Numerical Questions on Hierarchical Clustering

Numerical Questions on Hierarchical Clustering

1. **Distance Calculation**: Given the following points in a 2D space:

   - Point A: (1, 2)
   - Point B: (3, 4)
   - Point C: (6, 8)

   Calculate the Euclidean distance between:

   (a) Point A and Point B

   (b) Point A and Point C

   (c) Point B and Point C

2. **Single-Linkage Clustering**: Using the distances calculated in Question 1, apply single-linkage clustering:

   (a) If you treat each point as a separate cluster, show how clusters would merge step-by-step until all points belong to one cluster.

   (b) At which step does the first merge occur, and which points are merged?

3. **Complete-Linkage Clustering**: Using the same set of points from Question 1, perform complete-linkage clustering:

   (a) Determine the first two clusters that would merge based on the maximum distances calculated from the points.

   (b) What would the distance be between these clusters?

4. **Group Average Linkage**: Suppose we have two clusters formed after the first two merges in a hierarchical clustering process:

   - Cluster 1: A, B with points A (1, 2) and B (3, 4)
   - Cluster 2: C with point C (6, 8)

   Calculate the average distance between the points in Cluster 1 and Cluster 2. Use the Euclidean distance as your measure.

5. **Dendrogram Construction**: Based on the following distances between points A, B, and C:

   - Distance (A, B) = 2.24
   - Distance (A, C) = 7.21
   - Distance (B, C) = 4.47

   Construct a simple dendrogram that shows how these points would be clustered. Indicate the height at which each merge occurs.

6. **Silhouette Score Calculation**: Consider three clusters with the following average distances:

   - Cluster 1: Average distance within the cluster = 0.5
   - Cluster 2: Average distance within the cluster = 1.2
   - Cluster 3: Average distance to the nearest cluster (inter-cluster distance) = 1.5

Calculate the silhouette score for Cluster 1 using the formula:

$$\text{Silhouette Score} = \frac{b - a}{\max(a, b)}$$

where a is the average distance to points in the same cluster and b is the average distance to the nearest cluster.

7. **Cophenetic Correlation Coefficient**: Suppose you computed the cophenetic distances for a hierarchical clustering and obtained the following distances:

   - Observed distances: [2, 3, 5]
   - Cophenetic distances: [2.2, 3.1, 5.5]

   Calculate the cophenetic correlation coefficient using the formula:

$$r = \frac{\text{Cov}(\text{observed}, \text{cophenetic})}{\sigma_{\text{observed}} \sigma_{\text{cophenetic}}}$$

   (You can assume hypothetical values for the covariance and standard deviations for this calculation if needed).

8. **Non-Monotonicity Example**: Given a hypothetical dendrogram showing merges of three clusters (A, B, C) at different heights, analyze whether the dendrogram violates monotonicity.

   - List the merged clusters and their respective heights.
   - Identify if any clusters at higher levels are more similar than those at lower levels.

Answers to Selected Questions

1. **Distance Calculation**:

(a) $d(A, B) = \sqrt{(3-1)^2 + (4-2)^2} = \sqrt{2^2 + 2^2} = \sqrt{8} \approx 2.83$

(b) $d(A, C) = \sqrt{(6-1)^2 + (8-2)^2} = \sqrt{5^2 + 6^2} = \sqrt{25 + 36} = \sqrt{61} \approx 7.81$

(c) $d(B, C) = \sqrt{(6-3)^2 + (8-4)^2} = \sqrt{3^2 + 4^2} = \sqrt{9 + 16} = \sqrt{25} = 5$

2. **Single-Linkage Clustering**:

(a) First merge: A and B (distance = 2.83). Next merge: (A, B) with C (distance = 5).

(b) First merge occurs between A and B.

3. **Complete-Linkage Clustering**:

(a) First merge: A and B (distance = 2.83). Next merge: (A, B) with C (distance = 7.81).

(b) The distance between the merged clusters is 7.81.

4. **Group Average Linkage**:

- Average distance between Cluster 1 (A, B) and Cluster 2 (C):

$$\text{Average distance} = \frac{d(A, C) + d(B, C)}{2} = \frac{7.81 + 5}{2} \approx 6.41$$

5. **Dendrogram Construction**:

- Merges would occur at:
  - A and B at height 2.24.
  - A, B and C at height 4.47.

6. **Silhouette Score Calculation**:

- For Cluster 1: $a = 0.5$, $b = 1.5$

$$\text{Silhouette Score} = \frac{1.5 - 0.5}{\max(0.5, 1.5)} = \frac{1}{1.5} \approx 0.67$$

7. **Cophenetic Correlation Coefficient**:

- Calculate r with assumed covariance and standard deviations for observed and cophenetic distances.

8. **Non-Monotonicity Example**:

- Analyze and provide reasoning based on the heights and similarities presented.