700740862
Veera Venkata Satyanarayana
VXB08620

Assignment-6

(Q.1) Find out clustering representations, & dendrograms using single, complete, and average link proximity function in Hierarchial clustering technique?

Point	X-Coordinate	Y-Cordinate
P1	0.04005	0.5306
P2	0.2148	0.3854
P3	0.3457	0.3156
P4	0.2652	0.1875
P5	0.0789	0.4139
P6	0.0548	0.3022

Table-1

X-Y co-ordinate.

Distance Matrix.

	P1	P2	P3	P4	P5	P6
P1	0.000	0.2357	0.2218	0.3688	0.3421	0.2347
P2	0.2357	0.000	0.1483	0.2042	0.1388	0.2540
P3	0.2218	0.1403	0.000	0.1513	0.2843	0.1100
P4	0.3688	0.2042	0.1513	0.000	0.2932	0.2216
P5	0.3421	0.1388	0.2843	0.2932	0.000	0.3941
P6	0.2347	0.2540	0.1100	0.2216	0.3941	0.000

Table-2

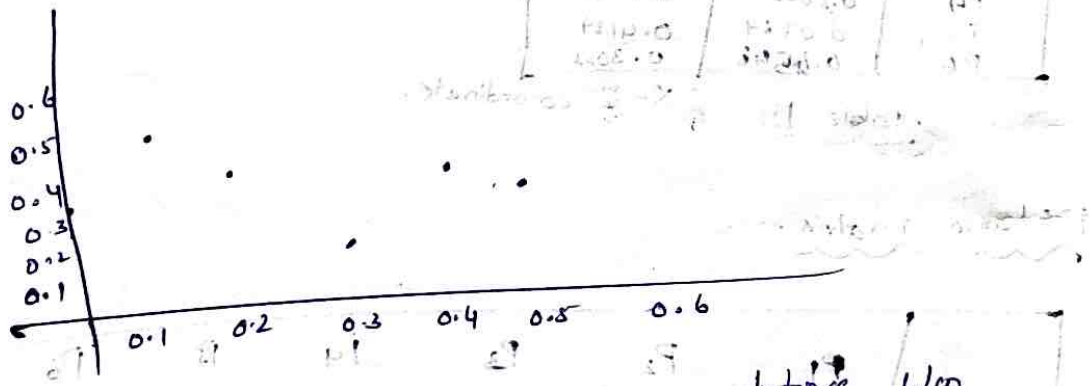
By single link :-

* For single link hirarchial clustering, the proximity of two clusters is minimum of the distance blw any two points in 2 different clusters.

* The single link technique is good for non-elliptical shapes, but sensitive to noise & outliers.

* Applying single link techniques to our example database set of six points.

set of six 2 dimensional points.



→ from table 1, we can observe distance blw P_3 & P_6 is 0.11.

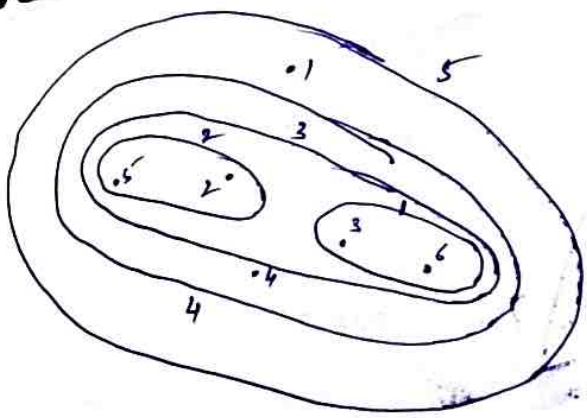
→ the height at which two clusters are merged can be represented as distance blw two clusters.

distance blw clusters $\{3,6\}$ & $\{2,5\}$ is given by

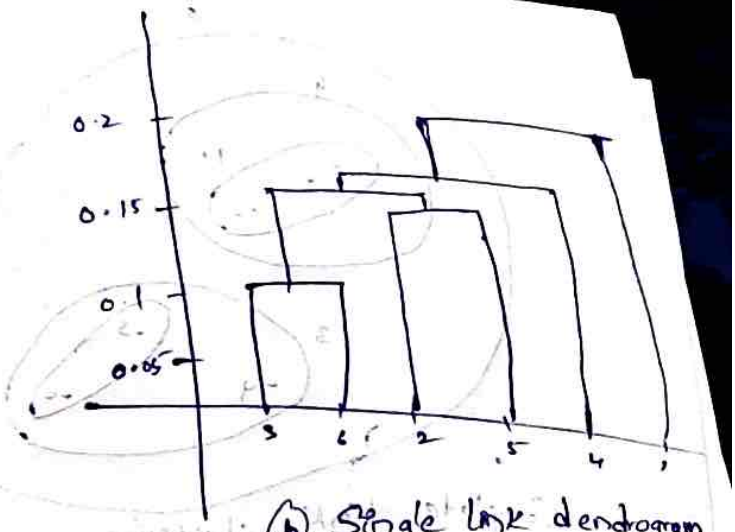
$dist(\{3,6\}, \{2,5\}) = \min(dist(3,2), dist(6,2), dist(3,5) * dis(6,5))$

$\Rightarrow \min(0.15, 0.25, 0.28, 0.39)$

$\Rightarrow 0.15$



Single link clustering



(b) Single link dendrogram

Complete link.

→ In Complete link of hierarchial clustering, the proximity of two clusters is defined as the maximum of the distances blw any two points in two different clusters.

→ Complete link is less susceptible to noise & outliers, but it can break large clusters & its favours globular shapers.

→ Below fig shows results of Applying Max to the sample data set of six points.

→ Here points 3 & 6 are merged first. {3,6} is merged with {4} instead of {2,5} or {1}. this is Because

$$dist(\{3,6\},\{4\}) = \max(dist(3,4), dist(6,4))$$

$$= \max(0.15, 0.22)$$

$$= 0.22$$

$$dist(\{3,6\},\{2,5\}) = \max(dist(3,2), dist(6,2), dist(3,5), dist(6,5)).$$

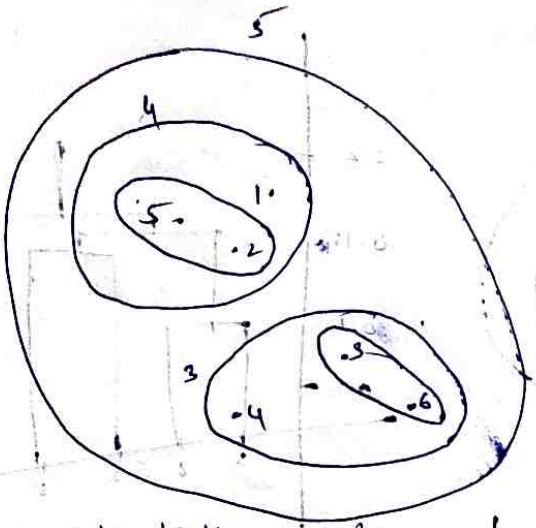
$$= \max(0.15, 0.25, 0.28, 0.39)$$

$$= 0.39.$$

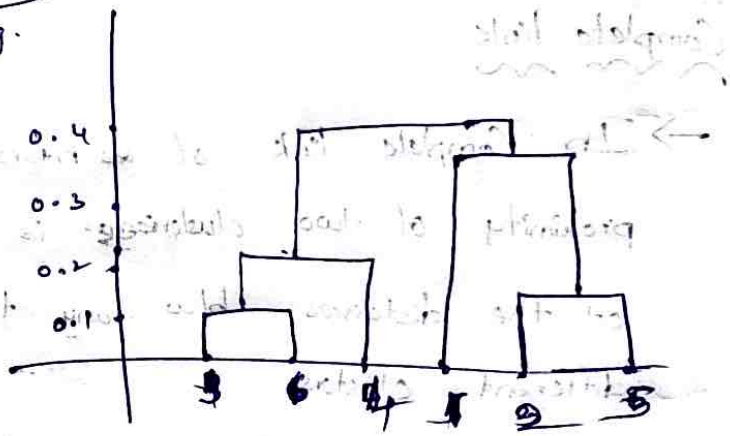
$$dist(\{3,6\},\{1\}) = \max(dist(3,1), dist(6,1))$$

$$= \max(0.22, 0.23)$$

$$= 0.23.$$



Complete link clustering.



Complete link dendrogram

Average link:

Below figure shows results after Applying the group Average approach to sample data of six points.

→ we calculate the distance blw some clusters.

→ proximity ⇒ $\text{proximity}(c_i, c_j) = \frac{\sum_{x \in c_i, y \in c_j} \text{proximity}(x,y)}{m_i \times m_j}$

$\text{dist}(\{3,6,4\}, \{1\}) = (0.22 + 0.37 + 0.23)/(3\times1)$

$= 0.28$

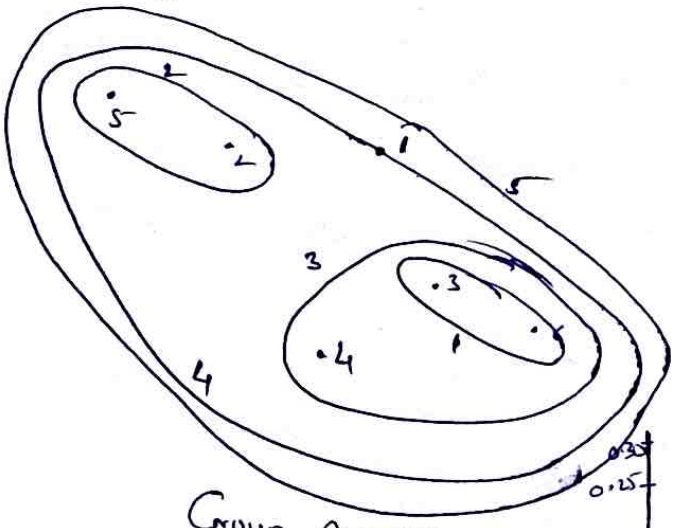
$\text{dist}(\{2,5\}, \{1\}) = (0.24 + 0.34)/(2\times1)$

$= 0.29$

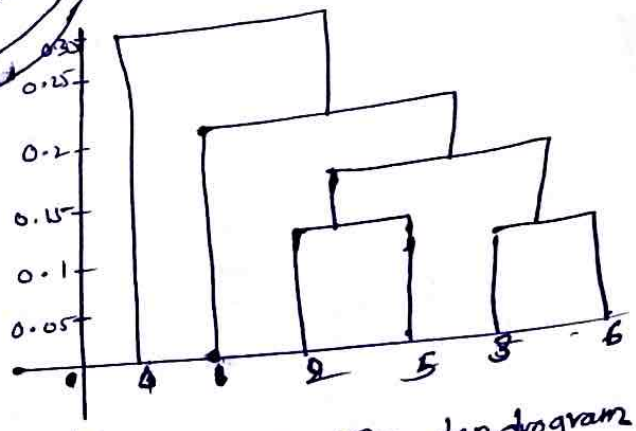
$\text{dist}(\{3,6,4\}, \{2,5\}) = (0.15 + 0.28 + 0.25 + 0.39 + 0.20 + 0.29)/(3\times2)$

$= 0.26$

Here Because dist $\{\{3,6,4\}, \{2,5\}\}$ is smaller than dist $(\{3,6,4\}, \{1\})$ and dist $\{(2,5), \{1\}\}$ clusters $\{3,6,4\}$ and $\{2,5\}$, are merged at the fourth stage.



Group Average clustering.



(b) group Average dendrogram

→ Average version of hierarchical clustering, the proximity of two clusters is defined as the average pairwise proximity among all pairs of points in the different clusters.

proximity proximity (C_i, C_j) of clusters C_i and C_j which are of size m_i and m_j respectively is

$$\text{proximity}(C_i, C_j) = \frac{\sum_{x \in C_i} \text{Proximity}(x, y)}{m_i \times m_j}$$