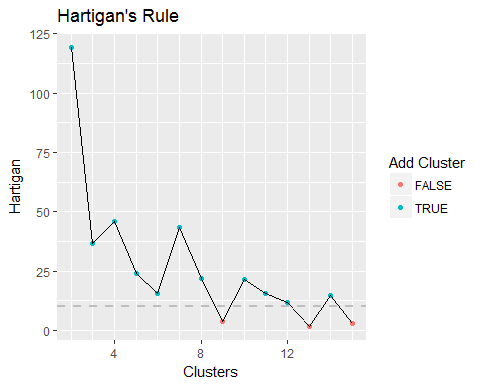
V6 <- Class attribute (categorical) 1=Low, 2=Medium, 3=High

Above information is obtained from the UCI website.

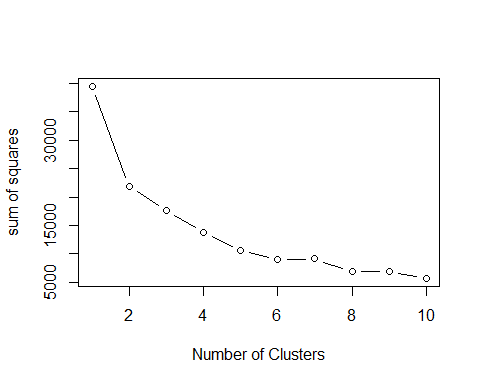
Below I am estimating number of clusters in my dataset for further analysis.

I am using Hartigan's Rule for this purpose.

# Hartigans's rule FitKMean (similarity)  
# require(useful)  
best<-FitKMeans(TAE,max.clusters=15, seed=111)   
PlotHartigan(best)



# Determining number of clusters   
sos <- (nrow(TAE)-1)\*sum(apply(TAE,2,var))  
for (i in 2:10) sos[i] <- sum(kmeans(TAE, centers=i)$withinss)  
plot(1:10, sos, type="b", xlab="Number of Clusters", ylab="sum of squares")



Scaling the data

TAE.scaled<-as.data.frame(lapply(TAE[,c(1:5)], scale))  
head(TAE.scaled)

## V1 V2 V3 V4 V5  
## 1 -2.0442691 1.3709228 -0.7269395 -2.3512468 -0.6877397  
## 2 0.4859328 0.1988954 -0.7269395 -2.3512468 -0.8428535  
## 3 -2.0442691 1.3709228 -0.7269395 0.4224897 1.6389674  
## 4 -2.0442691 -1.2661389 -0.8693103 0.4224897 0.3980570  
## 5 0.4859328 -0.9731320 0.4120267 0.4224897 2.1043089  
## 6 0.4859328 1.3709228 -0.7269395 -2.3512468 -0.6101828

Normalizing the numeric data in dataset with exception of cloumn V6 [target]. The normalization brings all values to a common format.

normalize <- function(x) {  
return ((x - min(x)) / (max(x) - min(x))) }  
  
TAE\_n <- as.data.frame(lapply(TAE[1:6], normalize))  
  
#Following summaries will show the normalization of numeric data  
summary(TAE\_n$V1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 1.0000 1.0000 0.8079 1.0000 1.0000

summary(TAE\_n$V2)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.2917 0.5000 0.5268 0.7917 1.0000

summary(TAE\_n$V3)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0800 0.1200 0.2842 0.5600 1.0000

summary(TAE\_n$V4)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 1.0000 1.0000 0.8477 1.0000 1.0000

summary(TAE\_n$V5)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.2540 0.3810 0.3947 0.5397 1.0000

summary(TAE\_n$V6)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.5000 0.5099 1.0000 1.0000

Forming training & test datasets from the normalized dataset

TAE\_train <- TAE\_n[1:75, 1:5] # First 75 rows  
head(TAE\_train)

## V1 V2 V3 V4 V5  
## 1 0 0.9166667 0.08 0 0.2539683  
## 2 1 0.5833333 0.08 0 0.2222222  
## 3 0 0.9166667 0.08 1 0.7301587  
## 4 0 0.1666667 0.04 1 0.4761905  
## 5 1 0.2500000 0.40 1 0.8253968  
## 6 1 0.9166667 0.08 0 0.2698413

TAE\_test <- TAE\_n[76:151, 1:5] # Next 76 rows  
head(TAE\_test)

## V1 V2 V3 V4 V5  
## 76 1 0.1250000 0.60 1 0.2857143  
## 77 1 0.1666667 0.04 1 0.7142857  
## 78 1 0.5416667 0.56 1 0.5555556  
## 79 0 0.9166667 0.08 0 0.3492063  
## 80 0 0.5000000 0.08 0 0.2222222  
## 81 1 0.6250000 0.72 1 0.1269841

The target variable which we have not included in our training and test data sets.

TAE\_train\_labels <- TAE[1:75, 6]  
TAE\_train\_labels

## [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1  
## [36] 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1  
## [71] 1 1 1 1 1

TAE\_test\_labels <- TAE[76:151, 6]  
TAE\_test\_labels

## [1] 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1  
## [36] 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1  
## [71] 1 1 1 1 1 1

k=2

TAE\_test\_pred <- knn(TAE\_train,TAE\_test,TAE\_train\_labels,k=2)  
TAE\_test\_pred

## [1] 2 1 2 3 1 3 1 2 3 3 2 3 3 3 3 3 3 3 3 2 2 3 3 3 3 3 3 3 3 3 1 1 1 3 3  
## [36] 3 2 3 2 1 3 1 1 2 3 1 2 3 3 2 3 3 1 1 1 3 3 3 3 3 3 3 1 2 3 1 3 3 1 1  
## [71] 1 1 3 3 2 2  
## Levels: 1 2 3

CrossTable(TAE\_test\_labels,TAE\_test\_pred, prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 76   
##   
##   
## | TAE\_test\_pred   
## TAE\_test\_labels | 1 | 2 | 3 | Row Total |   
## ----------------|-----------|-----------|-----------|-----------|  
## 1 | 11 | 8 | 11 | 30 |   
## | 0.367 | 0.267 | 0.367 | 0.395 |   
## | 0.579 | 0.571 | 0.256 | |   
## | 0.145 | 0.105 | 0.145 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## 2 | 5 | 1 | 16 | 22 |   
## | 0.227 | 0.045 | 0.727 | 0.289 |   
## | 0.263 | 0.071 | 0.372 | |   
## | 0.066 | 0.013 | 0.211 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## 3 | 3 | 5 | 16 | 24 |   
## | 0.125 | 0.208 | 0.667 | 0.316 |   
## | 0.158 | 0.357 | 0.372 | |   
## | 0.039 | 0.066 | 0.211 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## Column Total | 19 | 14 | 43 | 76 |   
## | 0.250 | 0.184 | 0.566 | |   
## ----------------|-----------|-----------|-----------|-----------|  
##   
##

cm <- table(TAE\_test\_labels,TAE\_test\_pred)  
cm

## TAE\_test\_pred  
## TAE\_test\_labels 1 2 3  
## 1 11 8 11  
## 2 5 1 16  
## 3 3 5 16

k=3

TAE\_test\_pred1 <- knn(TAE\_train,TAE\_test,TAE\_train\_labels,k=3)  
TAE\_test\_pred1

## [1] 2 1 2 3 3 3 1 2 3 3 2 3 3 3 3 3 2 3 3 2 2 3 3 3 1 3 3 1 2 3 1 1 1 3 3  
## [36] 2 2 3 2 1 2 1 2 2 3 1 2 3 3 2 3 3 1 1 1 2 3 3 3 2 3 3 2 2 2 1 2 3 1 1  
## [71] 2 1 2 3 2 2  
## Levels: 1 2 3

CrossTable(TAE\_test\_labels,TAE\_test\_pred1, prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 76   
##   
##   
## | TAE\_test\_pred1   
## TAE\_test\_labels | 1 | 2 | 3 | Row Total |   
## ----------------|-----------|-----------|-----------|-----------|  
## 1 | 9 | 15 | 6 | 30 |   
## | 0.300 | 0.500 | 0.200 | 0.395 |   
## | 0.529 | 0.577 | 0.182 | |   
## | 0.118 | 0.197 | 0.079 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## 2 | 6 | 5 | 11 | 22 |   
## | 0.273 | 0.227 | 0.500 | 0.289 |   
## | 0.353 | 0.192 | 0.333 | |   
## | 0.079 | 0.066 | 0.145 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## 3 | 2 | 6 | 16 | 24 |   
## | 0.083 | 0.250 | 0.667 | 0.316 |   
## | 0.118 | 0.231 | 0.485 | |   
## | 0.026 | 0.079 | 0.211 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## Column Total | 17 | 26 | 33 | 76 |   
## | 0.224 | 0.342 | 0.434 | |   
## ----------------|-----------|-----------|-----------|-----------|  
##   
##

cm1 <- table(TAE\_test\_labels,TAE\_test\_pred1)  
cm1

## TAE\_test\_pred1  
## TAE\_test\_labels 1 2 3  
## 1 9 15 6  
## 2 6 5 11  
## 3 2 6 16

k=5

TAE\_test\_pred2 <- knn(TAE\_train,TAE\_test,TAE\_train\_labels,k=5)  
TAE\_test\_pred2

## [1] 2 1 2 3 3 3 1 3 3 3 2 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 2 3 1 1 1 3 3  
## [36] 3 2 3 3 1 3 1 1 1 1 1 2 2 3 2 3 3 1 1 1 2 3 3 3 2 3 2 1 2 3 1 2 3 1 1  
## [71] 2 2 3 3 1 2  
## Levels: 1 2 3

CrossTable(TAE\_test\_labels,TAE\_test\_pred2, prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 76   
##   
##   
## | TAE\_test\_pred2   
## TAE\_test\_labels | 1 | 2 | 3 | Row Total |   
## ----------------|-----------|-----------|-----------|-----------|  
## 1 | 12 | 8 | 10 | 30 |   
## | 0.400 | 0.267 | 0.333 | 0.395 |   
## | 0.632 | 0.471 | 0.250 | |   
## | 0.158 | 0.105 | 0.132 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## 2 | 5 | 5 | 12 | 22 |   
## | 0.227 | 0.227 | 0.545 | 0.289 |   
## | 0.263 | 0.294 | 0.300 | |   
## | 0.066 | 0.066 | 0.158 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## 3 | 2 | 4 | 18 | 24 |   
## | 0.083 | 0.167 | 0.750 | 0.316 |   
## | 0.105 | 0.235 | 0.450 | |   
## | 0.026 | 0.053 | 0.237 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## Column Total | 19 | 17 | 40 | 76 |   
## | 0.250 | 0.224 | 0.526 | |   
## ----------------|-----------|-----------|-----------|-----------|  
##   
##

cm2 <- table(TAE\_test\_labels,TAE\_test\_pred2)  
cm2

## TAE\_test\_pred2  
## TAE\_test\_labels 1 2 3  
## 1 12 8 10  
## 2 5 5 12  
## 3 2 4 18

k=6

TAE\_test\_pred3 <- knn(TAE\_train,TAE\_test,TAE\_train\_labels,k=6)  
TAE\_test\_pred3

## [1] 2 1 2 3 3 3 1 3 3 3 2 3 3 3 3 3 3 3 3 2 2 3 3 3 3 3 3 3 2 3 1 1 1 3 3  
## [36] 3 2 3 3 1 3 1 1 2 1 1 2 2 3 2 3 3 1 1 1 2 3 3 3 3 3 2 1 2 3 1 2 3 2 1  
## [71] 2 2 3 3 1 2  
## Levels: 1 2 3

CrossTable(TAE\_test\_labels,TAE\_test\_pred3,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 76   
##   
##   
## | TAE\_test\_pred3   
## TAE\_test\_labels | 1 | 2 | 3 | Row Total |   
## ----------------|-----------|-----------|-----------|-----------|  
## 1 | 10 | 10 | 10 | 30 |   
## | 0.333 | 0.333 | 0.333 | 0.395 |   
## | 0.588 | 0.526 | 0.250 | |   
## | 0.132 | 0.132 | 0.132 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## 2 | 5 | 4 | 13 | 22 |   
## | 0.227 | 0.182 | 0.591 | 0.289 |   
## | 0.294 | 0.211 | 0.325 | |   
## | 0.066 | 0.053 | 0.171 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## 3 | 2 | 5 | 17 | 24 |   
## | 0.083 | 0.208 | 0.708 | 0.316 |   
## | 0.118 | 0.263 | 0.425 | |   
## | 0.026 | 0.066 | 0.224 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## Column Total | 17 | 19 | 40 | 76 |   
## | 0.224 | 0.250 | 0.526 | |   
## ----------------|-----------|-----------|-----------|-----------|  
##   
##

cm3 <- table(TAE\_test\_labels,TAE\_test\_pred3)  
cm3

## TAE\_test\_pred3  
## TAE\_test\_labels 1 2 3  
## 1 10 10 10  
## 2 5 4 13  
## 3 2 5 17

k=12

TAE\_test\_pred4 <- knn(TAE\_train,TAE\_test,TAE\_train\_labels,k=12)  
TAE\_test\_pred4

## [1] 2 1 2 3 3 2 1 3 2 3 2 3 3 3 2 3 3 2 3 2 2 2 2 3 3 3 3 3 2 3 1 1 1 3 3  
## [36] 3 2 3 3 3 2 1 1 2 3 1 2 2 3 2 3 3 1 1 2 2 3 2 2 3 2 2 2 2 3 1 3 2 1 1  
## [71] 2 3 3 3 1 2  
## Levels: 1 2 3

CrossTable(TAE\_test\_labels,TAE\_test\_pred4,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 76   
##   
##   
## | TAE\_test\_pred4   
## TAE\_test\_labels | 1 | 2 | 3 | Row Total |   
## ----------------|-----------|-----------|-----------|-----------|  
## 1 | 9 | 9 | 12 | 30 |   
## | 0.300 | 0.300 | 0.400 | 0.395 |   
## | 0.643 | 0.310 | 0.364 | |   
## | 0.118 | 0.118 | 0.158 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## 2 | 3 | 11 | 8 | 22 |   
## | 0.136 | 0.500 | 0.364 | 0.289 |   
## | 0.214 | 0.379 | 0.242 | |   
## | 0.039 | 0.145 | 0.105 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## 3 | 2 | 9 | 13 | 24 |   
## | 0.083 | 0.375 | 0.542 | 0.316 |   
## | 0.143 | 0.310 | 0.394 | |   
## | 0.026 | 0.118 | 0.171 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## Column Total | 14 | 29 | 33 | 76 |   
## | 0.184 | 0.382 | 0.434 | |   
## ----------------|-----------|-----------|-----------|-----------|  
##   
##

cm4 <- table(TAE\_test\_labels,TAE\_test\_pred4)  
cm4

## TAE\_test\_pred4  
## TAE\_test\_labels 1 2 3  
## 1 9 9 12  
## 2 3 11 8  
## 3 2 9 13

k=15

TAE\_test\_pred5 <- knn(TAE\_train,TAE\_test,TAE\_train\_labels,k=15)  
TAE\_test\_pred5

## [1] 2 1 2 3 3 2 1 3 3 3 2 3 3 2 2 3 3 2 3 2 2 3 3 2 1 2 3 3 2 3 1 1 1 2 1  
## [36] 1 2 3 3 2 2 1 1 2 1 1 2 3 3 2 3 3 1 1 2 2 3 2 2 2 3 2 1 2 3 1 2 2 1 1  
## [71] 1 3 3 1 1 2  
## Levels: 1 2 3

CrossTable(TAE\_test\_labels,TAE\_test\_pred5,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 76   
##   
##   
## | TAE\_test\_pred5   
## TAE\_test\_labels | 1 | 2 | 3 | Row Total |   
## ----------------|-----------|-----------|-----------|-----------|  
## 1 | 14 | 11 | 5 | 30 |   
## | 0.467 | 0.367 | 0.167 | 0.395 |   
## | 0.667 | 0.379 | 0.192 | |   
## | 0.184 | 0.145 | 0.066 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## 2 | 5 | 10 | 7 | 22 |   
## | 0.227 | 0.455 | 0.318 | 0.289 |   
## | 0.238 | 0.345 | 0.269 | |   
## | 0.066 | 0.132 | 0.092 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## 3 | 2 | 8 | 14 | 24 |   
## | 0.083 | 0.333 | 0.583 | 0.316 |   
## | 0.095 | 0.276 | 0.538 | |   
## | 0.026 | 0.105 | 0.184 | |   
## ----------------|-----------|-----------|-----------|-----------|  
## Column Total | 21 | 29 | 26 | 76 |   
## | 0.276 | 0.382 | 0.342 | |   
## ----------------|-----------|-----------|-----------|-----------|  
##   
##

cm5 <- table(TAE\_test\_labels,TAE\_test\_pred5)  
cm5

## TAE\_test\_pred5  
## TAE\_test\_labels 1 2 3  
## 1 14 11 5  
## 2 5 10 7  
## 3 2 8 14

For the model I have created crosstables for different values of k.

But, the accuracy of the predicted values as to whether they match up with the known values is necessary.

Accuracy of model for k = 2 is 36.84%

Accuracy of model for k = 3 is 39.47%

Accuracy of model for k = 5 is 46.05%

Accuracy of model for k = 6 is 40.78%

Accuracy of model for k = 12 is 43.42%

Accuracy of model for k = 15 is 50% [Highest accuracy]