Amjad Alqahtani
Assignment 2
Data Science

The following is a set of one-dimensional points: {6, 12, 18, 24, 30, 42, 48}.

a) Please conduct single-link agglomerative hierarchical clustering, and

i) (30%) Please compute the similarity matrix after each merging step.

Given points: {6,12,18,24,30,42,48}

Step 1:

Similarity Matrix calculate the initial pairwise distances between each point:

	6	12	18	24	30	42	48
6	0						
12	6	0					
18	12	6	0				
24	18	12	6	0			
30	24	18	12	6	0		
42	36	30	24	18	12	0	
48	42	36	30	24	18	6	0

Merge 6 and 12 closest pair: {6,12} has a distance 6. Merge into a cluster {6,12}.

Updated Distance Matrix:

	{6,12}	18	24	30	42	48
{6,12}	0					
18	6	0				
24	12	6	0			
30	18	12	6	0		
42	30	24	18	12	0	
48	36	30	24	18	6	0

Merge 18 and {6,12} closest pair: the distance of 6. Merge to form cluster {6,12,18}.

	{6,12,18}	24	30	42	48
{6,12,18}	0				
24	6	0			
30	12	6	0		
42	24	18	6	0	
48	30	24	18	6	0

Merge 24 and {6,12,18} closest pair: the distance of 6. Merge to form cluster {6,12,18,24}.

	{6,12,18,24}	30	42	48
{6,12,18,24}	0			
30	6	0		
42	18	12	0	
48	24	18	6	0

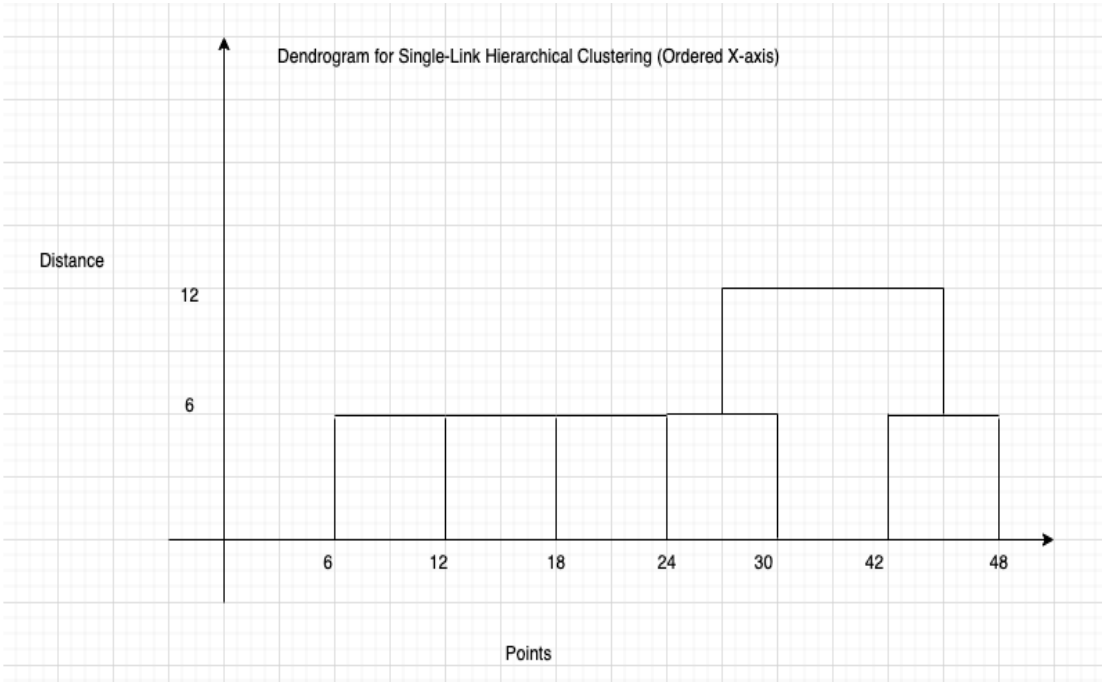
Merge 30 and {6,12,18,24} closest pair: the distance of 6. Merge to form cluster {6,12,18,24,30}.

	{6,12,18,24,30}	42	48
{6,12,18,24,30}	0		
42	12	0	
48	18	6	0

Final Merges Merge 42 and 48 with a distance of 6 into { 42 , 48 }. Merge {6,12,18,24,30} and {42,48} with a distance of 12 into a single cluster.

	{6,12,18,24,30}	{42,48}
{6,12,18,24,30}	0	
{42,48}	12	0

ii) (10%) Please show the final result by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.



iii) (10%) What are the two clusters produced by a single link?

For the data points {6,12,18,24,30,42,48}:

Cluster 1: {6,12,18,24,30}

Cluster 2: {42,48}

b) (40%) For each of the following sets of initial centroids, please create two clusters by assigning each point to the nearest centroid, and then calculate the total squared error for each set of two clusters. Please show both the clusters and the total squared error for each set of centroids.

i) {18,45}

We have these points {6,12,18,24,30,42,48} and assign them to the nearest centroid, then calculate total squared error.

Set 1: {18,45}:

For each point in {6,12,18,24,30,42,48}, calculate the distance to centroids 18 and 45, and assign the point to the nearest centroid:

Points closer to 18 will form Cluster 1. Points closer to 45 will form Cluster 2.

Cluster 1 (Centroid = 18): {6,12,18,24,30}

Cluster 2 (Centroid = 45): {42,48}

Calculate Total Squared Error: Cluster 1 : $(6-18)^2+ (12-18)^2+(18-18)^2+ (24-18)^2+ (30-18)^2 = 360$

Cluster 2: $(42-45)^2+ (48-45)^2 = 18$

The total of cluster 1 (TSE) is 378

ii) {15,40}

Cluster 1 (Centroid = 15): {6,12,18.24}

Cluster 2 (Centroid = 40): {30,42,48}

Calculate Total Squared Error:

Cluster 1: $(6-15)^2+ (12-15)^2+ (18-15)^2 +(24-15)^2 = 180$

Cluster 2: $(30-40)^2+(42-40)^2+(48-40)^2 = 168$

The total of cluster 1 (TSE) is 348

Final Results

Set 1 (Centroids = 18, 45): TSE = 378

Set 2 (Centroids = 15, 40): TSE = 348

c) (10%) Do both sets of centroids (given in part b) represent stable solutions; i.e., if the K-means algorithm was run on this set of points using the given centroids as the starting centroids, would there be any change in the clusters generated?

Set 1: {18,45}

Cluster 1 (Centroid = 18): {6,12,18,24,30}

Cluster 2 (Centroid = 45): {42,48}

Calculate New Centroids:

Cluster 1: (6+12+18+24+30)/5=18

New centroid for Cluster 2:

(42+48)/2=45

The centroids do not change upon recalculation, which indicates that Set 1 represents a stable solution.

Set 2: {15,40}

Cluster 1: (Centroid=15): {6,12,18,24}

Cluster 2 (Centroid=40): {30,42,48}

New centroid for Cluster 1: (6+12+18+24)/4=15

Cluster 2: (30+42+48)/3=40

The centroids remain the same after recalculating, indicating that Set 2 is also a stable solution.