Hierarchical Clustering Quiz Solutions

Zeqing Jin
zjin2017@berkeley.edu

Xianlin Shao
shayd@berkeley.edu

Yifei Zhang
yifei_zhang@berkeley.edu

Zilan Zhang
shilan@berkeley.edu

Quiz Questions

1. Describe a scenario in which you can apply hierarchical clustering. State your objective, data, definition of affinity, and so on.

   Answer: This is an open ended question. It is the most important kind of question which ensures students have the ability to reflect what they have learned in other settings.

2. Which of the following statements are correct with respect to Hierarchical Clustering?

   A. Hierarchical clustering always requires to specify the number of clusters.
   B. Hierarchical clustering has a global optimization objective.
   C. Divisive clustering may require higher time complexity compared to aggomerative clustering

   Answer: C. For A, a complete hierarchical clustering covers all the combinations of k values, so we may not define number of clusters.

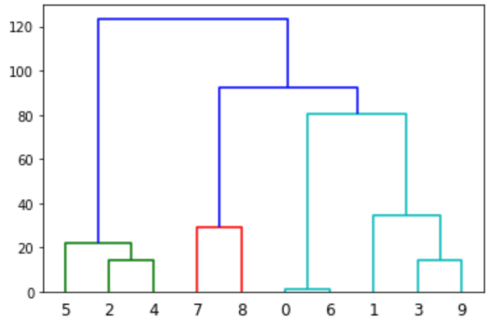
   Objective: This question examines students' basic understanding of hierarchical clustering.

3. For naive agglomerative clustering, calculate the time complexity.

   Answer: There are n iterations. In each iteration the complexity of dissimilarity has $O[n^2]$ complexity, so $O[n^3]$ in total.

   Objective: This question test students' basic understanding of agglomerative clustering process.

4. Answer the following questions based on the given Dendrogram.



i. Which two clusters are merged in the first iteration?

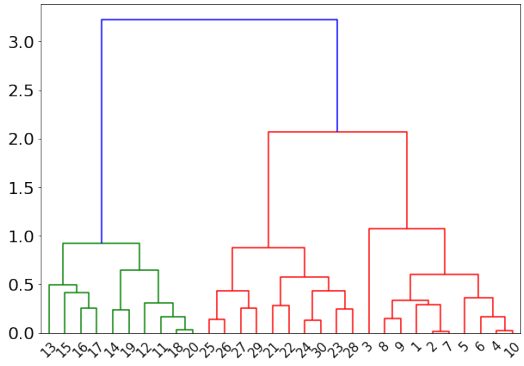
ii. If we break the cluster at dissimilarity level 60, how many clusters do we have ?

iii. What is the approximate dissimilarity between cluster containing 0, 6 and cluster containing 1, 3, 9?

Answer: 0 and 6; 4; around 80.

Objective: This question checks students' general understanding of dendrogram. They should know how to extract important clustering information from the dendrogram.

5. A dendrogram of hierarchical clustering using average linkage is given below. Roughly determine a reasonable number of cluster k.



Answer: Judging only from the dendrogram, it is reasonable to take k to be 2 or 3. However, note that since we are using average linkage, the average linkage between the merged cluster and other clusters would change. This could hardly be revealed from the dendrogram.

Objective: The question looks at students' understanding of dissimilarity from dendrogram. They should know how to generally determine the proper number of clusters.

6. The dissimilarity matrix of 5 points {1,2,3,4,5} is given below. Obviously we will first merge {3,4} because of minimum distance. Determine the two clusters merging in the next iteration using:

$$\begin{bmatrix} 0 & 12 & 40 & 35 & 24 \\ 12 & 0 & 21 & 66 & 27 \\ 40 & 21 & 0 & 8 & 18 \\ 35 & 66 & 8 & 0 & 10 \\ 24 & 27 & 18 & 10 & 0 \end{bmatrix}$$

i. Single linkage.

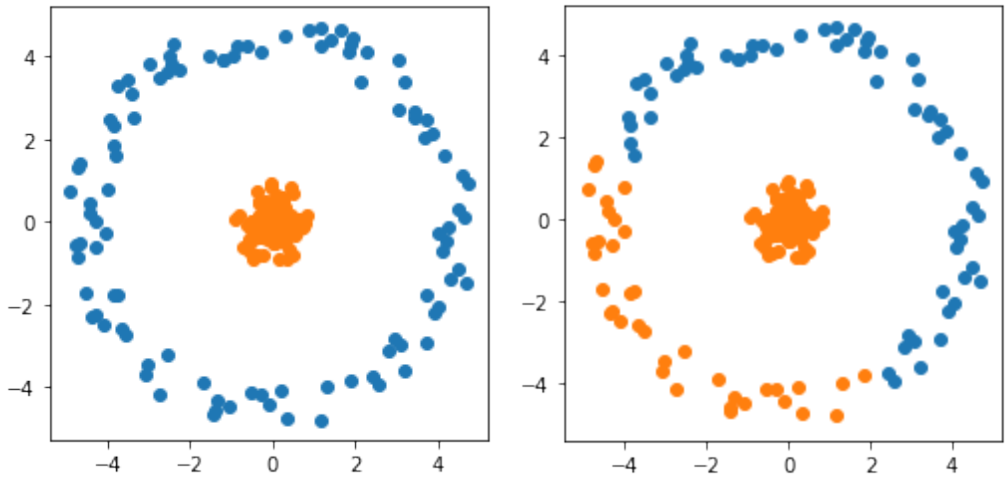
ii. Complete linkage.

iii. average linkage.

Answer: Single linkage: {3,4} and {5}, complete linkage: {1} and {2}, average linkage: {1} and {2}.

Objective: This question requires students to know the definition of different linkage algorithm and apply the algorithm to the dissimilarity matrix.

7. Determine which of the following clustering results ($k = 2$) is using single linkage and which one is using complete linkage.

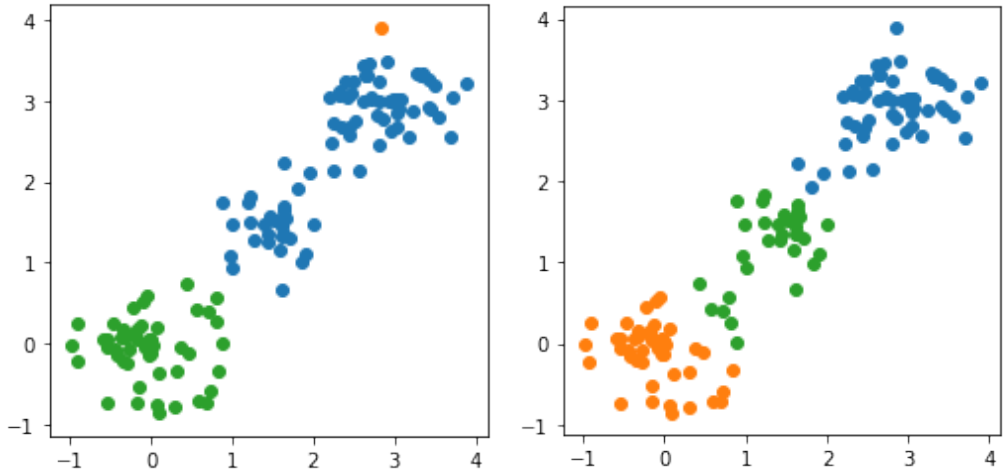


(a) (b)

Answer: (a) is single linkage, (b) is complete linkage.

Objective: The question tests students' understanding of single and complete linkage. They should know the feature of different linkages.

8. Determine which of the following clustering results ($k = 3$) is using single linkage and which one is using complete linkage.



(a) (b)

Answer: (a) is single linkage, (b) is complete linkage.

Objective: The question tests students' understanding of single and complete linkage. They should know the difference between different linkage functions.

9. Which of the following statements are correct with respect to Divisive Hierarchical Clustering?

   A. Divisive Hierarchical clustering splits n objects from 1 cluster into n clusters.

   B. When deciding which cluster to split, we choose the one with most objects.

   C. When splitting one cluster, the splinter cluster always starts from one object.

Answer: AC. When choosing the cluster to split, we look for the one with the highest diameter.

Objective: This question tests students' basic understanding of divisive hierarchical clustering algorithm. They should know how the naive algorithm works.

10. (Optional) Which of the following statements are correct with respect to CURE?
    A. CURE has a good application on data with outliers.
    B. The shrinking factor moves all the points in a cluster to the centroid.
    C. CURE effectively reduce the time complexity compared to agglomerative clustering and divisive clustering.

    Answer: A. The Shrinking factor only moves the representative points towards the centroid. CURE do not reduce too much calculations compared to agglomerative and divisive algorithms.

    Objective: This question tests students' basic understanding of CURE features and how it works.