CS5710 Machine Learning

Fall 2025 - Home Assignment 4.

Student name: Manikanth Reddy Devarapalli

Part A: Calculation

Q1. Find the cluster using the Average and MIN technique. Use Euclidean distance to build the complete distance matrix, updated the distance matrix to the final step and draw the dendrogram for each.

	X	Y
P1	0.4	0.5
P2	0.2	0.3
P3	0.1	0.08
P4	0.21	0.12
P5	0.6	0.16
P6	0.33	0.28
P7	0.11	0.15

ANS :

Step 1: Compute the Euclidean distance matrix

Euclidean distance between two points $(x_1, y_1)$and $(x_2, y_2)$is:

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

We calculate pairwise distances:

1. $d(P1, P2) = \sqrt{(0.4 - 0.2)^2 + (0.5 - 0.3)^2} = \sqrt{0.04 + 0.04} = \sqrt{0.08} \approx 0.283$
2. $d(P1, P3) = \sqrt{(0.4 - 0.1)^2 + (0.5 - 0.08)^2} = \sqrt{0.09 + 0.1764} \approx 0.53$
3. $d(P1, P4) = \sqrt{(0.4 - 0.21)^2 + (0.5 - 0.12)^2} = \sqrt{0.0361 + 0.1444} \approx 0.413$
4. $d(P1, P5) = \sqrt{(0.4 - 0.6)^2 + (0.5 - 0.16)^2} = \sqrt{0.04 + 0.1156} \approx 0.379$
5. $d(P1, P6) = \sqrt{(0.4 - 0.33)^2 + (0.5 - 0.28)^2} = \sqrt{0.0049 + 0.0484} \approx 0.226$
6. $d(P1, P7) = \sqrt{(0.4 - 0.11)^2 + (0.5 - 0.15)^2} = \sqrt{0.0841 + 0.1225} \approx 0.465$

	P1	P2	P3	P4	P5	P6	P7
P1	0	0.283	0.53	0.413	0.379	0.226	0.465
P2	0.283	0	0.242	0.188	0.452	0.13	0.203
P3	0.53	0.242	0	0.11	0.5	0.257	0.07
P4	0.413	0.188	0.11	0	0.395	0.159	0.1
P5	0.379	0.452	0.5	0.395	0	0.36	0.495
P6	0.226	0.13	0.257	0.159	0.36	0	0.252
P7	0.465	0.203	0.07	0.1	0.495	0.252	0

Step 2: Single Linkage (MIN)

Single linkage: distance between two clusters = **minimum distance between any pair of points in the clusters**.

1. Find smallest distance: **0.07 → P3 & P7**, merge as **C1 = {P3, P7}**
2. Update distances (minimum distance to cluster C1):

$$d(C1, P4) = \min\left(d(P3, P4), d(P7, P4)\right) = \min(0.11, 0.1) = 0.1$$
$$d(C1, P2) = \min\left(d(P3, P2), d(P7, P2)\right) = \min(0.242, 0.203) = 0.203$$

3. Next smallest: **0.1 → C1 & P4**, merge **C2 = {P3,P4,P7}**
4. Next smallest: **0.13 → C2 & P6**? Wait check distances:

$$d(C2, P6) = \min\left(d(P3, P6), d(P4, P6), d(P7, P6)\right) = \min(0.257, 0.159, 0.252) = 0.159$$

Next smallest distance: **0.159 → C2 & P6**, merge **C3 = {P3,P4,P6,P7}**

5. Next: **0.188 → P2 & P4** (but P4 now in C3), so distance:

$$d(C3, P2) = \min\left(d(P3, P2), d(P4, P2), d(P6, P2), d(P7, P2)\right)$$
$$= \min(0.242, 0.188, 0.13, 0.203) = 0.13$$

Merge **C4 = {P2,P3,P4,P6,P7}**

6. Next: smallest distance to remaining P1,P5:

$$d(C4, P1) = \min\left(d(P1, P2), d(P1, P3), d(P1, P4), d(P1, P6), d(P1, P7)\right)$$
$$= \min(0.283, 0.53, 0.413, 0.226, 0.465) = 0.226$$

Merge **C5 = {P1,P2,P3,P4,P6,P7}**

7. Finally merge **P5 → C6** at **0.36**

- Single linkage dendrogram: merges roughly in order:

1. P3+P7 (0.07)
2. C1+P4 (0.1)
3. C2+P6 (0.159)
4. C3+P2 (0.13)
5. C4+P1 (0.226)
6. C5+P5 (0.36)

Step 3: Average Linkage

- **Average linkage**: distance between two clusters = **average of distances between all pairs of points in the clusters**.

1. Smallest distance between points: **P3 & P7 → merge C1={P3,P7}**
2. Distance from C1 to others = **average distances**:

$$d(C1,P4) = \frac{d(P3,P4) + d(P7,P4)}{2} = \frac{0.11 + 0.1}{2} = 0.105$$
$$d(C1,P2) = \frac{0.242 + 0.203}{2} = 0.2225$$
$$d(C1,P6) = \frac{0.257 + 0.252}{2} = 0.2545$$

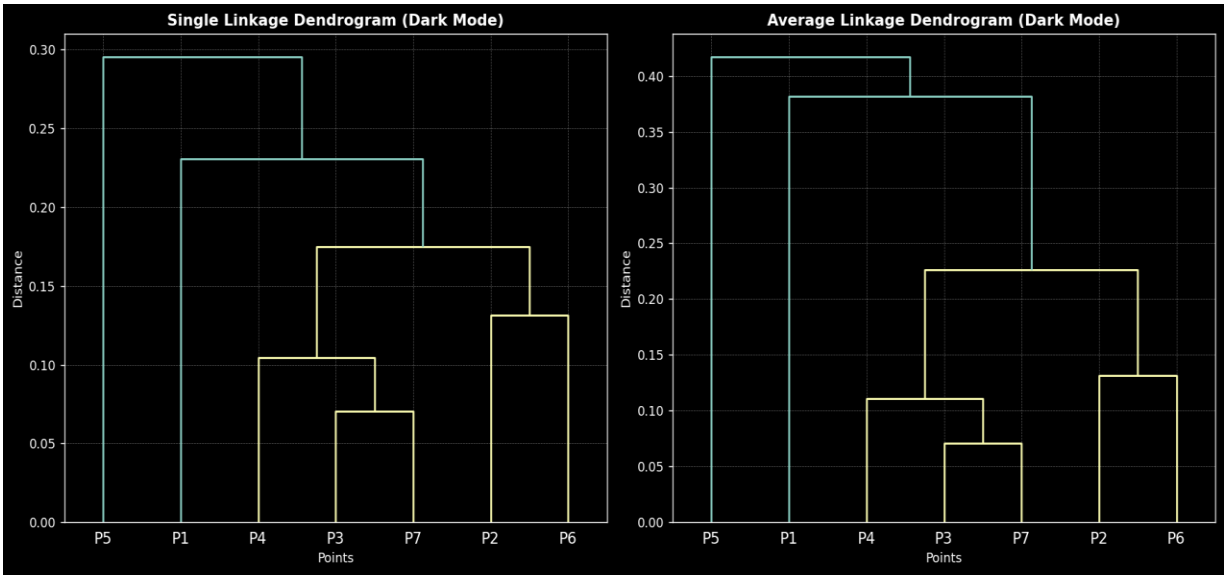
3. Merge **C1 & P4 → C2 = {P3,P4,P7}** (0.105)
4. Distance to P2:

$$d(C2,P2) = \frac{d(P2,P3) + d(P2,P4) + d(P2,P7)}{3} = \frac{0.242 + 0.188 + 0.203}{3} \approx 0.211$$

Distance to P6:

$$d(C2,P6) = \frac{0.257 + 0.159 + 0.252}{3} \approx 0.223$$

5. Merge **C2 & P2 → C3 = {P2,P3,P4,P7}** (0.211)
6. Next merge **C3 & P6 → C4 = {P2,P3,P4,P6,P7}** (0.223)
7. Next merge **C4 & P1 → C5** (average distance ~0.303)
8. Finally merge **C5 & P5 → C6** (average distance ~0.411)



Q2. We have the following 2D data points:

Points: (2,1), (3,1), (3, 3), (4, 1), (5, 1), (6,7), (1,3), (2,5)
for K =3:
Centroid1: (2,1)
Centroid 2: (4, 1)
Centroid 3: (5, 1)

Euclidean Distance:

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

Ans –
We are given eight 2-dimensional points:

$$(2,1), (3,1), (3,3), (4,1), (5,1), (6,7), (1,3), (2,5)$$

The initial centroids for $K = 3$are:

- $C_1 = (2,1)$
- $C_2 = (4,1)$
- $C_3 = (5,1)$

To assign each point to a cluster, we compute how far each point is from the three centroids using the Euclidean distance formula:

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

Step 1: Distance Calculation and Initial Assignment

Using the formula above, the distances from every point to the three centroids are summarised below:

Point	Dist → C1	Dist → C2	Dist → C3	Assigned Cluster
(2,1)	0.00	2.00	3.00	C1
(3,1)	1.00	1.00	2.00	C1 (tie → choose smaller index)
(3,3)	2.24	2.24	2.83	C1
(4,1)	2.00	0.00	1.00	C2
(5,1)	3.00	1.00	0.00	C3
(6,7)	7.21	6.32	6.08	C3
(1,3)	2.24	3.61	4.47	C1
(2,5)	4.00	4.47	5.00	C1

Initial Cluster Groups
After comparing all the distances:

- **Cluster 1 (C1 = 2,1):** (2,1), (3,1), (3,3), (1,3), (2,5)
- **Cluster 2 (C2 = 4,1):** (4,1)
- **Cluster 3 (C3 = 5,1):** (5,1), (6,7)

Step 2: Update the Centroids

Now we compute the mean of the points inside each cluster to get the updated centroids.

Updated Centroid for Cluster 1

Points: (2,1), (3,1), (3,3), (1,3), (2,5)

$$x_{avg} = \frac{2+3+3+1+2}{5} = 2.2$$
$$y_{avg} = \frac{1+1+3+3+5}{5} = 2.6$$

So the new centroid is:

$$C1' = (2.2,2.6)$$

Updated Centroid for Cluster 2

Only one point: (4,1)

$$C2' = (4,1)$$

Updated Centroid for Cluster 3

Points: (5,1) and (6,7)

$$x_{avg} = \frac{5+6}{2} = 5.5$$
$$y_{avg} = \frac{1+7}{2} = 4$$

So:

$$C3' = (5.5,4)$$

Final Centroids After One Iteration :

Cluster	New X	New Y
C1	2.2	2.6
C2	4.0	1.0
C3	5.5	4.0

Part B — Short-Answer

Q1.

a) Describe agglomerative hierarchical clustering.
b) Describe divisive hierarchical clustering.
c) Which one is more commonly used and why?

ANS –

a. Agglomerative clustering follows a bottom-up strategy. Each data point starts off as an individual cluster, and at every step the two closest clusters are joined together. This merging continues until all samples are combined into a single cluster or until a desired number of groups is reached.

b. Divisive clustering works in the opposite direction: it begins with all data points grouped together, and then repeatedly splits the cluster into smaller ones. The process keeps dividing the clusters until each point stands alone.

c. In practice, agglomerative methods are more widely applied because they are simpler to implement, require fewer assumptions, and are typically easier to compute than divisive approaches.

Q2.

a) To improve clustering quality, should inter-cluster distance be maximized or minimized?
b) Same question for intra-cluster distance — explain the reasoning.

a. For good clustering, the distance *between* clusters should be as large as possible. More separation means the clusters are well-defined and do not overlap.

b. The distance *inside* a cluster should be small. A low internal distance indicates that the points within the same cluster are closely related and similar to each other.

Q3.

a) Define single link, complete link, and average link.
b) Explain one strength and one weakness of single-link clustering.

a. Single Link: Measures the distance between two clusters by looking at the shortest distance between any pair of points belonging to the two clusters.

Complete Link: Uses the opposite strategy and considers the farthest distance between any two points across the two clusters.

Average Link: Calculates the mean of all pairwise distances between points from the two clusters.

b. Advantage: It can uncover clusters of irregular shapes, which makes it useful for datasets where clusters are not nicely spherical.

Drawback: It is vulnerable to the "chaining effect," where points or small groups act as bridges and incorrectly connect two clusters through long thin chains**.**

Q4.

a) What is the role of tokenization and give one example.
b) Compare stemming vs. lemmatization in terms of speed and accuracy.

a. Tokenization divides text into smaller meaningful units—such as words, terms, or subwords—that can be processed by an NLP model.
Example:
“Machine learning is interesting” → [“Machine”, “learning”, “is”, “interesting”]

b.

Feature	Stemming	Lemmatization
Definition	Removes word endings using heuristic rules.	Converts a word to its base/lemma using vocabulary and grammar rules.
Speed	Faster because it uses simple chopping rules.	Slower due to dictionary lookup and morphological analysis.
Accuracy	Less accurate; may produce non-words.	More accurate and outputs valid dictionary forms.
Example	"studies" → "studi"	"studies" → "study"

Q5.

a) Explain what word sense ambiguity is and provide an example.
b) Explain why pronoun reference ambiguity can confuse a model.

a. A word is ambiguous when it can represent multiple meanings depending on the context.
Example: "bat" can mean a flying mammal or sports equipment.

b. When a pronoun can refer to more than one possible noun, the model may not know which one is correct without deeper context.
Example: "Sarah called Emma because *she* was upset."
Here, *she* could refer to either person, causing confusion.

Q6.

a) Why can't NLP tasks like POS tagging be solved by predicting each token independently?
b) Give one example where decisions are mutually dependent in a sentence.

a. A word's part of speech depends heavily on surrounding words. Ignoring context would lead to incorrect predictions, since many words can function differently depending on the sentence.

b. In the sentence "I will park the car," the word *will* indicates that *park* is being used as a verb. Without considering the other words, the model could misinterpret its role.