Question 4.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a clustering model would be appropriate. List some (up to 5) predictors that you might use.

Answer:

Categorize which group the real estate belongs to. Based on the group prediction, target different marketing customer will be more efficiency.

Predictors: Building Area, Building Rent, Capital Value, Distance from airport, Distance from harbor, Distance from city center, population nearby, residence education degree nearby, residence average salary nearby, etc.

Question 4.2

The iris data set iris.txt contains 150 data points, each with four predictor variables and one categorical response. The predictors are the width and length of the sepal and petal of flowers and the response is the type of flower. The data is available from the R library datasets and can be accessed with iris once the library is loaded. It is also available at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Iris ). The response values are only given to see how well a specific method performed and should not be used to build the model.

Use the R function kmeans to cluster the points as well as possible. Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type.

1. Take a look at the correlation between predictors. Based on the corr matrix, the only predictors with a rather weak relationship with other predictors is sqpal.Width. The rest of the predictors has at least 1 strong relationship with others. So we can roughly predict sqpal.Width won’t contribute too much to the clustering.

And I tried scale the data, but it decreased the accuracy rate. So I am going to use raw data for the rest of the exercise.

library(corrplot)

res <- cor(Iris\_test, method = c("pearson", "kendall", "spearman"))

corrplot(res, method = "number")



1. Use all combination of 4 predictors, and loop k from 1 to 15, check how the wss goes.

no\_predictor <- dim(Iris\_test)[2]

no\_predictor

df\_result<- data.frame(totwithinss = numeric(), totss = numeric(), combination = character(), kvalue = integer(), accuracy = numeric())

colnames(df\_result) <- c('totwithinss', 'totss', 'combination', 'kvalue', 'accuracy')

i <- 1

k\_max <- 15

while (i <= no\_predictor){

selected\_predictor <- combn(seq(1,no\_predictor), i)

j <- 1

while (j <= dim(selected\_predictor)[2]){

wss <- sapply(1:k\_max, function(k){kmeans(Iris\_test[, c(selected\_predictor[, j])], k, nstart = 50, iter.max = 15)$tot.withinss})

totss <- sapply(1:k\_max, function(k){kmeans(Iris\_test[, c(selected\_predictor[, j])], k, nstart = 50, iter.max = 15)$totss})

k\_list <- c

kk <- as.data.frame(wss)

kk$tots <- totss

kk$combination <- paste("predictor = ", paste(c(selected\_predictor[, j]), collapse = ","))

kk$kvalue <- seq(1, length(wss))

df\_result <- rbind(df\_result, kk)

rm(kk)

j <- j+1

}

i <- i+1

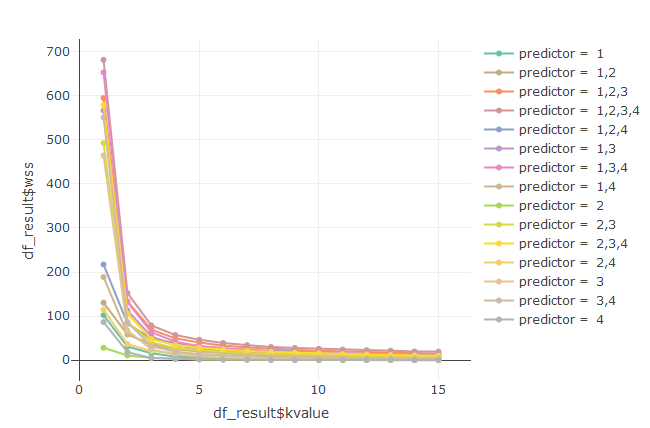
}

library(dplyr)

library(plotly)

plot\_ly(x=~df\_result$kvalue, y=~df\_result$wss, group=~df\_result$name, type="scatter",color=~df\_result$name, mode="lines+markers")

Note: 1 -> Sepal.Length, 2-> Swpal.Width, 3-> Petal.Length, 4-> Petal.Width



1. From step (b), we can see the margin stops changes a lot while k equal to 3 or 4, so here, we will choose the combination with wss\_ratio less than 12% to test the accuracy rate against the real category. Why I choose 12% is because only after 12%, wss\_ratio increase gap is larger than 1%.

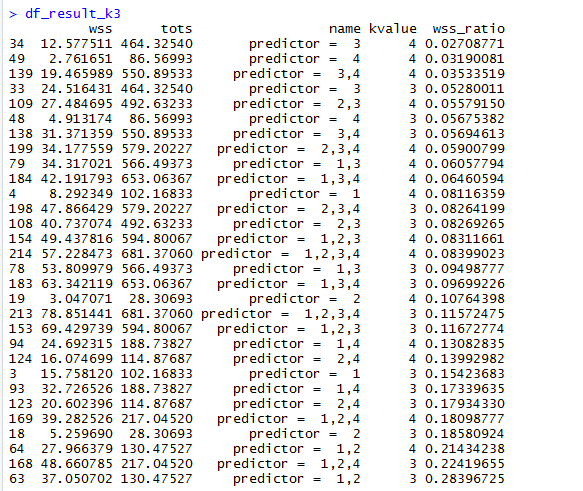
df\_result$wss\_ratio <- df\_result$wss/df\_result$tots

df\_result\_sort <- df\_result[order(df\_result$wss\_ratio), ]

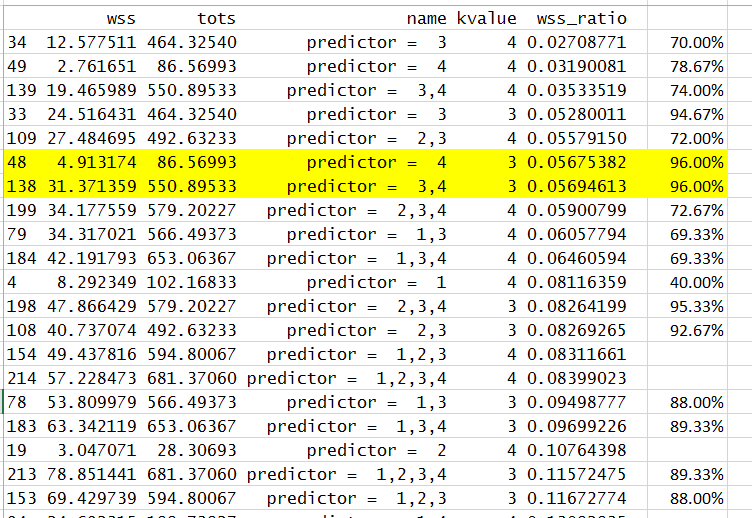
df\_result\_k3 <- df\_result\_sort[df\_result\_sort$kvalue==3|df\_result\_sort$kvalue==4, ]

df\_result\_k3





1. After calculate the accuracy rate (k=4 is always less than k=3). Best K value should be 3, and take correlation into consideration, best predictor combination should be (Petal.Length & Petal.Width).



Question 5.1

Using crime data from the file uscrime.txt (http://www.statsci.org/data/general/uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), test to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the grubbs.test function in the outliers package in R.

|  |  |
| --- | --- |
| Variable | Description |
| M | percentage of males aged 14–24 in total state population |
| So | indicator variable for a southern state |
| Ed | mean years of schooling of the population aged 25 years or over |
| Po1 | per capita expenditure on police protection in 1960 |
| Po2 | per capita expenditure on police protection in 1959 |
| LF | labour force participation rate of civilian urban males in the age-group 14-24 |
| M.F | number of males per 100 females |
| Pop | state population in 1960 in hundred thousands |
| NW | percentage of nonwhites in the population |
| U1 | unemployment rate of urban males 14–24 |
| U2 | unemployment rate of urban males 35–39 |
| Wealth | wealth: median value of transferable assets or family income |
| Ineq | income inequality: percentage of families earning below half the median income |
| Prob | probability of imprisonment: ratio of number of commitments to number of offenses |
| Time | average time in months served by offenders in state prisons before their first release |
| Crime | crime rate: number of offenses per 100,000 population in 1960 |

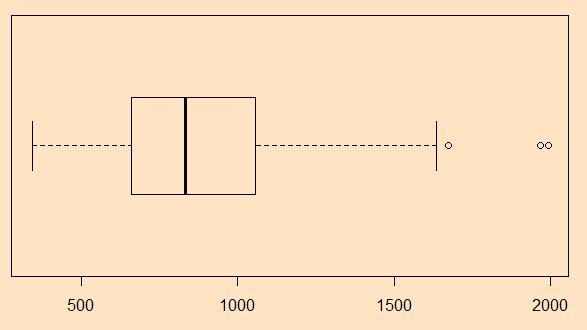
crime <- read.table("C:/Users/shero.chen/Desktop/NTU/GIT/Course/ISYE6501 - Introduction to Analytics Modeling/week2 - clustering/Temps.txt", stringsAsFactors = FALSE, header = TRUE)

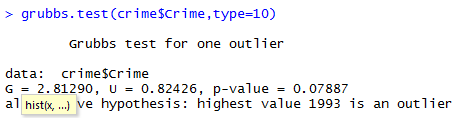
View(crime)

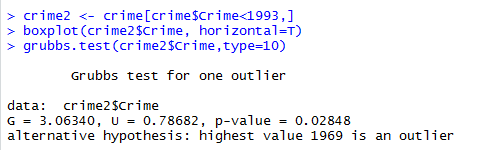
1. Use box plot to observe outlier in general. From box plot, there are 2 outstading outliers which is extremely large.

library(outliers)

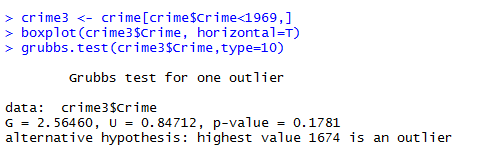
boxplot(crime$Crime, horizontal=T)



1. Use grubbs.test to validate the outlier. In 90% confidence level, value 1993 is outlier, but in 95% confidence level, it’s not. I am going to remove this record, and re-do the same test again.
2. After remove value 1993, from grubbs.text, 1969 has almost 98% one-tail confidence level to be sure it is an outlier. So I believe, both 1969 and 1993 should be outliers.



1. After remove 1969, re-do the test. There is no outlier.



Question 6.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a Change Detection model would be appropriate. Applying the CUSUM technique, how would you choose the critical value and the threshold?

In factory, if the machine temperature is too high, it means there is some problem with it. If ignore it, will cause damage and the cost is high amend fee and production line down. So we need to set up some temperature alarm of it. But if the alarm is too sensitive, will cause production line down frequently and require amenders to go and check, this is a cause too. When we set up the alarm, utility of CUSUM is one of popular methods. C is the value to exclude the noise, which means is a normal temperature of the functioning machine. And T is the threshold of the warning which we must inform maintenance.

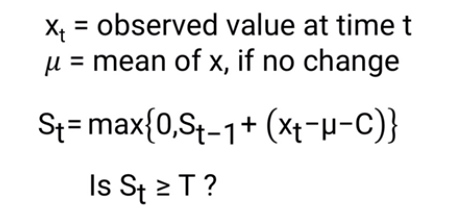
Question 6.2

1. Using July through October daily-high-temperature data for Atlanta for 1996 through 2015, use a CUSUM approach to identify when unofficial summer ends (i.e., when the weather starts cooling off) each year. You can get the data that you need from the file temps.txt or online, for example at http://www.iweathernet.com/atlanta-weather-records or https://www.wunderground.com/history/airport/KFTY/2015/7/1/CustomHistory.html . You can use R if you’d like, but it’s straightforward enough that an Excel spreadsheet can easily do the job too.

2. Use a CUSUM approach to make a judgment of whether Atlanta’s summer climate has gotten warmer in that time (and if so, when).

Answer:

I am using below formula to calculate cusum. If standard deviation is larger, then I will use a higher threshold (T) for the year. So different year, I have set C and T accordingly to predict the change. I have tried used only one set of C and T all the years, but sometimes, there is no change detection or sometimes too early detection. To find out more proper change of every year, I have adjusted the C and T accordingly every year.



Since I am using different T and C to detect change for different year accordingly, there would be bias if I evaluate the summer temperature only based on the changed temperature. Instead, I will use average temperature of the “official” summer period. In addition, I observed that in some years, there is a sudden big drop within the “official” summer period which we called it outlier, will be excluded from the average. And I got the result as below. So the summer temperature is getting warmer since 2010, and a drop in 2013, then climbs up again since 2014.

