EML Assignment 5 Problem 3

Muhammed Saeed and Ali Salaheldin Ali Ahmed

January 2023

Question 1

(a) The two clusters are the same in both cases. Hence, single linkage would result in a lower or equal dissimilarity to complete linkage. According to this, the fusion should occur at lower level in the single linkage case or, at most, at the same level as complete linkage.

(b) Since both clusters contain a single observation, both linkage methods produce the same dissimilarity which is just the dissimilarity between observations 5 and 6. Thus, both fusions should occur at the same level.

Question 2

Correlation-based distance can be used as a dissimilarity measure instead of Euclidean distance. It considers two observations similar if their features are highly correlated, even though they may be far in terms of Euclidean distance.

This measure is useful in situations where the shape of the observations rather than the magnitudes. For instance, clustering customers based on their shopping habits and items bought. In this task, the actual amount of items bought is not as important as the distribution of purchases across items.

Question 3

1. Which type of clustering is appropriate for the problem at hand? K-means and hierarchical clustering are exclusive clustering methods, which may not always be fitting.

2. Should the observations or features first be standardized in some way?

   For instance, maybe the variables should be scaled to have standard deviation one.

3. In the case of K-means clustering, how many clusters should we look for in the data?

4. In the case of hierarchical clustering, what dissimilarity measure and which type of linkage should be used?