

Workshop Week 8

COMP20008 2020S2

Clustering (& Linear regression)

Q1: Consider the 1-dimensional data set with 10 data points $\{1, 2, 3, \dots, 10\}$. Show the iterations of the k-means algorithm using Euclidean distance when $k = 2$, and the random seeds are initialized to $\{1, 2\}$.

- **Iteration 1** Data points: [1 2 3 4 5 6 7 8 9 10]
Assignments: [0, 1, 1, 1, 1, 1, 1, 1, 1, 1] Centroids: [1.0, 6.0]
 - 0 means 1 , 1 means cluster 2
- **Iteration 2** Data points: [1 2 3 4 5 6 7 8 9 10]
Assignments: [0, 0, 0, 1, 1, 1, 1, 1, 1, 1] Centroids: [2.0, 7.0]
- **Iteration 3** Data points: [1 2 3 4 5 6 7 8 9 10]
Assignments: [0, 0, 0, 0, 1, 1, 1, 1, 1, 1] Centroids: [2.5, 7.5]

Consider the 1-dimensional data set with 10 data points $\{1, 2, 3, \dots, 10\}$. Show the iterations of the k-means algorithm using Euclidean distance when $k = 2$, and the random seeds are initialized to $\{1, 2\}$.

- **Iteration 4** Data points: [1 2 3 4 5 6 7 8 9 10]
Assignments: [0, 0, 0, 0, 0, 1, 1, 1, 1, 1] Centroids: [3.0, 8.0]
- **Iteration 5** Data points: [1 2 3 4 5 6 7 8 9 10]
Assignments: [0, 0, 0, 0, 0, 1, 1, 1, 1, 1] Centroids: [3.0, 8.0]

Q2: Repeat Exercise 1 using agglomerative hierarchical clustering and Euclidean distance, with single linkage (min) criterion.

	1	2	3	4	5	6	7	8	9	10
1	0									
2	1	0								
3	2		0							
4	3			0						
5	4				0					
6	5					0				
7	6						0			
8	7							0		
9	8								0	
10	9									0

Initially, how many clusters do we have?



	1	2	3	4	5	6	7	8	9	10
1	0	1	2	3	4	5	6	7	8	9
2	1	0	1	2	3	4	5	6	7	8
3	2	1	0	1	2	3	4	5	6	7
4	3	2	1	0	1	2	3	4	5	6
5	4	3	2	1	0	1	2	3	4	5
6	5	4	3	2	1	0	1	2	3	4
7	6	5	4	3	2	1	0	1	2	3
8	7	6	5	4	3	2	1	0	1	2
9	8	7	6	5	4	3	2	1	0	1
10	9	8	7	6	5	4	3	2	1	0

Dissimilarity Matrix

Inter-point distance Matrix

Step1: Calculate Distances between every pair of observation: Euclidean Distance

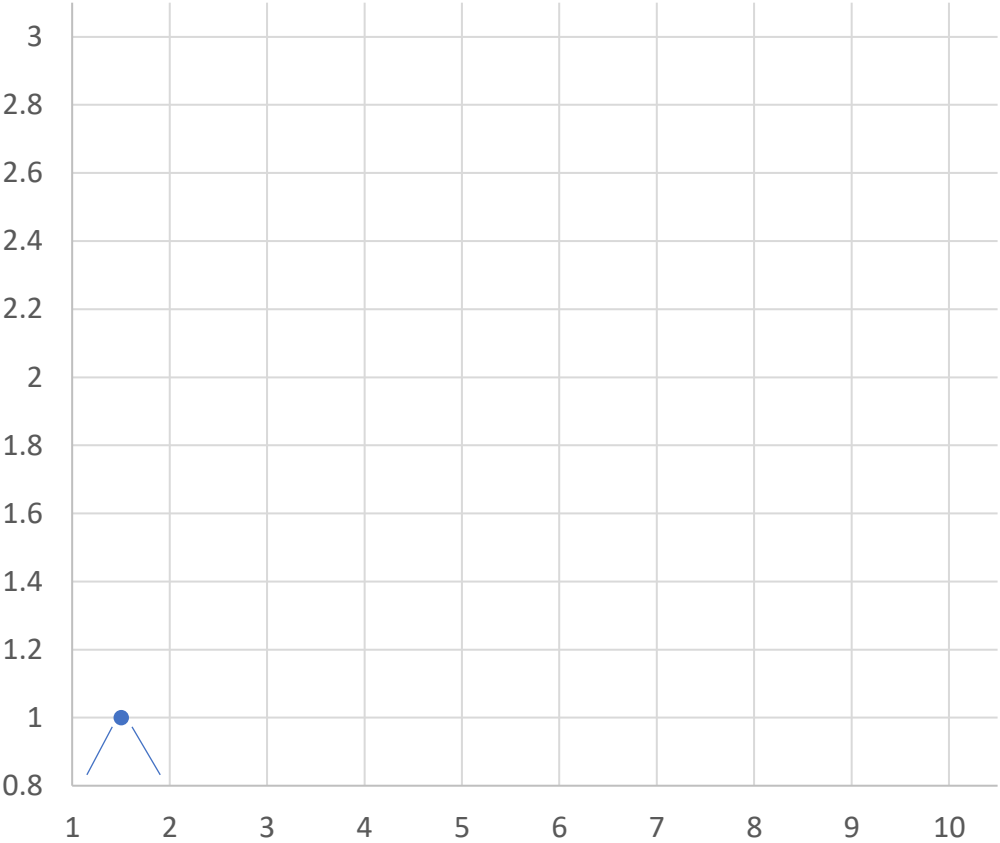
Dissimilarity Matrix

	1	2	3	4	5	6	7	8	9	10
1	0	1	2	3	4	5	6	7	8	9
2	1	0	1	2	3	4	5	6	7	8
3	2	1	0	1	2	3	4	5	6	7
4	3	2	1	0	1	2	3	4	5	6
5	4	3	2	1	0	1	2	3	4	5
6	5	4	3	2	1	0	1	2	3	4
7	6	5	4	3	2	1	0	1	2	3
8	7	6	5	4	3	2	1	0	1	2
9	8	7	6	5	4	3	2	1	0	1
10	9	8	7	6	5	4	3	2	1	0

Inter-point distance Matrix

Step 2: Choose the most similar two observations to merge (i.e. Closest)
(i.e. pair with the minimum distance in Dissimilarity Matrix)

Y-Values



Dendrogram Plot

X-axis → observations , Y-axis → distances

Dissimilarity Matrix

	1	2	3	4	5	6	7	8	9	10
1	0	1	2	3	4	5	6	7	8	9
2	1	0	1	2	3	4	5	6	7	8
3	2	1	0	1	2	3	4	5	6	7
4	3	2	1	0	1	2	3	4	5	6
5	4	3	2	1	0	1	2	3	4	5
6	5	4	3	2	1	0	1	2	3	4
7	6	5	4	3	2	1	0	1	2	3
8	7	6	5	4	3	2	1	0	1	2
9	8	7	6	5	4	3	2	1	0	1
10	9	8	7	6	5	4	3	2	1	0



	12	3	4	5	6	7	8	9	10
12	0	1							
3	1	0	1	2	3	4	5	6	7
4		1	0	1	2	3	4	5	6
5		2	1	0	1	2	3	4	5
6		3	2	1	0	1	2	3	4
7		4	3	2	1	0	1	2	3
8		5	4	3	2	1	0	1	2
9		6	5	4	3	2	1	0	1
10		7	6	5	4	3	2	1	0

Inter-point distance Matrix

Step 3: Update Dissimilarity Matrix: Calculate the distance between Cluster12 and all other observations (calculate linkage using min)

Dissimilarity Matrix

	1	2	3	4	5	6	7	8	9	10
1	0	1	2	3	4	5	6	7	8	9
2	1	0	1	2	3	4	5	6	7	8
3	2	1	0	1	2	3	4	5	6	7
4	3	2	1	0	1	2	3	4	5	6
5	4	3	2	1	0	1	2	3	4	5
6	5	4	3	2	1	0	1	2	3	4
7	6	5	4	3	2	1	0	1	2	3
8	7	6	5	4	3	2	1	0	1	2
9	8	7	6	5	4	3	2	1	0	1
10	9	8	7	6	5	4	3	2	1	0



	12	3	4	5	6	7	8	9	10
12	0	1	2	3	4	5	6	7	8
3	1	0	1	2	3	4	5	6	7
4	2	1	0	1	2	3	4	5	6
5	3	2	1	0	1	2	3	4	5
6	4	3	2	1	0	1	2	3	4
7	5	4	3	2	1	0	1	2	3
8	6	5	4	3	2	1	0	1	2
9	7	6	5	4	3	2	1	0	1
10	8	7	6	5	4	3	2	1	0

Inter-point distance Matrix

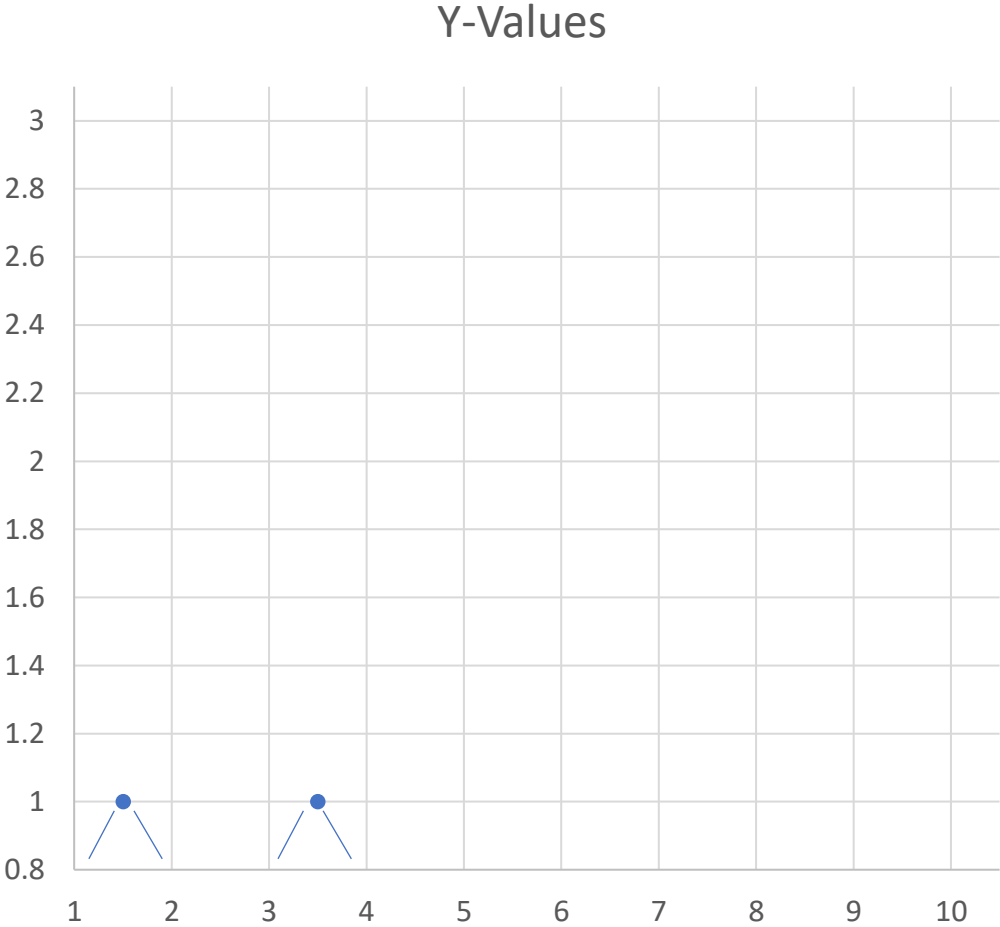
Step 3: Update Dissimilarity Matrix: Calculate the distance between Cluster12 and all other observations (calculate linkage using min)

How many clusters do we have now?

Updated Dissimilarity Matrix

	12	3	4	5	6	7	8	9	10
12	0	1	2	3	4	5	6	7	8
3	1	0	1	2	3	4	5	6	7
4	2	1	0	1	2	3	4	5	6
5	3	2	1	0	1	2	3	4	5
6	4	3	2	1	0	1	2	3	4
7	5	4	3	2	1	0	1	2	3
8	6	5	4	3	2	1	0	1	2
9	7	6	5	4	3	2	1	0	1
10	8	7	6	5	4	3	2	1	0

Updated distance Matrix



Repeat Step 2: Choose the most similar two observations to merge (i.e. Closest)
(i.e. pair with the minimum distance in Dissimilarity Matrix)

Dissimilarity Matrix

	12	3	4	5	6	7	8	9	10
12	0	1	2	3	4	5	6	7	8
3	1	0	1	2	3	4	5	6	7
4	2	1	0	1	2	3	4	5	6
5	3	2	1	0	1	2	3	4	5
6	4	3	2	1	0	1	2	3	4
7	5	4	3	2	1	0	1	2	3
8	6	5	4	3	2	1	0	1	2
9	7	6	5	4	3	2	1	0	1
10	8	7	6	5	4	3	2	1	0



	12	34	5	6	7	8	9	10
12	0		3	4	5	6	7	8
34		0						
5	3		0	1	2	3	4	5
6	4		1	0	1	2	3	4
7	5		2	1	0	1	2	3
8	6		3	2	1	0	1	2
9	7		4	3	2	1	0	1
10	8		5	4	3	2	1	0

Inter-point distance Matrix

Repeat Step 3: Update Dissimilarity Matrix: Calculate the distance between Cluster12 and all other observations (calculate single linkage using min)

Dissimilarity Matrix

	12	3	4	5	6	7	8	9	10
12	0	1	2	3	4	5	6	7	8
3	1	0	1	2	3	4	5	6	7
4	2	1	0	1	2	3	4	5	6
5	3	2	1	0	1	2	3	4	5
6	4	3	2	1	0	1	2	3	4
7	5	4	3	2	1	0	1	2	3
8	6	5	4	3	2	1	0	1	2
9	7	6	5	4	3	2	1	0	1
10	8	7	6	5	4	3	2	1	0



	12	34	5	6	7	8	9	10
12	0	1	3	4	5	6	7	8
34	1	0	1	2	3	4	5	6
5	3	1	0	1	2	3	4	5
6	4	2	1	0	1	2	3	4
7	5	3	2	1	0	1	2	3
8	6	4	3	2	1	0	1	2
9	7	5	4	3	2	1	0	1
10	8	6	5	4	3	2	1	0

Inter-point distance Matrix

Let's see some python code

Repeat Step 3: Update Dissimilarity Matrix: Calculate the distance between Cluster12 and all other observations (calculate linkage using min)