**Assignment 4**

By Harsh Sharma

Part 3

**Part III. Clustering**

**Question 1.** Given the initial cluster centroids and the distances provided, identify which cluster each observation belongs to. Justify your answer with the distance values.

Solution: The data points will belong to the cluster which they have the closest distance. From the table given we can find the cluster data points.

Centroid 1: P1, P2, P5, P11, P12

Centroid 2: P4, P6, P7, P8, P15

Centroid 3: P3, P9, P10, P13, P14

**Question 2**. After assigning each observation to the nearest cluster based on the provided distances, calculate the new centroids for each cluster. Present the new centroid values.

Solution: Let’s calculate distances.

Centroid 1 (Pressure) = (- 0.392 - 0.251 - 0.597 - 0.736 - 0.288) / 5

= -0.4528

Centroid 1 (Temperature) = ( -1.258 - 1.781 - 1.694 - 2.116 - 1.577) / 5

= -1.6852

Centroid 1 (Volume) = (-0.666 - 1.495 - 0.686 - 1.165 - 0.618) / 5

= -0.926

**Centroid 1 = 〈-0.4528, -1.6852, −0.926〉**

Centroid 2 (pressure) = (-0.917 - 1.204 - 0.778 - 1.075 - 1.28) / 5

= -1.0508

Centroid 2 (Temperature) = (-0.961 - 0.605 - 0.436 - 1.199 - 1.188) / 5

= -0.8778

Centroid 2 (Volume) = (0.055 + 0.351 - 0.22 - 0.141 + 0.053) / 5

= 0.0196

**Centroid 2 = 〈-1.0508, -0.8778, 0.0196〉**

Centroid 3 (Pressure) = (-0.823 – 0.854 – 1.027 – 1.113 -0.849) / 5

= -0.9332

Centroid 3 (Temperature) = (-0.042 – 0.654 – 0.269 -0.271 -0.430)

= -0.3332

Centroid 3 (Volume) = (1.254 + 0.771 + 0.893 + 0.930 + 0.612) / 5

= 0.892

**Centroid 3 = 〈-0.9332, --0.3332, 0.892〉**

**Question 3.** Evaluate the homogeneity of the clusters formed in the first iteration using (i) the Within-Cluster Sum of Squares and (ii) the average Silhouette Coefficient. Which cluster appears to be the most homogeneous, and why?

Answer: I’ve written a python script to calculate Euclidean distance between each point and then to calculate Sum of squares within each cluster. Here are the results of the script. I’ll be attaching the script as well in the submission.

Distance between Centroid (-0.4528, -1.6852, -0.926) and point 1 (-0.392, -1.258, -0.066) in this Cluster: 0.9621831842222146

Distance between Centroid (-0.4528, -1.6852, -0.926) and point 2 (-0.251, -1.781, -1.495) in this Cluster: 0.6112788888878791

Distance between Centroid (-0.4528, -1.6852, -0.926) and point 3 (-0.736, -1.694, -0.686) in this Cluster: 0.3713215318292221

Distance between Centroid (-0.4528, -1.6852, -0.926) and point 4 (-0.288, -2.116, -1.165) in this Cluster: 0.5194888641732371

Distance between Centroid (-0.4528, -1.6852, -0.926) and point 5 (-0.597, -1.577, -0.618) in this Cluster: 0.35688216542718976

Distance between Centroid (-1.0508, -0.8778, 0.0196) and point 1 (0.917, -0.961, 0.055) in this Cluster: 1.9698761991556728

Distance between Centroid (-1.0508, -0.8778, 0.0196) and point 2 (1.204, -0.605, 0.351) in this Cluster: 2.2952927569266626

Distance between Centroid (-1.0508, -0.8778, 0.0196) and point 3 (0.778, -0.436, -0.22) in this Cluster: 1.8966035009985613

Distance between Centroid (-1.0508, -0.8778, 0.0196) and point 4 (1.075, -1.199, -0.141) in this Cluster: 2.1559191636051662

Distance between Centroid (-1.0508, -0.8778, 0.0196) and point 5 (1.28, -1.188, 0.053) in this Cluster: 2.351588450388375

Distance between Centroid (-0.9332, -0.3332, 0.892) and point 1 (-0.823, -0.042, 1.254) in this Cluster: 0.4774782508135842

Distance between Centroid (-0.9332, -0.3332, 0.892) and point 2 (-0.854, -0.654, 0.771) in this Cluster: 0.35188958495528116

Distance between Centroid (-0.9332, -0.3332, 0.892) and point 3 (-1.027, -0.269, 0.893) in this Cluster: 0.11367092856135194

Distance between Centroid (-0.9332, -0.3332, 0.892) and point 4 (-1.113, -0.271, 0.93) in this Cluster: 0.1940125769119105

Distance between Centroid (-0.9332, -0.3332, 0.892) and point 5 (-0.849, -0.43, 0.612) in this Cluster: 0.3079933116156908

Sum of squares for Cluster 1: 1.8345716

Sum of squares for Cluster 2: 22.9238416

Sum of squares for Cluster 3: 0.49723360000000005

According to **Within-cluster Sum of Squares (WCSS) cluster 3 is the most homogeneous cluster as it has the least value of 0.49 representing that the coordinates are very close to the median.**

ii) the average Silhouette Coefficient

To calculate the average Silhouette Coefficient I’ve also written a python script which uses sklearn for the calculation. I’ll be attaching the code as well for it. Here are the average Silhouette Coefficient for each of the three medians.

The silhouette score for the provided coordinates using median (-0.4528, -1.6852, -0.926) is: 0.4794025641653915

The silhouette score for the provided coordinates using median (-1.0508, -0.8778, 0.0196) is: 0.2695749022983073

The silhouette score for the provided coordinates using median (-0.9332, -0.3332, 0.892) is: 0.4598320525989384

According to average Silhouette Coefficient, the most homogeneous cluster is Cluster 1 as it has the maximum value of 0.479, closest to 1.

**Overall, I think cluster 1 is the most homogeneous cluster as it has better overall score in both the performance metrices.**

**Question 4**. Assuming the new centroids have been calculated, describe how you would perform the next iteration of k-means clustering. What changes would you expect in the cluster memberships?

**Answer**: We have already calculated new medians for the next iteration. And so, we’ll reassign the data points to the nearest centroids and recalculate the centroids based on these new assignments. This process continues iteratively until convergence, typically when the centroids no longer change significantly or when a predefined number of iterations is reached.

It’s possible that in the next iteration some points will change their centroid association. How will this association change will depend on the Euclidean distance.

**Question 5.** Based on the distances to the centroids, identify any potential outliers in the dataset. Discuss how outliers can influence the formation of clusters and centroid calculation.

**Answer**: Using Euclidean Distance, **Point 15** is the outlier which belongs to the **second cluster**. It has the maximum distance from its median among all coordinates and it is 2.351588.

The outlier will try to pull the mean towards it. Centroids in algorithms like K-means are calculated as the mean position of all points within a cluster. Outliers, being significantly different from other points, can skew the mean, pulling the centroid closer to them. This results in centroids that might not represent the cluster well, especially if the outlier is an erroneous or irrelevant data point.