Clustering (Instruction manual)

CIS 400/600 Fundamentals of Data and Knowledge Mining

1. Installing `clustertend` package and loading the same

   ```
   > install.packages("clustertend")
   > library(clustertend)
   ```

2. Reading `WINE` dataset and performing **Hopkins Statistic** with 10% sample

   ```
   > wine <- read.csv("wines.csv")
   > hopkins(wine, 0.1 * nrow(wine))
   ```

3. Shuffle data as follows

   ```
   > wine <- wine[sample(nrow(wine)), ]
   ```

4. Running `K-Means` algorithm with `K = 5`

   ```
   > kmeans_5 <- kmeans(wine, 5)
   ```

   Attributes under `kmeans()` method can be explored using

   ```
   > attribute(kmeans_5)
   ```

   Some of these attributes include

   (a) `cluster`

   (b) `centers`

(c) `totss`

(d) `tot.withinss`

5. Printing confusion matrix for K-Means clustering with `K = 5` is as follows

```
> table(wine[,1], kmeans_5$cluster)
```

6. Hierarchical clustering with **Euclidean distance** as distance metric and **Single - Link** as cluster proximity measure

```
> d <- dist(wine_stand, method = "euclidean")
> hier <- hclust(d, method = "single")
```

Plotting dendrogram and cutting it to 4 clusters is as follows

```
> plot(hier)
> clusters <- cutree(hier, k = 4)
> rect.hclust(hier, k = 4, border = "red")
```

Printing confusion matrix for hierarchical clustering

```
> table(wine[,1], clusters)
```