700740867
Sai Pranitha Gutlikonda

Assignment-6 11/16/2022

Question-1

Find out clustering representations & Dendrogram using single, complete and Average link proximity function in Hierarchial clustering technique.

| point | x-coordinate | y co-ordinate |
|---|---|---|
| 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 | 0.4548 | 0.3022 |

table1

X-Y coordinates.

Distance matrix:

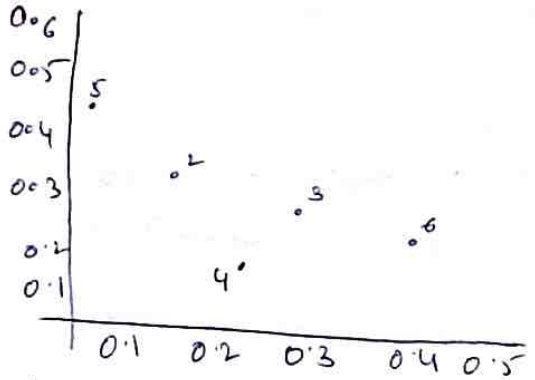
| | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $P_6$ |
|---|---|---|---|---|---|---|
| $P_1$ | 0.00 | 0.2357 | 0.2218 | 0.3688 | 0.3421 | 0.2347 |
| $P_2$ | 0.2357 | 0.000 | 0.1483 | 0.2042 | 0.1388 | 0.2540 |
| $P_3$ | 0.2218 | 0.1483 | 0.000 | 0.1513 | 0.2843 | 0.1100 |
| $P_4$ | 0.3688 | 0.2042 | 0.1513 | 0.000 | 0.2932 | 0.2216 |
| $P_5$ | 0.3421 | 0.1388 | 0.2843 | 0.2932 | 0.000 | 0.3921 |
| $P_6$ | 0.2347 | 0.2540 | 0.1100 | 0.2216 | 0.3921 | 0.000 |

table2

By Single Link:-

- For Single Link Hierarchial clustering, the proximity of two clusters is minimum of the distance between any two points in 2 different cluster.
- The single link technique is good for non elliptical shapes, but sensitive to noise & outliers

• Applying Single Link technique to our example data set of six point set of six 2 dimensional points.




→ From table 1, we can observe distance between $P_2$ & $P_6$ is 0.11

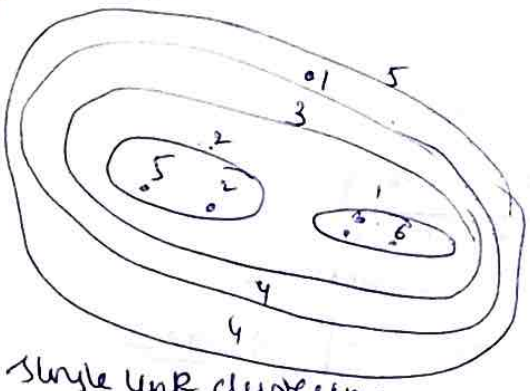
→ The height at which two clusters are merged can be represented as distance between two clusters.

distance between cluster {3,6} & {2,5} is given by

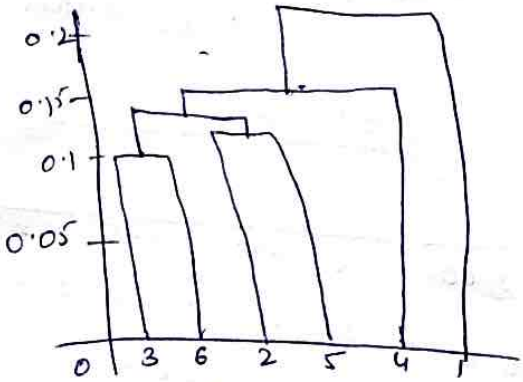
dist({3,6}, {2,5}) = Min (dist(3,2), dist(6,2), dist(3,5), dist(6,5))

= Min (0.15, 0.25, 0.28, 0.39)

= 0.15




Single link clustering




(b) single link dendogram

## Complete link

→ In complete link of hierarchical clustering, the proximity of two clusters is defined as the maximum of the distance between any two points in two different clusters

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

→ Below is shown results of applying max to the sample data set of six points

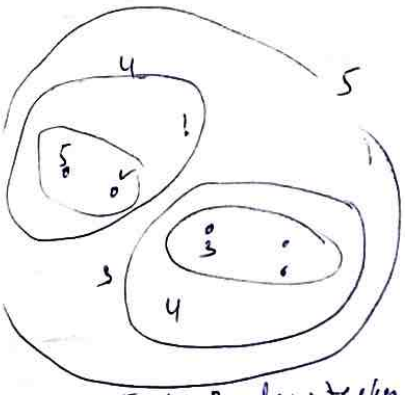
→ three points 3 and 6 are merged first.

{3,6} is merged with {4} instead of {2,5} or {1} this is because

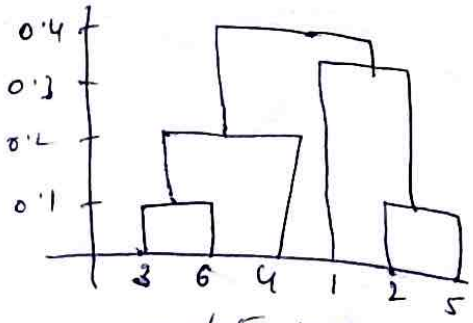
dist({3,6,4}, {4}) = max(dist(3,4), dist(6,4))
= max(0.15, 0.22)
= 0.22

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({3,6,4}, {1}) = max(dist(3,1), dist(6,1))
= max(0.22, 0.23) = 0.23



complete link clusters



complete link dendrogram

Average link:-

Below figure shows results after applying the group average approach to sample data of six points.

Here calculate the distance between some clusters.

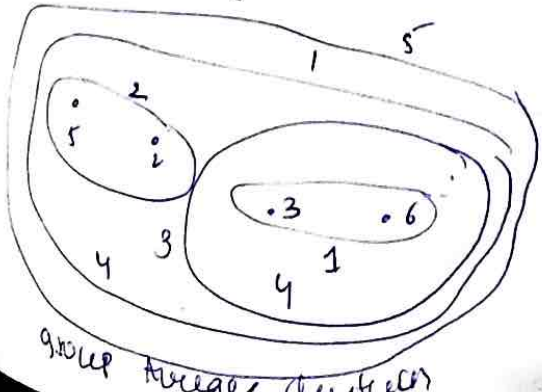
proximity = proximity $(C_i, C_j) = \frac{\sum_{x \in C_i, y \in C_j} \text{proximity}(x,y)}{m_i \times m_j}$

dist({3,6,4}, {1}) = (0.22 + 0.37 + 0.23)/(3×1)
= 0.28

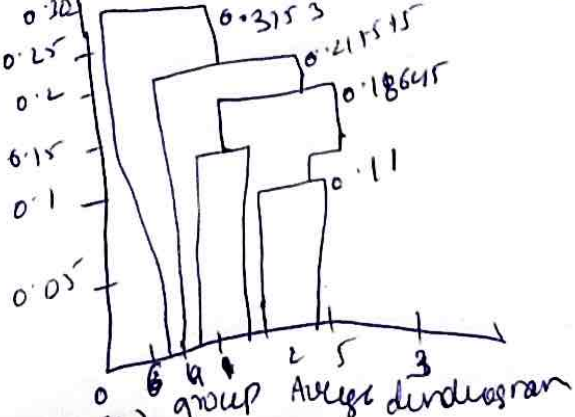
dist({2,5}, {1}) = (0.24 + 0.34)/(2×1) = 0.29

dist({3,6,4}, {2,5}) = (0.15 + 0.28 + 0.25 + 0.39 + 0.20 + 0.29)/(3×2)
= 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 clusters



(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 cluster $C_i$ and $C_j$ which are of size $m_i$ and $m_j$ respectively is

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