ITM 6285 Data Mining Lab - Clustering

K - Means Clustering (expected time - 0.75 hours)

Learning objective:

1. Perform K-means clustering in R

2. Interpret the R outputs

Task 1: Import the dataset

Import the data called "Car_Identification.csv", remember to change the Category field as factors. First, take a look at the data. The four features (acceleration, body roll, down force, and braking) are provided for each car. The category column shows the real car brand.

Then do a summary of the data set using function "summary". Short answer what are the median of the acceleration, body roll, down force, and braking. How many categories do you see, and how many cars are there in each category?

Task 2: Standardize the data

Before doing standardization, remember to **remove the category column**, because the "kmeans" function takes each column as one feature by design.

You may use DataSetName$Category=NULL to remove the category column

Using the "scale" function to standardize the data. Answer why we need to standardize the data?

Task 3: K-Means clustering

Do a K-Means clustering with 3 clusters, and check the output. Then run the K-Means clustering again, and check the output again. Are the two k-means clustering outputs the same? Why?

Copy the outputs from your last run on your answer sheet. Based on your latest run, what is the average acceleration for each of the three clusters? Which cluster does the last data point (150^{th} car) belong to?