CS3002: Clustering Lab - ASSESSED

This worksheet must be assessed by a GTA. All assessed exercises must be completed and checked by a GTA in the lab. This is PASS / FAIL and all assessed sheets must be passed in order to pass the coursework for this module. You can make multiple attempts but do not ask to be assessed until you are ready.

In this lab you will be exploring the use of various R commands to investigate different clustering methods applied to a dataset.

Datasets:

- Spaeth Simulated Data – this is simple 2 dimensional data to help plot clusters.
- Seeds Data - This dataset comprises of measurements taken from three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment. The 7 features are:
    1. Area
    2. Perimeter
    3. Compactness
    4. Length of kernel
    5. Width of kernel
    6. Asymmetry coefficient
    7. Length of kernel groove

    Taken from paper: M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, S. Zak, 'A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images', in: Information Technologies in Biomedicine, Ewa Pietka, Jacek Kawa (eds.), Springer-Verlag, Berlin-Heidelberg, 2010, pp. 15-24.

R Functions:

- `read.csv`
- `plot`
- `dist`
- `hclust`
- `cutree`
- `kmeans`
- `WK_R (User defined)`

Hierarchical Clustering

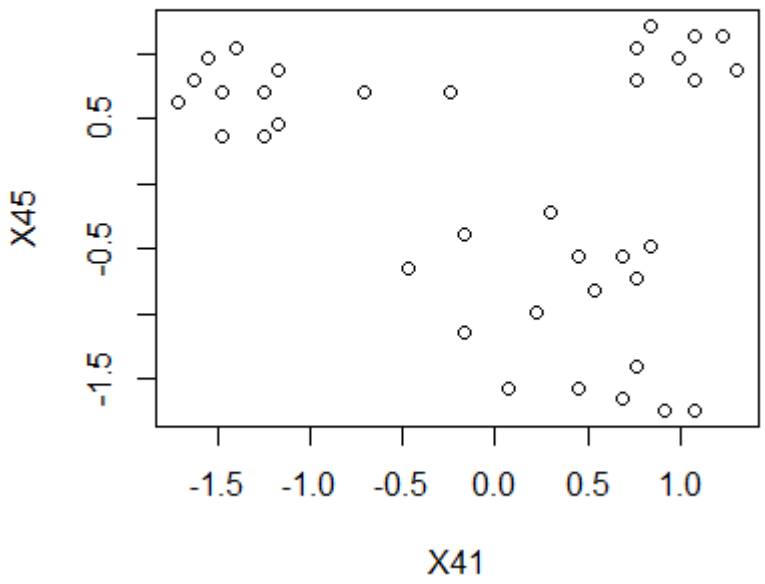
Download the *Spaeth_01.csv* data from blackboard and save it to your home directory. This is a toy datasets with two variables and 37 datapoints.

Type in the following with the correct pathname to load in the data to R:

```
mydata = read.csv('H:\\spaeth_01.csv', sep=",")
```

Then plot it with:

```
plot(mydata)
```



Now do some pre-processing on the data:

```
# optional: Prepare Data
mydata = na.omit(mydata) # deletion of missing data
mydata = scale(mydata) # standardize variables
```

Before performing Hierarchical Clustering, we first generate a matrix of Euclidean distances between each datapoint:

```
d <- dist(mydata, method = "euclidean") # distance matrix
```

and then use `hclust` to perform the clustering (here we use average link clustering) and plot to see the dendrogram:

```
fit <- hclust(d, method="average")
plot(fit) # display dendogram
```

We can now create the clusters by cutting this dendrogram

```
Hgroups <- cutree(fit, k=5) # cut tree into 5 clusters
```

We can also draw the dendrogram with red borders around these 5 clusters

```
rect.hclust(fit, k=5, border="red")
```

We can also draw a scatterplot with the assigned clusters as colours

```
plot(mydata, col=Hgroups)
```

- Do the clusters look reasonable? What happens if you change the number of clusters?

  Try some other forms of hierarchical by using the `hclust` command but replacing 'average' with 'complete' or 'single'

- Try loading in some of the other "spaeth" datasets and see how they cluster

K-means Clustering

Now we can use the K-means clustering technique to generate our cluster allocations. We use the kmeans command with k = 5.

```
fit <- kmeans(mydata, 5) # 5 cluster solution
```

We can get different statistics on the clusters e.g. the mean value for each variable:

```
aggregate(mydata,by=list(fit$cluster),FUN=mean)
```

We get the assignments by using the $ command:

```
Kgroups = fit$cluster
```

And we can draw the scatterplot as we did with hierarchical:

```
plot(mydata, col=Kgroups)
```

In order to use Weighted Kappa (WK) we can load in some R code from blackboard. Download "WK_R.r" and place it in your H drive then type:

```
source("H:\\WK_R.r")
```

WK takes in two clusterings and returns the Weighted Kappa between them so to calculate the WK value between the hierarchical and k-means methods use:

```
wk = WK_R(Kgroups, Hgroups)
```

ASSESSED EXERCISE:

Using the Weighted Kappa function, WK_R.r, in blackboard, explore how the hierarchical clustering compares to k means for the seeds dataset "seeds_dataset.csv" given the correct clusters in "seeds_real.csv".

You need to:

- Calculate the WK for K means clustering with different values of K
- Try Hierarchical with different linkage measures: single, complete and average
- Use scatterplots to illustrate the different clusterings.

- Plot the dendrograms where necessary.
- Plot the different Weighted Kappa values on an appropriate graph.

Association Rule Mining

```
#ASSOCIATION RULES

#Adapted from https://towardsdatascience.com/association-rule-mining-in-r-
ddf2d044ae50

install.packages("arules")
install.packages("arulesViz")
library(arules)

#Read in Groceries data

data('Groceries')
#inspect first 3 transactions
inspect(head(Groceries, 3))

#Apriori Algorithm - generte rules

grocery_rules <- apriori(Groceries, parameter = list(support = 0.01, confi
dence = 0.5))
#inspect first 3 sorted by confidence
inspect(head(sort(grocery_rules, by = 'confidence'), 3))
```

SAMPLE EXAM QUESTIONS:

Q1) Given the following three attributes:

ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Age	23	57	18	32	26	71	66	43	52	18
Blood Pressure	95	150	85	110	100	80	100	105	95	90
BMI	25	20	27	23	18	30	27	25	35	29

Choose any 2 patients and calculate:

a) What is the Euclidean distance between each of them?
b) What is the Manhattan distance between each of them?

Q2) Briefly describe three different forms of hierarchical clustering methods

Q3) Describe two clustering methods and their advantages / disadvantages