QMST 3339 Homework 2

Tahir Ekin[1]

[1]McCoy College of Business, Texas State University

Spring 2018

Guidelines

1. A hard copy (print-out) of your report that summarizes the results and includes the supporting graphs is due to the beginning of the lecture on 2/28/2018, Wednesday.
2. Do not include data in your report.
3. This is a group-based homework. However, all team members are expected to understand the assignment, relevant script and the submission thoroughly.
4. This is graded out of 5 points, and accounts for 5% of your overall grade. See the syllabus for more details.
5. Submission of any work, for which unauthorized help has been received, is termed as academic dishonesty and will be grounds for a failing grade in the course. See "Academic Honesty" section of the syllabus for more details.

1 Partitional Clustering

A dataset containing 13 chemical measurements on 178 Italian wine samples is of interest. The data originally come from the UCI Machine Learning Repository but we will access it via the rattle package. Be aware that installing this package may take up to a minute. Then you can use *data(wine, package="rattle")* command to access the data in R. More information about this dataset can be retrieved by using *?wine* command. The attributes are:

1) Type
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10) Color intensity
11) Hue
12) OD280/OD315 of diluted wines
13) Proline

1. Determine the numerical variables and scale them using the default scale() function in R. 0.6 pts

2. Using the scaled data set, utilize k-means clustering analysis for 25 randomly selected starting points, with initial seed as 1234. Answer the questions below. 2.4 pts, 0.6 pts each.

   (a) Using WSS method, determine the number of clusters.

(b) Report the within cluster sum of squares by cluster. What is the ratio of between sum of squares and total sum of squares? Comment on its meaning, does it correspond to having high quality clusters?

(c) What is the size of each cluster? Share the centroid means of each cluster.

(d) Report the counts of wine types in each cluster. What type of wine do you see most in the first cluster?

2 Hierarchical Clustering

We will use the *iris* dataset from the *datasets* library. You can use *library(datasets); data(iris);* to access the data set in R. This famous (Fisher's (1936) or Anderson's (1935)) *iris* data set provides the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 150 flowers (50 flowers from each of the 3 species of iris). The species are Iris setosa, versicolor, and virginica. *iris* is a data frame with 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.

We are interested in a sample of 40 records from the iris data. We aim to utilize hierarchical clustering with the variables of Sepal.Length, Sepal.Width, Petal.Length, Petal.Width.

1. Share the R code to draw 40 randomly selected records from the iris data set that only includes the numerical variables of interest. Use the seed as 123. 0.5 pts

2. Using the Ward' method, create a dendrogram. Share the dendrogram that is labelled with respect to species. 1 pt

3. Share the plot of dendrogram with three clusters. 0.5 pts