Assignment 4

Chee Kay Cheong (cc4778)

2023-02-08

knitr::opts\_chunk$set(message = FALSE, warning = FALSE)  
  
library(tidyverse)  
library(stats)  
library(factoextra)  
library(cluster)  
library(caret)  
library(modelr)  
  
set.seed(123)

# Part I: Implementing a Simple Prediction Pipeline

The New York City Department of Health administered a questionnaire on general health and physical activity among residents. Using the dataset class4\_p1.csv, fit and evaluate two prediction models using linear regression. The aim of the models are to predict the number of days in a month an individual reported having good physical health (feature name: healthydays).

### Step 1: Load and clean dataset

class4 = read\_csv("./Data/class4\_p1.csv") %>%   
 janitor::clean\_names() %>%   
 select(-x1) %>%   
 mutate(  
 chronic1 = as\_factor(chronic1),  
 chronic3 = as\_factor(chronic3),  
 chronic4 = as\_factor(chronic4),  
 tobacco1 = as\_factor(tobacco1),  
 alcohol1 = as\_factor(alcohol1),  
 habits5 = as\_factor(habits5),  
 habits7 = as\_factor(habits7),  
 agegroup = as\_factor(agegroup),  
 dem3 = as\_factor(dem3),  
 dem4 = as\_factor(dem4),  
 dem8 = as\_factor(dem8),  
 povertygroup = as\_factor(povertygroup)) %>%   
 drop\_na()

### Step 2: Partition data into training and testing (use a 70/30 split)

set.seed(123)  
  
train.index = createDataPartition(class4$healthydays, p = 0.7, list = FALSE)  
  
class4\_train = class4[train.index, ]  
class4\_test = class4[-train.index, ]

## Problem 1

Fit two prediction models using different subsets of the features in the training data.

* **Model 1**
  + Outcome: healthydays
  + Predictors: chronic4, gpaq8totmin, gpaq11days, habits5, habits7 & agegroup
* **Model 2**
  + Outcome: healthydays
  + Predictors: bmi, tobacco1, alcohol1, habits5 & habits7

model\_1 = lm(healthydays ~ chronic4 + gpaq8totmin + gpaq11days + habits5 + habits7 + agegroup, data = class4\_train)  
  
model\_2 = lm(healthydays ~ bmi + tobacco1 + alcohol1 + habits5 + habits7, data = class4\_train)

## Problem 2

Apply both models within the test data and determine which model is the preferred prediction model using the appropriate evaluation metric(s).

We will be using **Root Mean Square Error (RMSE)** as the appropriate evaluation metric to compare how well the model is predicting against the actual values.

rmse(model\_1, class4\_test)

## [1] 7.218366

rmse(model\_2, class4\_test)

## [1] 7.401421

Based on the result, it appears that *Model 1* is a better model in predicting the number of days in a month an individual reported having good physical health because it has a lower RMSE value compared to Model 2.

## Problem 3

One setting where the implementation of *Model 1* would be useful is when we hope to predict a person’s overall perceived health in a senior community.

# Part II: Conducting an Unsupervised Analysis

Using the dataset from the Group assignment Part 3 (USArrests), identify clusters using hierarchical analysis. Use an agglomerative algorithm for hierarchical clustering. Use a Euclidian distance measure to construct your dissimilarity matrix.

### Step 1: Load dataset & prepare for analysis

data("USArrests")  
# Checked no missing data.  
  
# Check means and SDs to determine if scaling is necessary:  
colMeans(USArrests, na.rm = TRUE)

## Murder Assault UrbanPop Rape   
## 7.788 170.760 65.540 21.232

apply(USArrests, 2, sd, na.rm = TRUE)

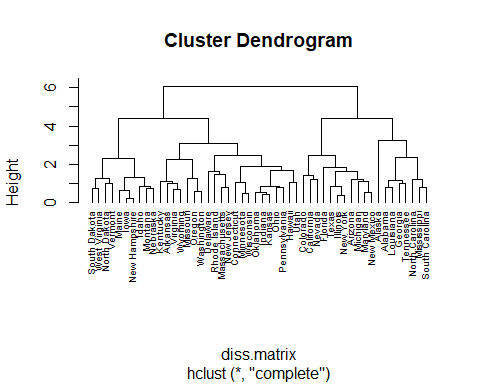
## Murder Assault UrbanPop Rape   
## 4.355510 83.337661 14.474763 9.366385

# Means and standard deviations are very different from each other. Scaling is needed.  
US\_Arrests = scale(USArrests)

## Problem 4

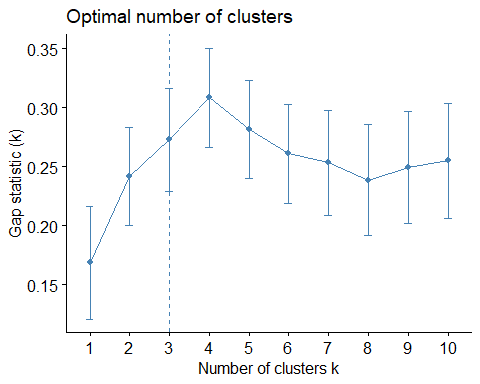
We will be conducting a hierarchical clustering analysis using the ***complete*** linkage.

# Create Dissimilarity matrix  
diss.matrix = dist(US\_Arrests, method = "euclidean")  
  
# Hierarchical clustering using Complete Linkage  
clusters.h = hclust(diss.matrix, method = "complete" )  
  
# Plot the obtained dendrogram  
plot(clusters.h, cex = 0.6, hang = -1)



#### a) Determine the optimal number of clusters using ***gap-statistic*** analysis.

gap\_stat = clusGap(US\_Arrests, FUN = hcut, K.max = 10, B = 50)  
fviz\_gap\_stat(gap\_stat)



Based on the result, the optimal number of clusters is *3*.

#### b) Describe the composition of each cluster in terms of the original input features

clusters = kmeans(US\_Arrests, 3, nstart = 25)  
  
clusters

## K-means clustering with 3 clusters of sizes 20, 13, 17  
##   
## Cluster means:  
## Murder Assault UrbanPop Rape  
## 1 1.0049340 1.0138274 0.1975853 0.8469650  
## 2 -0.9615407 -1.1066010 -0.9301069 -0.9667633  
## 3 -0.4469795 -0.3465138 0.4788049 -0.2571398  
##   
## Clustering vector:  
## Alabama Alaska Arizona Arkansas California   
## 1 1 1 3 1   
## Colorado Connecticut Delaware Florida Georgia   
## 1 3 3 1 1   
## Hawaii Idaho Illinois Indiana Iowa   
## 3 2 1 3 2   
## Kansas Kentucky Louisiana Maine Maryland   
## 3 2 1 2 1   
## Massachusetts Michigan Minnesota Mississippi Missouri   
## 3 1 2 1 1   
## Montana Nebraska Nevada New Hampshire New Jersey   
## 2 2 1 2 3   
## New Mexico New York North Carolina North Dakota Ohio   
## 1 1 1 2 3   
## Oklahoma Oregon Pennsylvania Rhode Island South Carolina   
## 3 3 3 3 1   
## South Dakota Tennessee Texas Utah Vermont   
## 2 1 1 3 2   
## Virginia Washington West Virginia Wisconsin Wyoming   
## 3 3 2 2 3   
##   
## Within cluster sum of squares by cluster:  
## [1] 46.74796 11.95246 19.62285  
## (between\_SS / total\_SS = 60.0 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

*Cluster 1* has a medium urban population size, but they have the greatest rate in murder, assault, and rape because all the values for each features are positive (>0) and are the highest among all three clusters. *Cluster 2* has the smallest proportion of urban population, as well as the lowest crime rate because all the values for each features are negative (<0) and are the smallest among all three clusters. *Cluster 3* has the biggest urban population. Its crime rate (murder, assault, rape) is not the lowest but still lower than average.

## Problem 5

One research question that can be addressed using the newly identified clusters is: What are some contributing elements to the high crime rate in the Cluster 1 states?

Before using these clusters for the above research question, we should be careful for not including the state name in the cluster to avoid defaming certain states.