DAMI 2019

Homework Assignment 2

Unsupervised Learning – Clustering

These instructions should be used together with the provided assignment R code. To complete the assignment, you can choose either to write the code by yourself completely, or to use the provided R code and fill in the missing parts. In this assignment, you need to use the synthetic data that has already been generated and you can download it (***cluster.RData***) from the course website. ONLY submit the complete R code (.R file). Screen shots, plots, word or pdf documents sent will not be graded.

Part 1: Distance Calculation

The Euclidean distance between two vectors, *a* and *b* of length *n* can be calculated as

$$d_{euc}(a,b) = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\cdots+(a_n-b_n)^2} = \sqrt{\sum_{i=1}^{n}(a_i-b_i)^2}$$

The Manhattan distance between these two vectors can be calculated as

$$d_{man}(a,b) = |a_1-b_1| + |a_2-b_2| + \cdots + |a_n-b_n| = \sum_{i=1}^{n}|a_i-b_i|$$

[YOUR TASK] *Write a function that can compute Euclidean and Manhattan distance between any two given points (use aforementioned equations for computing distances). For example the Euclidean distance and Manhattan distance between two points (4.2, 7.1) and (3.1, 5.8) should be 1.702939 and 2.4 respectively.* [4 POINTS]

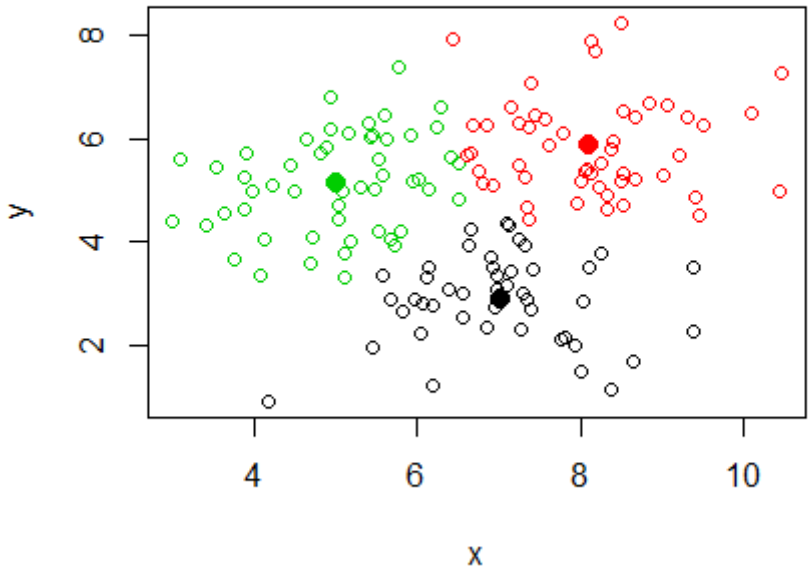
Part 2: K-Means Clustering

K-means clustering is a simple approach to separate a dataset into k unique and non-overlapping clusters. To perform k-means clustering, we must first specify the number of clusters, k; then the algorithm will assign each observation to one of the k clusters.

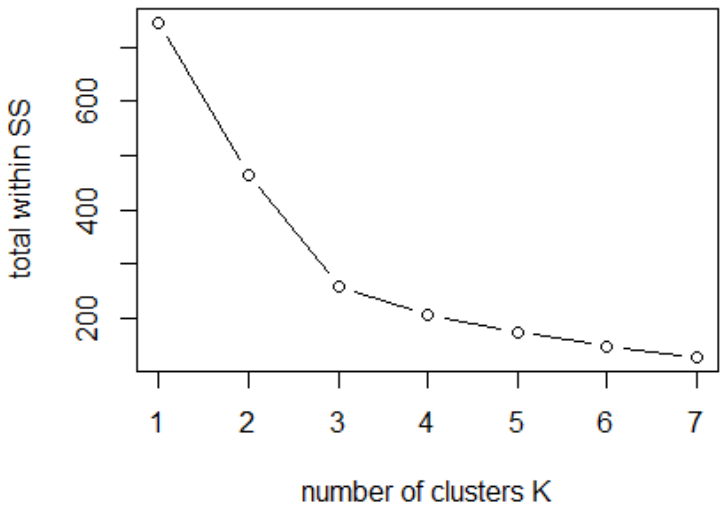
Algorithm of k-means clustering

1. Randomly assign a number between 1 to k to each of the observations. This serves as initial random cluster assignment for the observations.
2. Iterate until the cluster assignments stop changing:
    a. For each of k clusters, compute the cluster centroid. The kth cluster centroid is the mean vector of observations in this cluster.
    b. Assign each observation to the cluster whose centroid is closest. Euclidean distance is used to measure the distance.

[YOUR TASK 2a] *Write your own function in R to perform K-Means clustering on a given dataset. Use this function to cluster the given dataset "cluster.RData" into 3 clusters (K=3). The final plot should look similar to the one shown below.* [6 POINTS]



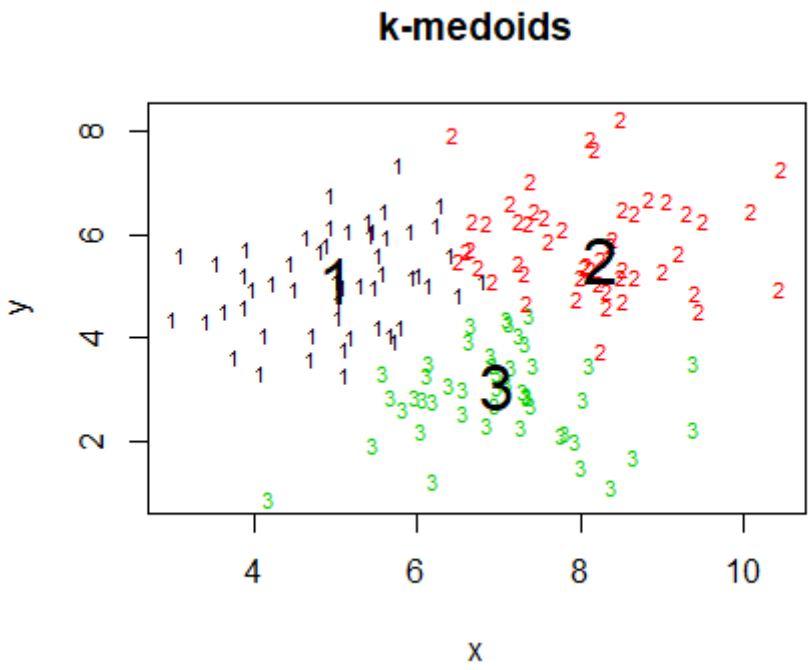
[YOUR TASK 2b] *Use kmeans() function in base R library to perform K-Means clustering for this part of task. Make a plot of number of clusters (K) vs. Total within Sum of Squares among clusters for different values of K ranging from 1 to 7. Determine the best value of K from the plot. The plot should look similar to the one shown below.* [5 POINTS]



Part 3: K-Medoid clustering

The k-medoid clustering is a more robust clustering algorithm compared to k-means. The most common realization of k-medoid clustering is the Partitioning Around Medoids (PAM) algorithm. In R, the function *pam()* in the "cluster" package implements the algorithm. Install the package and read the help file for more details on how to use *pam()* function.

[YOUR TASK 3a] *Install package "cluster" and use pam() to perform k-medoids with value of K set to 3. Use the synthetic dataset that has been provided. Use Manhattan distance as a metric. Plot the points and color them according to the cluster they are assigned to. Locate the centroids in the plot. The plot should look similar to the one shown below*. [4 POINTS]

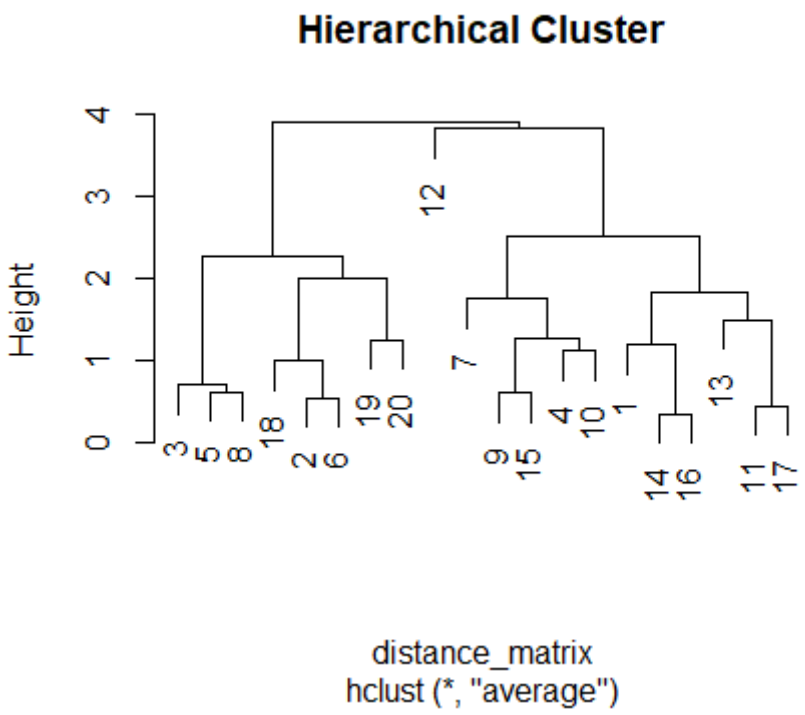


[YOUR TASK 3b] *Find the Silhouette value of the kmedoid clustering performed by assigning value of K as 2, 3, 4 and 5 in pam() and determine the best value of K. Follow the R codes for hints.* [3 POINTS]

Part 4: Hierarchical Clustering

One potential disadvantage of k-means clustering is that number of clusters, K need to be specified beforehand. Alternatively, hierarchical clustering does not require this and it results in a tree like representation of the observations, dendrogram.

[YOUR TASK 4a] *Perform a hierarchical clustering on smaller subset of the given synthetic data. Only 20 observations are randomly selected for this task. Use Euclidean distance metric and average-link approach for computing inter-cluster distance. Plot a dendrogram for this clustering which should look similar to the one shown below.* [4 POINTS]



[YOUR TASK 4b] *Cut the dendrogram tree obtained in 4a to get three clusters. Plot the points and color them according to the clusters they belong to. The final plot should look similar to the one shown below.* [4 POINTS]