Ray Beecham
Cory Singer
JP Fernandez
Athena Mills
Cade Northcutt QMST 3339 February 24, 2018

Homework #2

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

sapply(wine, is.numeric) # Which Variables are numeric using logical operator. Type is False.

scWine <- scale(wine[-1]) # I am scaling the variables minus the first variable 'Type' since it is not a part of the numerical variables.

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.

Ray Beecham
Cory Singer
JP Fernandez
Athena Mills
Cade Northcutt QMST 3339 February 24, 2018

Using the Within cluster sum of squares method it was determined that there are 15 clusters but looking at the plot the best number of clusters is 3.

(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?

Cluster 1, 2, 3

```
385.6983 558.6971 326.3537
```

I would say according to the 44.8% the quality of clusters is average or does not correspond to having high quality clusters primary because being closer to 100% is the optimum quality

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

The Size of each cluster is: 62, 65, 51

```
  cluster  Alcohol    Malic      Ash Alcalinity Magnesium  Phenols Flavanoids Nonflavanoids P
roanthocyanins    Color
1       1 13.67677 1.997903 2.466290   17.46290 107.96774 2.847581  3.0032258     0.2920968
1.922097 5.453548
2       2 12.25092 1.897385 2.231231   20.06308  92.73846 2.247692  2.0500000     0.3576923
1.624154 2.973077
3       3 13.13412 3.307255 2.417647   21.24118  98.66667 1.683922  0.8188235     0.4519608
1.145882 7.234706
        Hue Dilution   Proline
1 1.0654839 3.163387 1100.2258
2 1.0627077 2.803385  510.1692
3 0.6919608 1.696667  619.0588
```

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?

```
    1  2  3
  1 59  3  0
  2  0 65  0
```

Judging by the data we are seeing 59 the most in the first cluster out of 62. (59 + 2)

2 Hierarchical Clustering

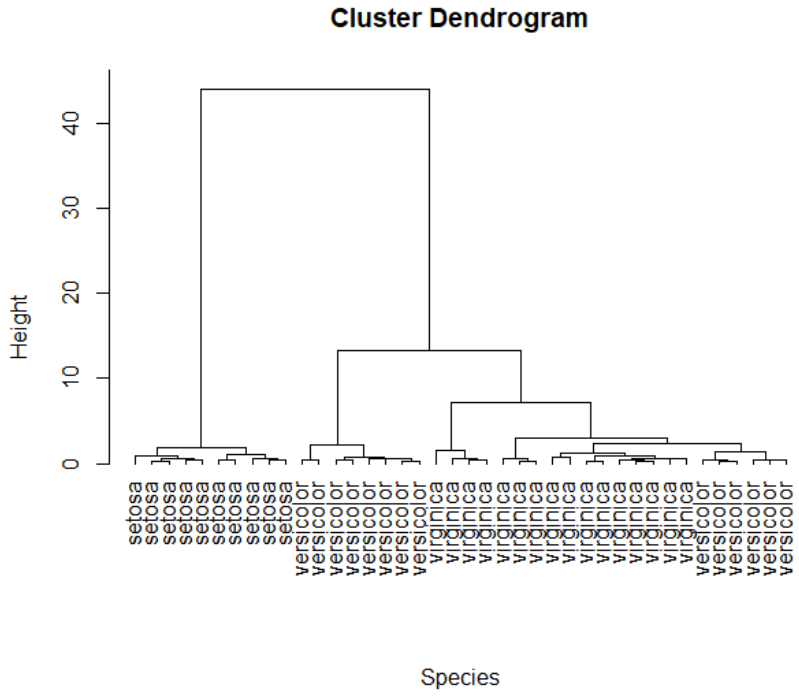
We will use the iris dataset from the datasets library. You can use library(datasets); data(iris); to access the data set in R. This famous (Fisher's (1936) or Anderson's (1935)) iris data set provides the measurements in centimeters of the variables sepal length and width and petal length and width, 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.

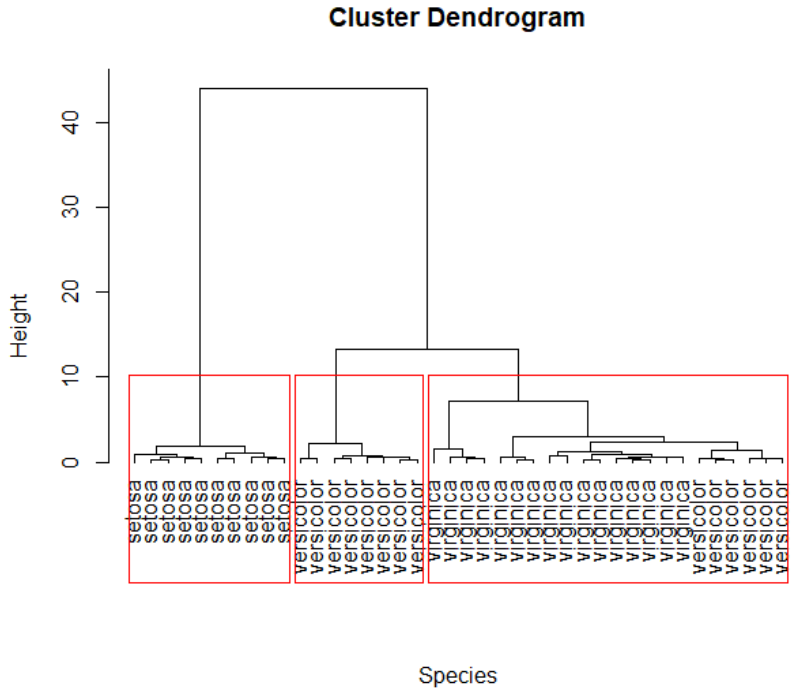
```
set.seed(123)
myD <- sample(1:dim(iris)[1], 40)
with40 <- iris[myD,1:5]
comp40 <- iris[myD,1:4]
```

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

Ray Beecham
Cory Singer
JP Fernandez
Athena Mills
Cade Northcutt



3. Share the plot of dendrogram with three clusters.



Ray Beecham
Cory Singer
JP Fernandez
Athena Mills
Cade Northcutt

QMST 3339

February 24, 2018