ADELEKE ADEMUYIWA 02073331
MATH 522 ADVANCED MATHEMATICAL STATISTICS

PROJECT 5

**Title, Author, Date.**

An Introduction to Statistical Learning.

Gareth James. Daniela Witten. Trevor Hastie. Robert Tibshirani.

## Description of data used

K-Means Clustering. Function kmeans () solves k-means clustering in R. An example is a simple simulation of two clusters in data.

## Abstract.

To perform kmeans () function in R studio with various initial cluster assignments, nstart argument is used. If value of nstart greater than one is used. Here compared using nstart = 1 to nstart = 20.

```
set.seed (4)
km.out <- kmeans (x, 3, nstart = 1)
km.out$tot.withinss
[1] 104.3319
Km.out <- kmeans (x, 3, nstart = 20)
Km.out$tot.withinss
[1] 97.9793
```

Km.out$tot.withinss is total within cluster squares which is seeked to minimize by performing k-means clustering. Individual within cluster sum of squares contained in vector km.out$withinss. Run k-means clustering with a large value of nstart, 50 or 20. Hence an undesirable local optimum is obtained. When k-means clustering, using various initial cluster assignments. It is important to set a random seed utilizing set.seed () function. Like this, initial cluster assignments are replicated, k-means output is fully reproducible.

## Introduction & Background.

To perform k-means clustering, using various initial cluster assignments. It is vital to set a random seed utilizing set.seed () function. In this way, initial cluster assignments in step 1 is replicated, k-means output is reproducible. Clustering in Hierarchy. hclust () function implements hierarchical clustering in R. The data from previous lab to plot hierarchical clustering dendrogram utilizing complete, single, and average linkage clustering, with Euclidean distance as dissimilarity measure. Start by clustering observations utilizing complete linkage. dist () function is utilized to compute 50 x 50 inter observation Euclidean distance matrix.

```
hc.complete <- hclust (dist (x), method = “complete”)
```

This could easily perform hierarchical clustering with average or single linkage instead.

```
hc.complete <- hclust (dist (x), method = “average”)
hc.single <- hclust (dist (x), method = “single”)
```

**Does the background and significance have a logical organization?** Does it move from the general to the specific? The background of k-means clustering function kmeans () solves kmeans in R studio in the simple simulation of 2 clusters in data. K-means clustering moving from the beginning of nstart argument is utilized. nstart value is greater than one is utilized. Comparison of nstart = 1 to nstart = 20 output is 97.9793. The specific is to set a random seed set.seed() function. Step 1 is repliacated, then k-means output is reproducible. 50 x 50 inter observation Euclidean distance matrix is provided.

**Has sufficient background been provided to understand the report?** Yes, these results obtained are performed in hierarchy on the full data set. Often performing clustering on first few principal score vectors give better results than the output on full data. Principal component step as one of signifying data. K-means clustering in first few principal component score vectors rather than full data set.

**Does this section end with statements about the goals of the report?** No. It ended with exercises of using hierarchical clustering, scaling the variables, involvement of k-means clustering algorithm.

```
hc.complete <- hclust (dist (x), method = “complete”)
```

<u>Methods.</u>

Now plot dendrograms obtained using usual plot () function. Numbers at bottom of plot identify each observation.

```
par (mfrow = c (1, 3) )
plot (hc.complete, main = “complete linkage”,
          xlab = “ ”, sub = “ “, cex = .9)
plot (hc.average, main = “average linkage”,
          xlab = “ “, sub =” “, cex = .9)
plot (hc.single, main = “single linkage”,
          xlab = “ “, sub = “ “, cex = .9)
```

Labels determined for each observation associated with a given cut of dendrogram, we use cutree () function.

```
cutree (hc.complete, 2)
[1]    1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2
[ 30 ] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

cutree (hc.average, 2)
[1]   1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2
[30] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

cutree (hc.single, 2)
[1]  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[30] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Assessment. Could the study be repeated based on the information given here? Is the material organized into logically and clearly with technical terms clearly explained? Yes. It is organized logically and technical terms explained in the numbers at bottom of plot cex = .9 identifies each observation of

```
plot(hc.complete, main = “complete linkage”
          xlab = “ “, sub = “ “, cex = .9
plot(hc.average, main = “ “, average linkage”,
          xlab “ “, sub=” “, cex = .9
plot(hc.single, main = “sungle linkage”,
          xlab = “ “, sub = “ “, cex = .9
```

identifies each observation each.

Second argument to cutree () is number of clusters obtained. This data, complete, average linkage separate observations into correct groups. Thence, single linkage identifies one point as belonging to its own cluster. A more sensible answer is obtained when four clusters are selected, hence, there are still two singletons.

```
cutree (hc.single, 4)
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 3 3 3 3
[30]  3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

Scale variables before performing hierarchical clustering of observations, use scale () function

```
Xsc <- scale (x)
plot (hclust (dist (xsc), method = “complete”),
          main = “Hierarchical clustering with scaled features”)
```

Correlation-based distance is computed using as.dist () function, this converts an arbitrary square symmetric matrix into a form that the hclust () function recognizes as a distance matrix. Hence, only makes data sense for data with at least three features hence the absolute correlation between any two observations of measurements on two features is always 1. Hence, we cluster a three dimensional data set. This data set does not contain any true clusters.

```
x <- matrix (rnorm (30 * 3), ncol = 3)
dd <- as.dist (1 – cor (t (x) ) )
plot (hclust (dd, method = “complete”),
          main = “complete linkage with correlation-based distance”,
          xlab = “ “, sub = “ “)
```

## Results

K-means clustering makes the attempt to group exact types of objects in clusters form. It effectively gathers same qualities between objects and groups them into clusters. Example. Going grocery shopping to purchase plantains. There you see various kinds of plantains. You notice there that the plantains are arranged in a group of their kinds. Like all plantains are kept in one place, eggs are kept with its kinds. If you notice here then you find that they are forming a group or cluster, where each of plantains is kept within its kind of formed group clusters.

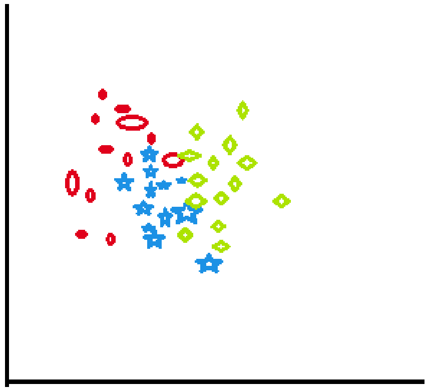

fig 1: before applying k-means clustering

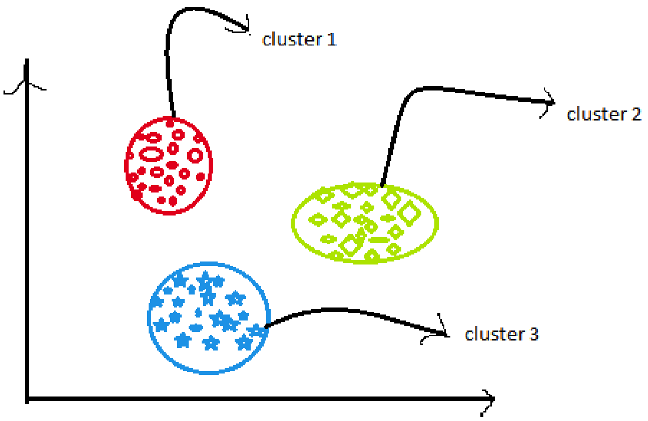

fig 2: After applying K-means clustering

From the two figures. The observation shows the first figure. The first figure on the left shows data before k-means clustering algorithm is applied. Here these three different categories are disconfigured. Then you see such data in real world, you are not able to figure out the different categories. The second figure shows data after k-means clustering algorithm is applied. It is obvious that all three different items are organised into three different categories called clusters.

Assessment. **Is the content appropriate for a results section? Is there a clear description of the results?** Yes. It is appropriate in how each observation is identified. Yes.

**Are the results/data analyzed well? Given the data in each figure/table is the interpretation accurate and logical? Is the analysis of the data thorough? Is anything relevant ignored?** Yes. The results are analyzed well in order of before and after the k-means clustering algorithm is applied. The data analysis is thorough. None ignored.

**Are the figures/tables appropriate for the data being discussed? Are the figure legends and titles clear and concise?** Yes. The figures are appropriate in how linked and clustered they are for the dendrogram. Yes the titles are clear in the order of k-means clustering data and functions of two simple simulations.

<u>Discussion & Conclusions.</u>

<u>Objective restated and connections drawn between your analyses and objective.</u>

The algorithm of k-means explained clearly.

1. Select number K to decide amount of clusters.
2. Select random K points.
3. Assign each data point to their closest points, which forms predefined K clusters.
4. Calculate variance and place a new point of each cluster.

5. Repeat step 3, this reassign each datapoint to new closest point of each cluster.
6. If any reassignment happens, go to step 4 else complete.
7. Model is ready.

Limitations

1. Output is hugely influenced by original input, the amount of clusters.
2. In a awhile, it is tough to forecast amount of clusters, or k value.
3. An array of data substantially hits concluding outcomes.
4. In some scenario, clusters show complex spatial views, execution of clustering is not a good option.
5. Rescaling is sometimes conscious, it is not done by normalization of data points, output gets changed completely.

Concluding sentence summarizing key findings and impact on field

K-means clustering is unsupervised machine learning algorithm that is part of a wide pool of data operations and techniques in the realm of data science. It is the fastest and most efficient algorithm to organize data points into groups even when little data is available about data. Taking notes of the right algorithm, can save time and efforts and assist in gathering more accurate outcome.

Possible areas for future research to better investigate your research question

K-means is a persuasive algorithm that demonstrates a type of context in data science, this approach incorporates two parts in its procedure. K-means is expectation to maximization (EM) algorithm.

1. Re-run to transform.
2. Assume some cluster centres.

E-step: Appoint data points to the closest cluster centre.
M-step: Introduce cluster centres to the mean.

Here E-step is expectation step, this comprises upgrading forecasts of associating data point with respective cluster.

M-step is maximization step, it includes maximizing some features that specify the cluster centres region, for this maximization, this is expressed by considering the mean of data points of each cluster.

Each reiteration of E-step and M-step algorithm always yield in terms of improved estimation of clusters characteristics.

Assessment. **Do you clearly state whether the results answer any questions posed?** Yes. The result of each reiteration answers the questions of if cluster centres region is maximized**. Were specific data cited from the results to support each interpretation?** Yes. E-step and M-step. **Do you clearly articulate the basics for supporting or rejecting any hypotheses?** Yes.

References.

Assessment. **Are references appropriate and of adequate quality?** Yes. **Are the cited property (both in the text and at the end of the paper)?** Yes.

Aman Ravi. (2021). Understanding K-means clustering in machine learning by.

Gareth James. Daniela Witten. Trevor Hastie. Robert Tibshirani. (2021). An Introduction to Statistical Learning with Applications in R.