Clustering

In this assignment you will complete a variety of tasks related to binary classification with neural networks. The dataset that we will be using is related to criminal justice and deals specifically with parole violations.

Deliverable: All of your work for this assignment should be done in an R Markdown document. Knit your document into a Word file and submit the Word file as the deliverable for this assignment.

Libraries: For this assignment you will need the following libraries: tidyverse, cluster, factoextend, and dendextend.

Before beginning the assignment tasks, you should read-in the data for the assignment into a data frame called "trucks". In this dataset, Driver_ID is a unique identifer for each delivery driver, Distance is the average mileage driven by each driver in a day, and Speeding is the percentage of the driver's time in which he is driving at least 5 miles per hour over the speed limit.

Task 1: Plot the relationship between Distance and Speeding. Describe this relationship. Does there appear to be any natural clustering of drivers?

Task 2: Create a new data frame (called trucks2) that excludes the Driver_ID variable and includes scaled versions of the Distance and Speeding variables. **NOTE: Wrap the scale(trucks2) command in an as.data.frame command to ensure that the resulting object is a data frame. By default, scale converts data frames to lists**

Task 3 Use k-Means clustering with two clusters (k=2) to cluster the trucks2 data frame. Use a random number seed of 1234. Visualize the clusters using the fviz_cluster function. Comment on the clusters.

Task 4: Use the two methods from the k-Means lecture to identify the optimal number of clusters. Use a random number seed of 123 for these methods. Is there consensus between these two methods as the optimal number of clusters?

Task 5: Use the optimal number of clusters that you identified in Task 4 to create k-Means clusters. Use a random number seed of 1234. Use the fviz_cluster function to visualize the clusters.

Task 6: In words, how would you characterize the clusters you created in Task 5?

Before starting Task 7, read in the "wineprice.csv" file into a data frame called wine. This is a small dataset containing wine characteristics and the price of wine at auction. WinterRain refers to the amount of rain received in winter, AGST refers to the average growing season temperature, HarvestRain refers to the amount of rain received in the harvest season, Age refers to the age of the wine when sold at auction, and FrancePop refers to the population of France at the time the wine was sold at auction.

Create a new data frame called wine2 that removes the Year and FrancePop variables and scales the other variables.

Task 7: Use the two methods from Task 4 to determine the optimal number of k-Means clusters for this data. Use a random number seed of 123. Is there consensus between these two methods as the optimal number of clusters?

Task 8: Use the optimal number of clusters that you identified in Task 4 to create k-Means clusters. Use a random number seed of 1234. Use the fviz_cluster function to visualize the clusters.

Task 9: Use agglomerative clustering to develop a dendogram for the scaled wine data. Follow the same process from the lecture where we used a custom function to identify the distance metric that maximizes the "agglomerative coefficient". Plot the dendogram.

Task 10: Repeat Task 9, but with divisive clustering.