Assignment 9: Clustering Algorithms

Your Name

April 8, 2024

Sometimes we perform hierarchical agglomerative clustering to discover k clusters: we can cut the dendrogram at the appropriate distance from the root to obtain a specified number of clusters. Suppose we want to generate **two** clusters from a set of **six** tuples below:

$$A1(7,3), A2(6,4), A3(3,6), B1(5,1), B2(9,5), B3(10,4)$$

Let the distance function be Euclidean distance.

1. K-means clustering [9 Points]

For each of the following pairs of initial centroids, calculate distances from each point, and assign each point to a corresponding cluster (C1 or C2). Compute SSE for each solution.

1.1. First iteration [2 points]

A. Initial centroids: A1(7,3), B1(5,1)

	Cluster 1 or 2	A1(7,3)	B1(5, 1)
A1(7,3)			
A2(6, 4)			
A3(3, 6)			
B1(5, 1)			
B2(9, 5)			
B3(10, 4)			

B. Initial centroids: A2(6,4), B2(9,5)

	Cluster 1 or 2	A2(6, 4)	B2(9, 5)
A1(7,3)			
A2(6, 4)			
A3(3, 6)			
B1(5, 1)			
B2(9, 5)			
B3(10, 4)			

1.2. Second iteration [4 points]

Compute new centroids, and recompute new distances.

A. New centroids:

$C_1 =$
$C_2 =$
New distances:

	Cluster 1 or 2	C_1	C_2
A1(7,3)			
A2(6, 4)			
A3(3, 6)			
B1(5, 1)			
B2(9, 5)			
B3(10, 4)			

B. New centroids:

$C_1 =$
$C_2 =$
New distances:

	Cluster 1 or 2	C_1	C_2
A1(7,3)			
A2(6, 4)			
A3(3, 6)			
B1(5, 1)			
B2(9, 5)			
B3(10, 4)			

1.3. Comparing two clustering results [3 points]

Compare the total SSE of two clusters obtained after the second iteration of K-means for A and B:

$$SSE_A = \sum_{i=1}^{K} \sum_{x \in C_i} [\text{dist}(m_i, x)]^2 =$$

$$SSE_B = \sum_{i=1}^{K} \sum_{x \in C_i} [\text{dist}(m_i, x)]^2 =$$

Explain what the difference in SSE tells us about the quality of the clusters in each case.

2. Hierarchical clustering [10 points]

2.1. Full hierarchical clustering [6 points]

Use the same points as above to perform full hierarchical clustering:

$$A1(7,3), A2(6,4), A3(3,6), B1(5,1), B2(9,5), B3(10,4)$$

Fill in the original proximity matrix:

	A1(7,3)	A2(6, 4)	A3(3, 6)	B1(5, 1)	B2(9, 5)	B3(10,4)
A1(7,3)	0					
A2(6, 4)		0				
A3(3, 6)			0			
B1(5, 1)				0		
B2(9, 5)					0	
B3(10,4)						0

Use the **MAX** as inter-cluster distance.

Show every step and the updated proximity matrix at each step. Also draw a final dendrogram. After that cut the dendrogram to obtain two clusters.

2.2. Cluster quality comparisons [4 points]

Compute SSE of these two clusters, and compare their quality to the clusters obtained after two steps of K-means in question 1.

$$SSE_{hierarchical} = \sum_{i=1}^{K} \sum_{x \in C_i} [\text{dist}(m_i, x)]^2 =$$