Aravya Mendem
700765926

Assignment -4

Part A : Calculation

1) Given :

Point	x	y
P_1	0.4	0.5
P_2	0.2	0.3
P_3	0.1	0.08
P_4	0.21	0.12
P_5	0.6	0.16
P_6	0.33	0.28
P_7	0.11	0.15

	P_1	P_2	P_3	P_4	P_5	P_6	P_7
P_1	0						
P_2	0.282	0					
P_3	0.516	0.241	0				
P_4	0.424	0.180	0.117	0			
P_5	0.394	0.423	0.506	0.392	0		
P_6	0.230	0.131	0.305	0.2	0.295	0	
P_7	0.454	0.175	0.070	0.104	0.49	0.356	0

Distance :

$(P_1, P_2) = \sqrt{(0.2)^2 + (0.2)^2} = 0.2\sqrt{2} = 0.282$

$(P_1, P_3) = \sqrt{(0.4-0.1)^2 + (0.5-0.08)^2} = \sqrt{(0.3)^2 + (0.42)^2} = 0.516$

$(P_1, P_4) = \sqrt{(0.4-0.21)^2 + (0.5-0.12)^2} = 0.424$

$(P_1, P_5) = \sqrt{(0.4-0.6)^2 + (0.5-0.16)^2} = 0.394$

$(P_1, P_6) = \sqrt{(0.4-0.33)^2 + (0.5-0.28)^2} = 0.230$

$(P_1, P_7) = \sqrt{(0.4-0.11)^2 + (0.5-0.15)^2} = 0.454$

$(P_2, P_3) = \sqrt{(0.2-0.1)^2+(0.3-0.08)^2} = 0.241$

$(P_2, P_4) = \sqrt{(0.2-0.21)^2+(0.3-0.12)^2} = 0.1803$

$(P_2, P_5) = \sqrt{(0.2-0.6)^2+(0.3-0.16)^2} = 0.4237$

$(P_2, P_6) = \sqrt{(0.2-0.33)^2+(0.3-0.28)^2} = 0.131$

$(P_2, P_7) = \sqrt{(0.2-0.11)^2+(0.3-0.15)^2} = 0.175$

$(P_3, P_4) = \sqrt{(0.1-0.21)^2+(0.08-0.12)^2} = 0.117$

$(P_3, P_5) = \sqrt{(0.1-0.6)^2+(0.08-0.16)^2} = 0.506$

$(P_3, P_6) = \sqrt{(0.1-0.33)^2+(0.08-0.28)^2} = 0.305$

$(P_3, P_7) = \sqrt{(0.1-0.11)^2+(0.08-0.15)^2} = 0.070$

$(P_4, P_5) = \sqrt{(0.21-0.6)^2+(0.12-0.16)^2} = 0.392$

$(P_4, P_6) = \sqrt{(0.21-0.33)^2+(0.12-0.28)^2} = 0.2$

$(P_4, P_7) = \sqrt{(0.21-0.11)^2+(0.12-0.15)^2} = 0.104$

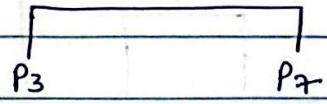
$(P_5, P_6) = \sqrt{(0.6-0.33)^2+(0.16-0.28)^2} = 0.295$

$(P_5, P_7) = \sqrt{(0.6-0.11)^2+(0.16-0.15)^2} = 0.49$

$(P_6, P_7) = \sqrt{(0.33-0.11)^2+(0.28-0.15)^2} = 0.256$

MIN LINK :

closest pair is (P_3, P_7) cluster A.



$d(A, P_1) = \min[d(P_1, P_3), d(P_1, P_7)]$

$= 0.516, 0.454 = 0.454$

$d(A, P_2) = \min[d(P_3, P_2), d(P_7, P_2)]$

$= 0.175$

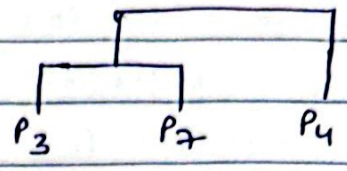
$d(A, P_4) = 0.117$

$d(A, P_5) = 0.49$

$d(A, P_6) = 0.2$

Smallest is $d(A, P_4) = 0.117$

Cluster B



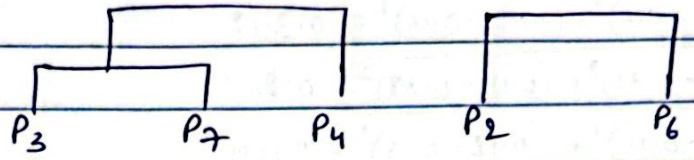
$d(B, P_1) = \min[(P_3, P_4, P_7), P_1] = 0.424$

$d(B, P_2) = \min[(P_3, P_4, P_7), P_2] = 0.175$

$d(B, P_5) = \min[(P_3, P_4, P_7), P_5] = 0.392$

$d(B, P_6) = \min[(P_3, P_4, P_7), P_6] = 0.2$

$d(P_2, P_6) = 0.131$ is minimum

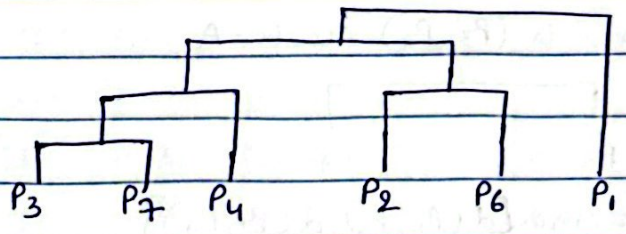


Cluster C $[P_3, P_7, P_4, P_2, P_6]$

$d(C, P_1) = 0.230$

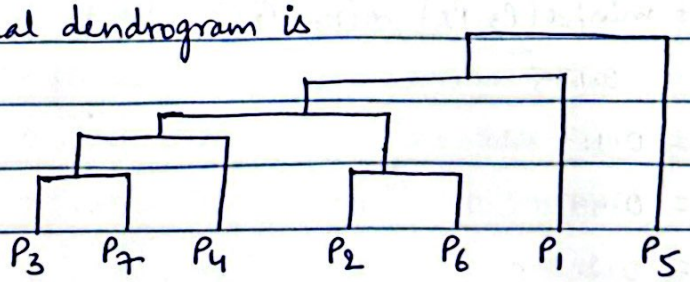
$d(C, P_5) = 0.295$

P_1 is the least value.

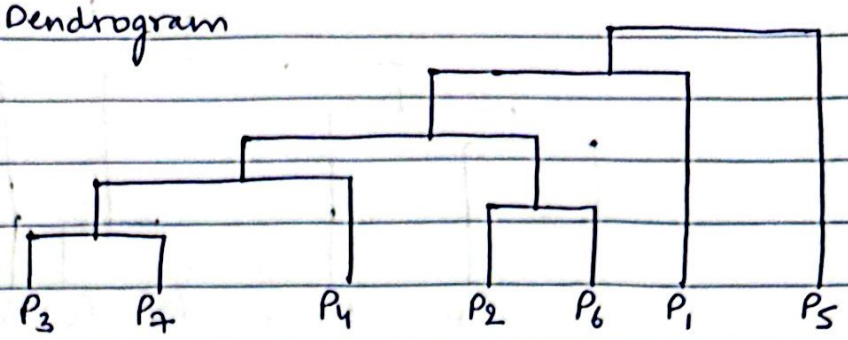


Next P_5

Final dendrogram is

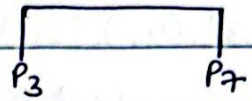


MIN Dendrogram



Average link

Closest pair (P_3, P_7) cluster A



$$d(A, P_1) = \frac{0.516 + 0.454}{2} = \frac{0.97}{2} = 0.485$$

$$d(A, P_2) = \frac{0.282 + 0.175}{2} = 0.2285$$

$$d(A, P_4) = \frac{0.117 + 0.104}{2} = 0.1105$$

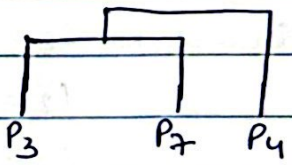
$$d(A, P_5) = \frac{0.506 + 0.49}{2} = 0.498$$

$$d(A, P_6) = \frac{0.305 + 0.256}{2} = 0.2805$$

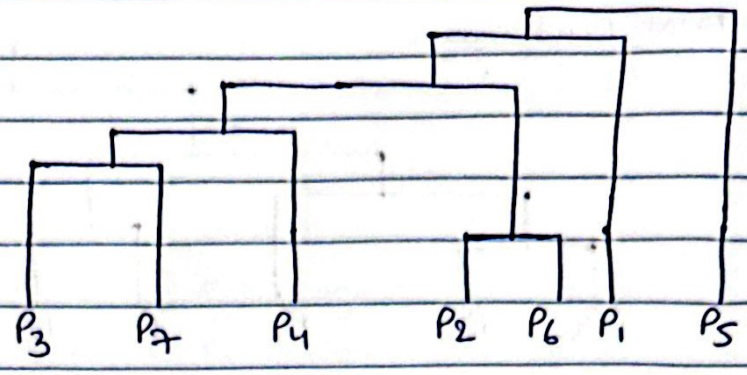
~~$d(A, P_4) = \frac{0.282 + 0.175}{2} = 0.2805$~~

$d(A, P_4)$ is least

Cluster B



After going through all the vertices we get the final dendrogram.



2) Given :

Points = (2,1), (3,1), (3,3), (4,1), (5,1), (6,7), (1,3), (2,5)

k = 3

Centroid 1 = (2,1)

Centroid 2 = (4,1)

Centroid 3 = (5,1)

Euclidean distance = $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

Point (2,1)

Distance to $C_1 = 0$

$C_2 = 2$ Assign to C_1

$C_3 = 3$

Point (3,1)

Distance to $C_1 = 1$

$C_2 = 1$ Assign to C_1

$C_3 = 2$

Point (3,3)

Distance to $C_1 = \sqrt{5} = 2.236$

$C_2 = \sqrt{5} = 2.236$ Assign to C_1

$C_3 = \sqrt{8} = 2.828$

Point (4,1)

Distance to $C_1 = 2$

$C_2 = 0$ Assign to C_2

$C_3 = 1$

Point (5, 1)

Distance to $C_1 = 3$

$C_2 = 1$ Assign to C_3

$C_3 = 0$

Point (6, 7)

Distance to $C_1 = \sqrt{52} = 7.211$

$C_2 = \sqrt{40} = 6.324$ Assign to C_3

$C_3 = \sqrt{37} = 6.082$

Point (1, 3)

Distance to $C_1 = \sqrt{5} = 2.236$

$C_2 = \sqrt{13} = 3.605$ Assign to C_1

$C_3 = \sqrt{20} = 4.472$

Point (2, 5)

Distance to $C_1 = 4$

$C_2 = \sqrt{20} = 4.472$ Assign to C_1

$C_3 = 5$

New centroids :

$$C_1 = \left(\frac{2+3+3+1+2}{5} , \frac{1+1+3+3+5}{5} \right)$$

$$= (2.2, 2.6)$$

$C_2 = (4, 1)$

$C_3 = (5.5, 4)$

Part B : Short-Answer Questions

1) a) Agglomerative hierarchical clustering :
A Bottom-up approach. Start with each data point as its own cluster and iteratively merge the two closest clusters (by some linkage rule) until all points are in a single cluster or a stopping condition is met. The sequence of merges forms a dendrogram.

b) Divisive hierarchical clustering :
Start with all points in one cluster and iteratively split clusters (often using a flat clustering method or heuristics) until each point is alone or a stopping condition is reached. Splits are chosen to maximize separation within the cluster that's being divided.

c) Agglomerative is more common because it is :
* Simpler to implement
* Requires fewer decisions (only merging, not splitting)
* Computationally easier and more stable in practice.

2) a) Inter-cluster distance should be maximized. The reason can be specified as the clusters should be far apart so they are clearly separated.

b) Intra-cluster distance should be minimized. The reason can be specified as points in the same cluster should be close to each other (high similarity).

3) a) * Single link (minimum/nearest neighbor) : Distance between the two closest points in the clusters.

* Complete link (maximum/farthest neighbor) : Distance between the two farthest points in the clusters.
* Average link : Average pairwise distance between all points across two clusters.

b) Single-link clustering :
* strength : Good at finding long, chain-shaped clusters; flexible with shapes.
* Weakness : Sensitive to noise/outliers and can cause "chaining effect".

4) a) Tokenization role : Tokenization splits raw text into smaller units (tokens), typically words or subwords; it is the first preprocessing step for most NLP tasks.
Example : Converting "Don't stop!" → tokens ["Do", "n't", "stop", "!"] (or using word tokens ["Don't", "stop", "!"] depending on tokenizer).

b) Speed : Stemming is generally faster (rule-based chopping of word endings; e.g., Porter stemmer).
Accuracy : Lemmatization is more accurate semantically because it returns valid dictionary lemmas using morphological analysis and POS info (e.g., "better" → "good" with lemmatization), whereas stemming may produce nonwords ("running" → "run" vs "runni" depending on stemmer). So stemming is faster; lemmatization is more accurate.

5) a) Word sense Ambiguity :
A word has multiple meanings, and context decides which is correct.

Example: "bank" = river bank or financial bank.

b) Pronoun reference Ambiguity:
Pronouns like "he/she/it/they" may refer to multiple possible nouns. Models get confused because it is unclear which entity the pronoun attaches to.
Example: "Alex told Jordan that he would help." Who is "he"? Alex or Jordan? Ambiguity confuses the model if insufficient context or world knowledge exists.

6) a) Because POS tags depend on sentence context – a token's correct tag often depends on neighboring tokens and global syntactic structure. Predicting each token in isolation ignores these dependencies and will mis-tag ambiguous tokens (e.g., "record" can be noun or verb).

b) In the sentence "Time flies like an arrow," whether "flies" is tagged as noun or verb depends on surrounding words; similarly, deciding that "like" is a preposition vs verb depends on the parse of the entire sentence. Another example: Subject–verb agreement (choosing singular vs plural verb form depends on the subject token's number).