CS422-01: Homework #8

Vijay K. Gurbani

Sujad Patel

December 4, 2022

Part I

1. Exercises

1.1 Tan, Chapter 7 (Cluster Analysis: Basic Concepts and Algorithms)

* 1

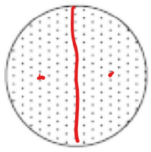
2. Find all well-separated clusters in the set of points shown in Figure 7.35

The first group of points are closest to each other, therefore it is the most well-separated. The second group of points are farther apart in relation to each other than the first group, so it is less well separated and finally the third group of points are closer to each sub group than others points the same group, therefore it is not well separated. The sets of points can be clustered into 3 different groups or 5 groups.

* 2

6.

(a) k=2. Assuming that the points are uniformly distributed in the circle, how many possible ways are there (in theory) to partition the points into two clusters? What can you say about the positions of the two centroids?



(a)

With k = 2, circle can be partitioned into equal halves and the each would have semi-circle centroids.

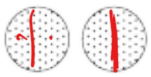
(b) k =3. The distance between the edges of the circles is slightly greater than the radii of the circles.

(b)

One of the circles would house two clusters.

* 2 continued on next page. . .

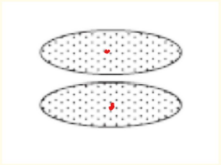
(c) k=3. The distance between the edges of the circles is much less than the radii of the circles.



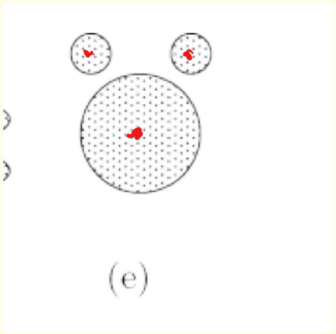
(c)

One of the circles would house two clusters.

(d) k=2.



(e) k=3.



(e)

Since the K matches number of circles, each circle is a cluster with centroid of that circle.

* 3

7. For the given data set, which of the following should occur in order to minimize the squared error when finding K clusters?

B is the correct choice. Since the dense region is closely packed but spread out region has much larger distance to its centroid than dense region. Therefore, more centroids in spread out region will help minimize that distance.

* 4

11. Total SSE is the sum of the SSE for each separate attribute. What does it mean if the SSE for one variable is low for all clusters? Low for just one cluster? High for all clusters? High for just one cluster? How could you use the per variable SSE information to improve your clustering?

** 4 continued on next page...*

Low SSE for one variable for all cluster means that the variable is almost a constant, and the constant variable is not used for separating the data into clusters.
Low SSE for one cluster mean that low variation from the data sets.
Higher SSE for all cluster means that there is high variation between the data sets. The variable can be a noise.
High SSE for one cluster means that its has strong association with a cluster in one data set but cannot be replicated in other data sets. SSE is used to eliminate variable with poor relationship between clusters.

* 5

12.

(a) What are the advantages and disadvantages of the leader algorithm as compared to K-means?

The advantages are:
single scan of data is required for the leader algorithm. Hence, it is more computationally efficient than the K-means algorithm.
The leadership algorithm is an incremental clustering algorithm.
The disadvantages are:
It is impossible to set the count of the resulting clusters as in the case of the leadership algorithm.
When compared to the K-means algorithm, the leadership algorithm is not efficient to produce clusters with better quality.

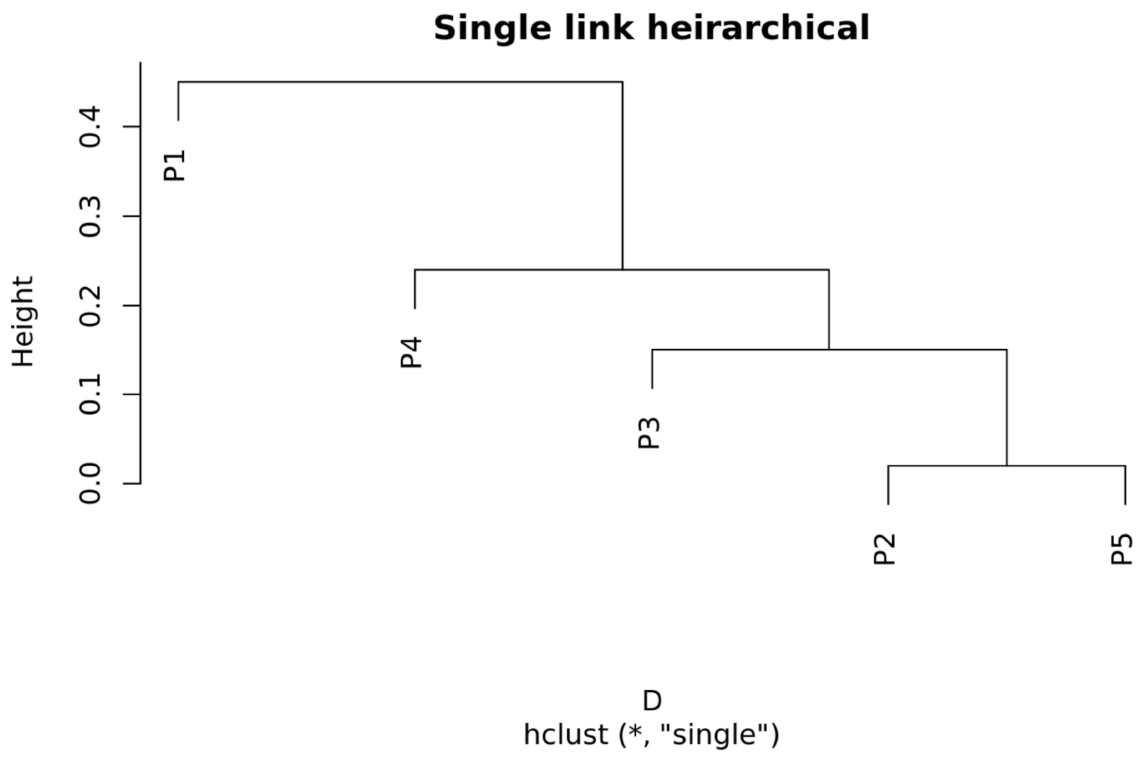
(b) Suggest ways in which the leader algorithm might be improved.

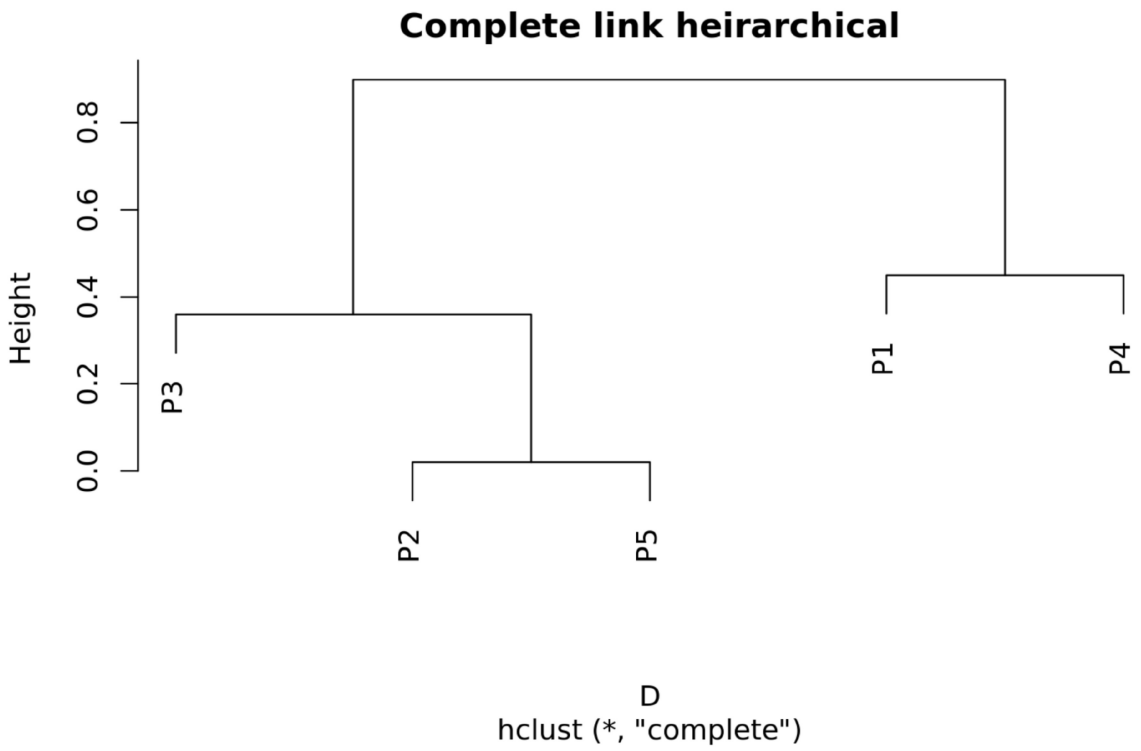
Sample can be used for determining the distribution of distances among the points and it can be utilized more effectively to set the threshold value.

* 6

16. Use the similarity matrix in Table 7.13 to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

** 6 continued on next page...*





* 6 continued on next page...