**Implementing the first two epochs of a Clustering Algorithm**



# ****Step 1 (Initialisation)****

The first step to be carried out as part of this assignment, is to enter the eight data records as data points onto a graph in **Appendix Five** as per **Appendix One;** with the initial clusters set as shown in **Appendix two**, using the distance matrix table as per **Appendix Three**, to make an initial decision as to what data point belongs to one of the three clusters in **Appendix Four.**

# Step 2.1 – Calculations of Epochs

Now the data points have been plotted into a cartesian plane via an x and y coordinate graph, applied the initial centroids, and a preliminary clustering of data points to said centroids is complete; the next phase is an attempt to compute new centroids for each of these clusters using the equation below for each cluster [1, eq. (4-2)]:

Then, to recalculate the Euclidean Distance of each data point against these new centroids as shown in the below equation [1, eq. (4-1)], [3, Fig. 25], [4, Fig. 29]:

Also shown as the equation below [2, p. 145]

Once both calculations have been carried out for the centroids and its associated data points, a plot will be created to highlight the potential change of clustering performed. In the next section, a flow chart will display the method in which the algorithm will take place to conduct the K-means Clustering algorithm, based on the step after the initialisation phase.

# Step 2.2 – Flowchart

The flowchart shown in Figure 1 below demonstrates as to how the k-means clustering algorithm can be implemented.

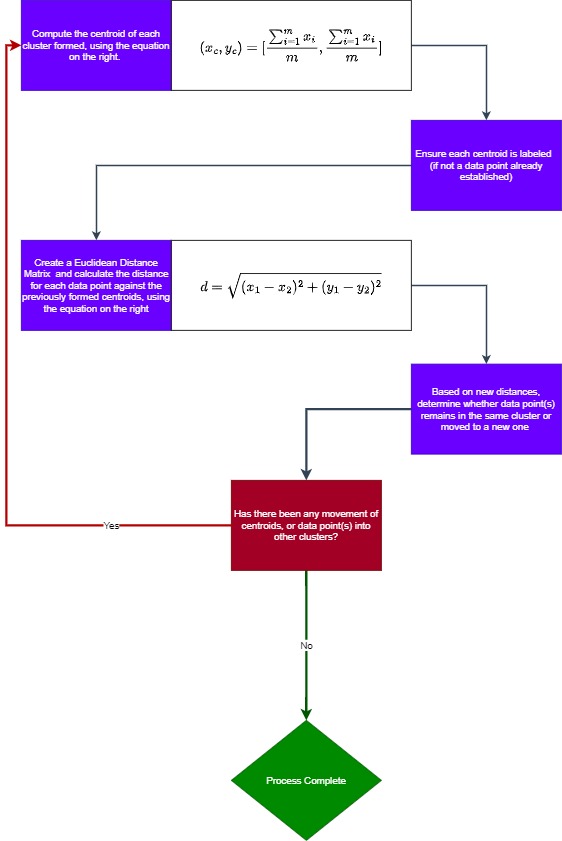


Figure 1 - Flowchart of K-means Clustering Algorithm.

# Step 2.3 – Epoch One Calculations

## Step 2.3.1 – Compute Centroid Creation for all three clusters

The content provided in Figure 2, represents the creation and computation required for centroids to be implemented within the three individual clusters previously recorded. The content presented can also be found in the attached Excel Document in “Epoch One” sheet attached alongside this word document.



Figure 2 - Epoch One Centroid Computation for all Three Clusters

## Step 2.3.2 – Compute Euclidean Distance Matrix

The information shown in Figure 3, demonstrates the computation table used for centroids of each cluster. The content presented can also be found in the attached Excel Document in “Epoch One Euclidean Distance” sheet attached alongside this word document. 

Figure 3 - Epoch One Euclidean Distance Matrix calculations between data points and centroids.

## Step 2.3.3 – Epoch One Cluster Plot

Figure 4 shows the plot on the output of the clusters formed, since the introduction of new centroids onto the data.

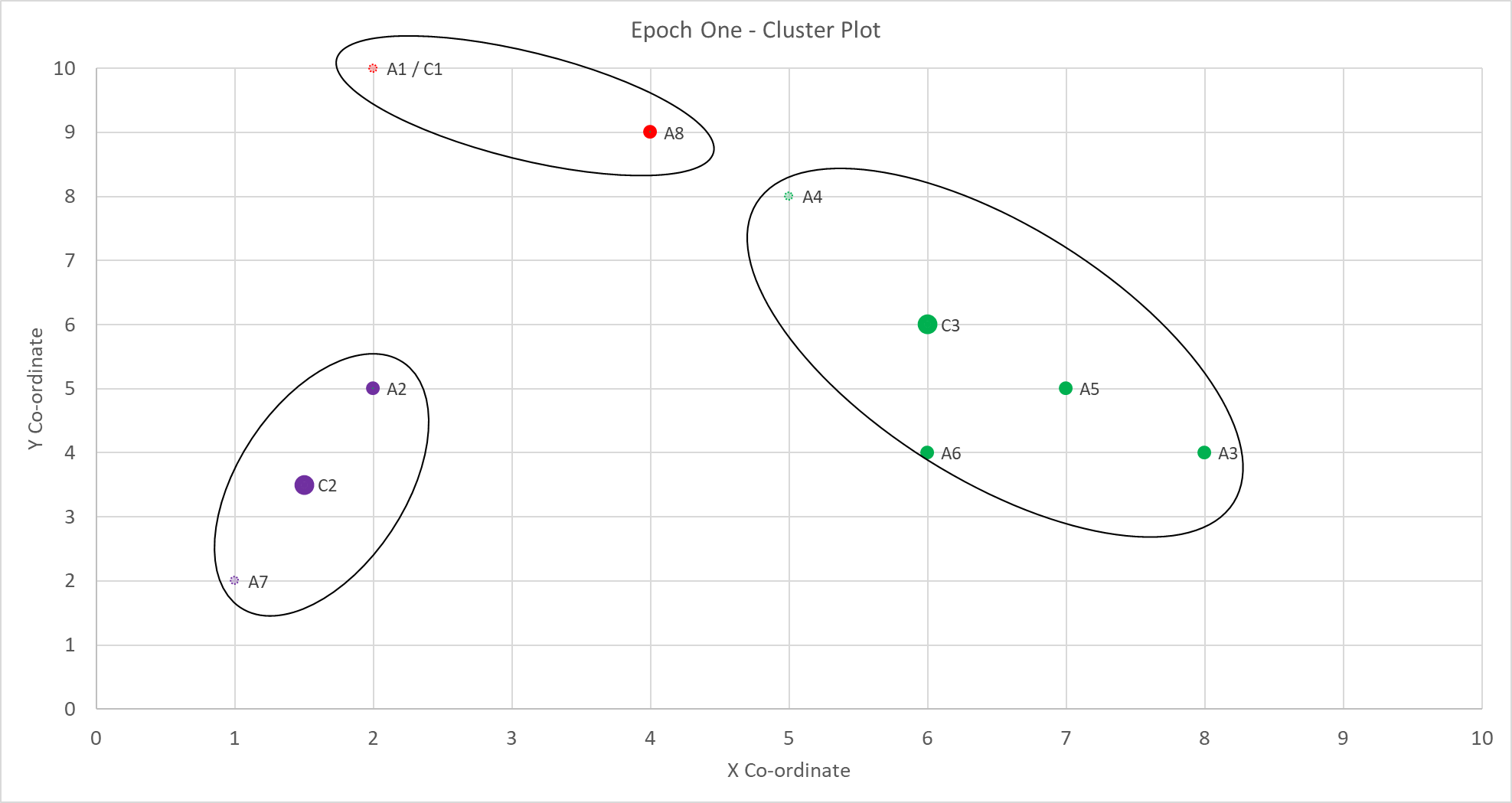


Figure 4 - Plot of output from cluster in Epoch One.

As it shows, the new centroids are colour-coded and in a larger size compared to the data points provided. The previous initial centroids as part of the initialisation in Step 1, are in a lighter colour with a dashed border. This chart is also available in the Excel workbook attached alongside this Word document, under the sheet “Epoch One”. The next step is to perform the relevant calculations and show the results from Epoch Two.

# Step 2.4 – Epoch Two

## Step 2.4.1 – Compute Centroid Creation for all three clusters

The content provided in Figure 5, represents the creation and computation required for centroids to be implemented within the three individual clusters previously recorded. The content presented can also be found in the attached Excel Document in “Epoch Two” sheet attached alongside this word document. As these are new centroids based on the clusters found in Epoch One results, the previous computed centroids will be visible in the final plot output, but not used in the computation of their creation itself.



Figure 5 - Epoch Two Centroid Computation for all Three Clusters

## Step 2.4.2 – Compute Euclidean Distance Matrix

The information shown in Figure 6, demonstrates the computation table used for centroids of each cluster. The content presented can also be found in the attached Excel Document in “Epoch Two Euclidean Distance” sheet attached alongside this word document.



Figure 6 - Epoch Two Euclidean Distance Matrix calculations between Data Points and Centroids

## Step 2.4.3 – Epoch Two Cluster Plot

Figure 7 shows the plot on the output of the clusters formed, since the introduction of new centroids onto the data.

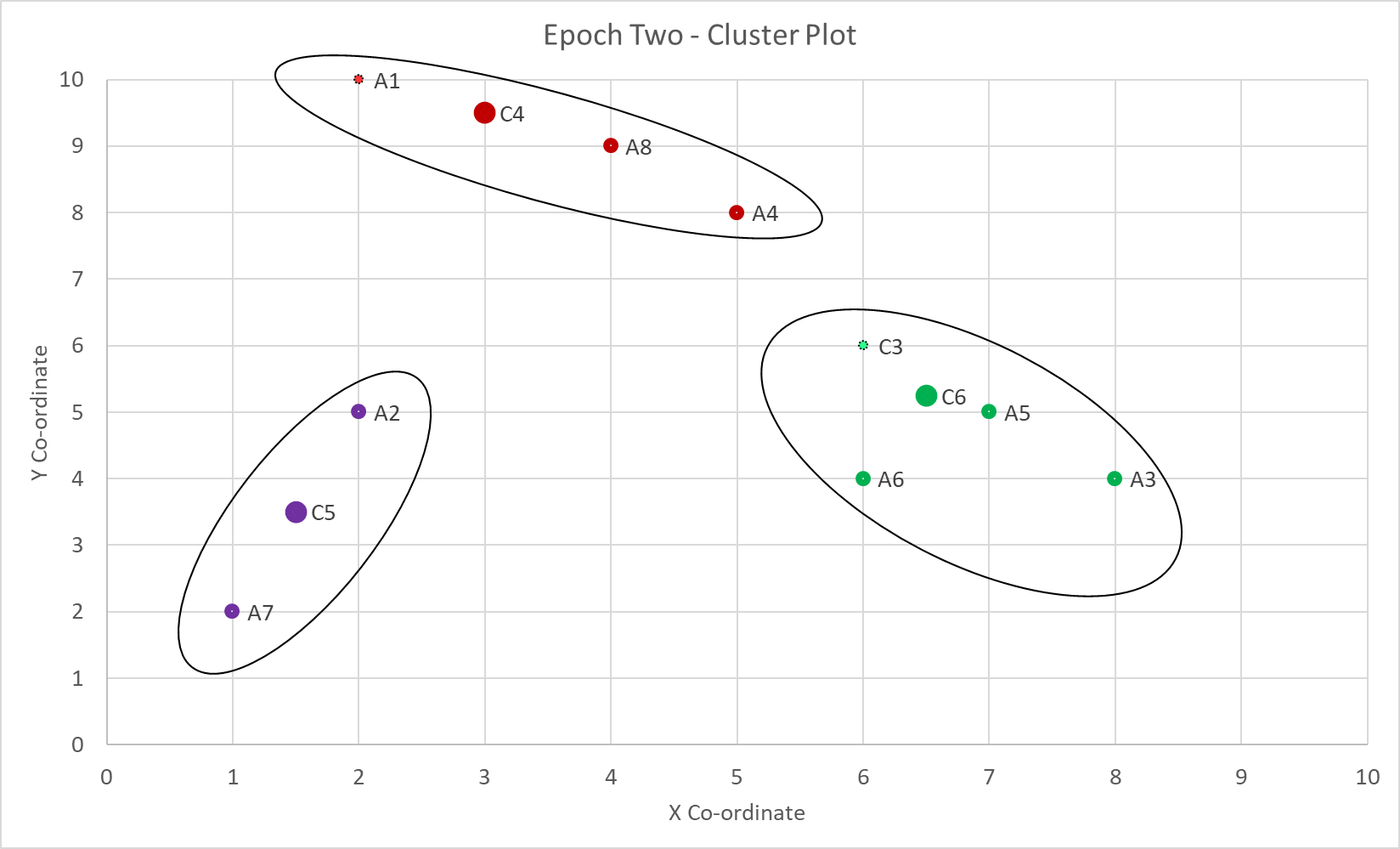


Figure 7 - Plot of output from cluster in Epoch Two.

Since the output from the last epoch, there are elements worth noting. Firstly, there has been no change in centroids between C5 and C2 from the previous epoch. Secondly, there has been a shift of data point(s) from the green cluster to the red cluster, which appears to balance out.

# Step 3 - How many more iterations are needed to converge?

Based on the examination of the previous effort carried out against this data, a good guide initially would be to run this algorithm over 23 more epochs (making 5 Epochs in total). Whilst 10 would intuitively feel like a solid amount to go for, 5 would suffice due to the little movement expected after the initial 2 epochs have been declared.

# References

[1] D. Dietrich, B. Heller, and B. Yang, Data science & big data analytics : discovering, analyzing, visualizing and presenting data. Indianapolis, In: Wiley, 2015, p. 121.

[2] S. Sedkaoui, Data analytics and big data. London: Iste Ltd / John Wiley & Sons, Inc, 2018, p. 145.

[3] O. Theobald, Machine learning for absolute beginners : a plain English introduction. United States: The Author, 2017, p. 67.

[4] O. Theobald, Statistics for absolute beginners. S.L.: Scatterplot Press, 2017, p. 134.

# Appendix One – Data Records

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **X Co-ordinate** | **Y Co-ordinate** | **X/Y Co-ordinates** |
| A1 | 2 | 10 | (2,10) |
| A2 | 2 | 5 | (2,5) |
| A3 | 8 | 4 | (8,4) |
| A4 | 5 | 8 | (5,8) |
| A5 | 7 | 5 | (7,5) |
| A6 | 6 | 4 | (6,4) |
| A7 | 1 | 2 | (1,2) |
| A8 | 4 | 9 | (4,9) |

# Appendix Two – Initial Centroids

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **X Co-ordinate** | **Y Co-ordinate** | **X/Y Co-ordinates** |
| A1 | 2 | 10 | (2,10) |
| A4 | 5 | 8 | (5,8) |
| A7 | 1 | 2 | (1,2) |

# Appendix Three – Euclidean Distance matrix table

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field** | **A1** | **A2** | **A3** | **A4** | **A5** | **A6** | **A7** | **A8** |
| A1 | 0 | √25=5 | √36=6 | √13 | √50 | √52 | √65 | √5 |
| A2 |  | 0 | √37 | √18 | √25=5 | √17 | √10 | √20 |
| A3 |  |  | 0 | √25 | √2 | √2 | √53 | √41 |
| A4 |  |  |  | 0 | √13 | √17 | √52 | √2 |
| A5 |  |  |  |  | 0 | √2 | √45 | √25=5 |
| A6 |  |  |  |  |  | 0 | √29 | √29 |
| A7 |  |  |  |  |  |  | 0 | √58 |
| A8 |  |  |  |  |  |  |  | 0 |

# Appendix Four – Selection of Cluster for each Data Point

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Field** | **A1 Distance** | **A4 Distance** | **A7 Distance** | **Selection** |
| A2 | √25 | √18 | √10 | A7 |
| A3 | √36 | √25 | √53 | A4 |
| A5 | √50 | √13 | √45 | A4 |
| A6 | √52 | √25 | √29 | A4 |
| A8 | √5 | √2 | √58 | A4 |

# Appendix Five – Initial Clustering of data points