



Agglomerative Hierarchical Clustering

Points

A(1,2), B(2,3), C(6,7), D(7,8)

Step 1: Calculate Euclidean distance for computing initial distance matrix.

$$d(p, q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}$$

$$d(A, B) = \sqrt{(1-2)^2 + (2-3)^2} = \sqrt{1+1} = 1.41$$

$$d(A, C) = \sqrt{(1-6)^2 + (2-7)^2} = \sqrt{25+25} = 7.07$$

$$d(A, D) = \sqrt{(1-7)^2 + (2-8)^2} = \sqrt{36+36} = 8.49$$

$$d(B, C) = \sqrt{(2-6)^2 + (3-7)^2} = \sqrt{16+16} = 5.66$$

$$d(B, D) = \sqrt{(2-7)^2 + (3-8)^2} = \sqrt{25+25} = 7.07$$

$$d(C, D) = \sqrt{(6-7)^2 + (7-8)^2} = \sqrt{1+1} = 1.41$$

Distance Matrix

	A	B	C	D
A	0	1.41	7.07	8.49
B	1.41	0	5.66	7.07
C	7.07	5.66	0	1.41
D	8.49	7.07	1.41	0

Step 2

Merging ①

Smallest distance = 1.41

occurs between A-B and C-D

Merge A & B first

Available clusters

$\{AB\}, \{C\}, \{D\}$

Step 3

Update distances using any one linkage method. Here we are using single linkage.

$$\begin{aligned} \text{Distance between } \{AB\} \text{ and } C &= \text{minimum}(d(A,C), d(B,C)) \\ &= \min(7.07, 5.66) = 5.66 \end{aligned}$$

$$\begin{aligned} \text{Distance between } \{A,B\} \text{ and } \{D\} &= \text{minimum}(d(A,D), d(B,D)) \\ &= \min(8.49, 7.07) = 7.07 \end{aligned}$$

Update distance matrix

	AB	C	D
AB	0	5.66	7.07
C	5.66	0	(1.41)
D	7.07	(1.41)	0

Step 4.

Next merging operation

Smallest distance = 1.41 (C-D)

Merge C & D

clusters AB, CD.

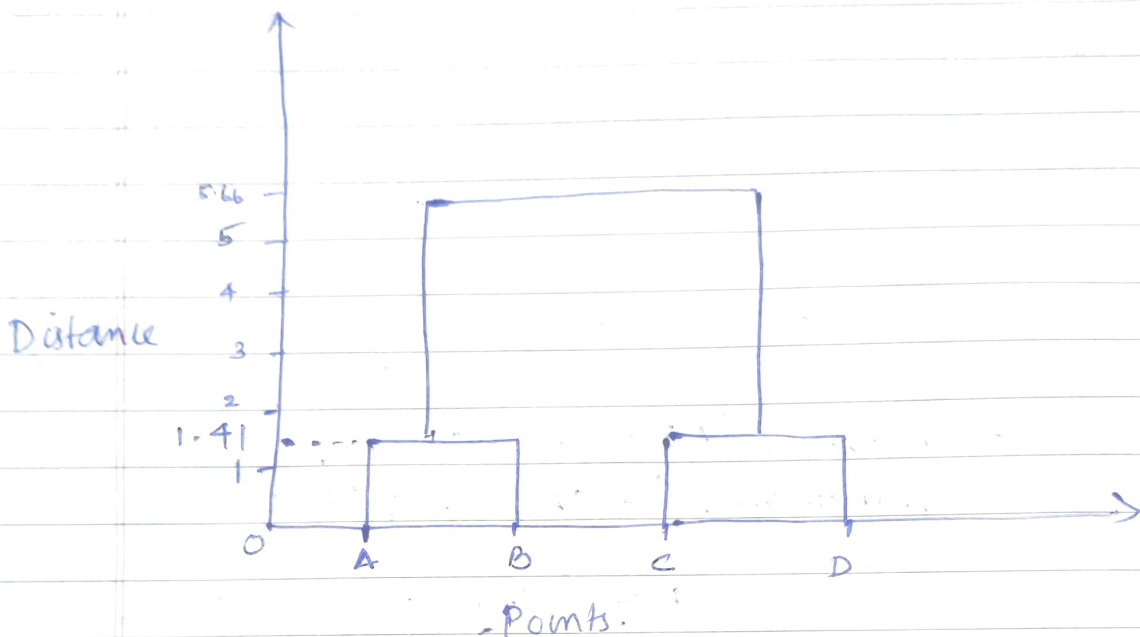
Step 5 Final merge.

$$\begin{aligned} \text{Distance between } AB \text{ &\& } CD &= \min(d(A,C), d(A,D), \\ &d(B,C), d(B,D)) \\ &= \min(7.07, 8.49, 5.66, 7.07) = 5.66 \end{aligned}$$

Date



Join $\{AB\}$ and $\{C, D\}$ at 5.66 distance

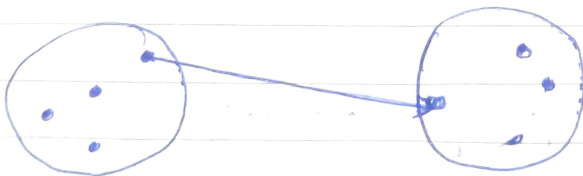


Linkage Methods

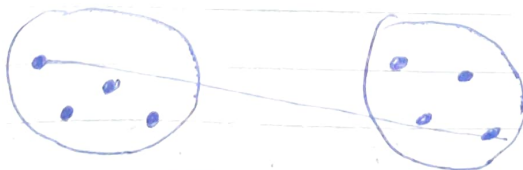
→ tells us how to compute the distance between two clusters.

a) Single Linkage

→ we took the closest pair by calculating minimum distance



b) Complete Linkage

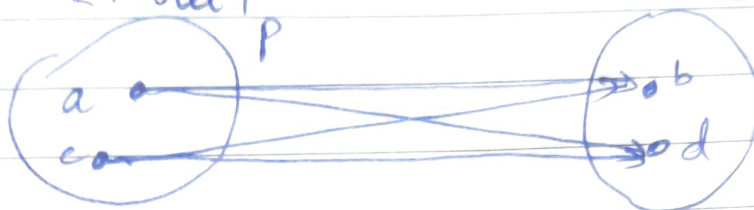


maximum distance b/w elements in clusters

Date _____

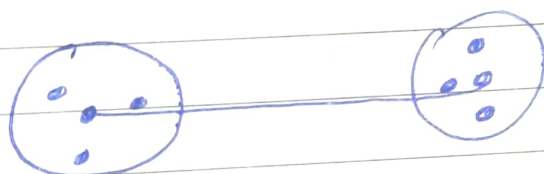


c) Average linkage - avg of the distance of the pairs



$$d(P, Q) = \frac{1}{|P| \cdot |Q|} \sum_{i \in P} \sum_{j \in Q} d(i, j)$$

d) Centroid Method - combining clusters with min. distance b/n the centroids of 2 clusters



Average linkage e.g.

$$A - C = 7.07$$

$$A - D = 8.49$$

$$B - C = 5.66$$

$$B - D = 7.07$$

$$d((AB), (CD)) = \frac{7.07 + 8.49 + 5.66 + 7.07}{4}$$

$$= \underline{\underline{7.07}}$$

Hierarchical Clustering (Divisive)

Data set

A(1,2) B(2,1) C(6,5) D(7,6)

step 1:

Start with all points in one cluster

Initial cluster: {A, B, C, D}

step 2: Pairwise Euclidean Distances

$$d(p, q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}$$

$$d(A, B) = \sqrt{(1-2)^2 + (2-1)^2} = 1.41$$

$$d(A, C) = \sqrt{(1-6)^2 + (2-5)^2} = 5.83$$

$$d(A, D) = \sqrt{(1-7)^2 + (2-6)^2} = 7.21$$

$$d(B, C) = \sqrt{(2-6)^2 + (1-5)^2} = 5.66$$

$$d(B, D) = \sqrt{(2-7)^2 + (1-6)^2} = 7.07$$

$$d(C, D) = \sqrt{(6-7)^2 + (5-6)^2} = 1.41$$

Closest pairs are (A, B), (C, D) both at 1.41

So the algorithm decides the 1st split as

$$\{A, B, C, D\} \Rightarrow \{A, B\}, \{C, D\}$$

Date

Step 3: Inspect each cluster for further splitting.

Now we have 2 clusters $\{A, B\}$ and $\{C, D\}$

cluster $\{A, B\}$

Distance between A & B is quite small.

If we want to split then split $\{A\}, \{B\}$

cluster $\{C, D\}$

$\Rightarrow \{C\}, \{D\}$

Dendrogram

