SAMPOERNA UNIVERSITY

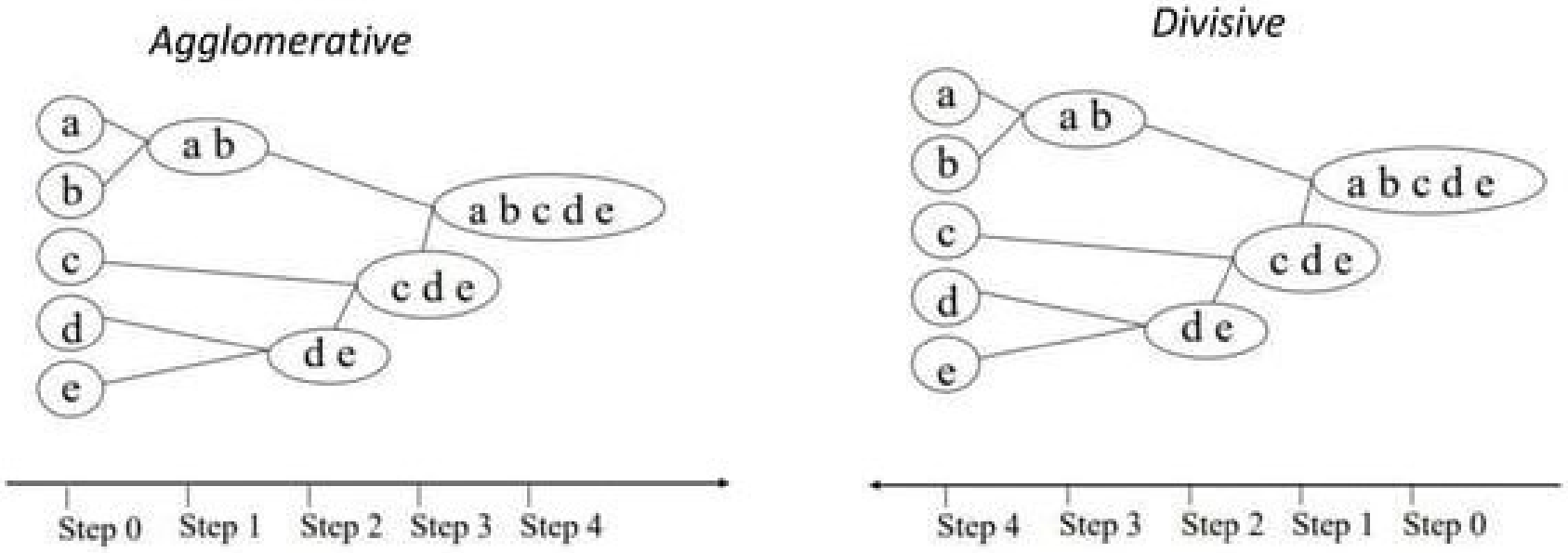
GENERAL COURSE

COMP3315: Artificial Intelligence

Unsupervised Learning – Divisive Hierarchical Clustering

By: Gilang Raka Rayuda Dewa

Reminder: Agglomerative vs Divisive



Initial Dataset

A = (1, 2)
B = (2, 1)
C = (4, 5)
D = (5, 4)
E = (6, 5)

We will use Divisive clustering.

Good news! We can utilize K-Means to do that.

Step 0: Initial Cluster

Put them all dataset into 1 cluster

$$Cluster\_0 = [A, B, C, D, E]$$

Step 1: Calculate ED via the K-Means Rule

SAMPOERNA UNIVERSITY

We randomly initialize 2 centroids. Let's choose:

- Centroid_1 = A = (1, 2)
- Centroid_2 = C = (4, 5)

Calculate ED (Euclidean distance) of each points

Point	Dist to (1,2)	Dist to (4,5)	Assign to
A(1, 2)	0	$\sqrt{[(3^2 + 3^2)]} = 4.24$	C1
B(2, 1)	$\sqrt{[(1^2 + 1^2)]} = 1.41$	$\sqrt{[(2^2 + 4^2)]} = 4.47$	C1
C(4,5)	$\sqrt{[(3^2 + 3^2)]} = 4.24$	0	C2
D(5,4)	$\sqrt{[(4^2 + 2^2)]} = 4.47$	$\sqrt{[(1^2 + 1^2)]} = 1.41$	C2
E(6,5)	$\sqrt{[(5^2 + 3^2)]} = 5.83$	$\sqrt{[(2^2 + 0^2)]} = 2.00$	C2

Step 2: Ensure Convergence

Recalculate the Centroid and Recalculate ED again to ensure convergence results.

- **Cluster_1 (A, B):**
  Mean x = (1 + 2)/2 = 1.5
  Mean y = (2 + 1)/2 = 1.5
  → New Centroid_1 = (1.5, 1.5)

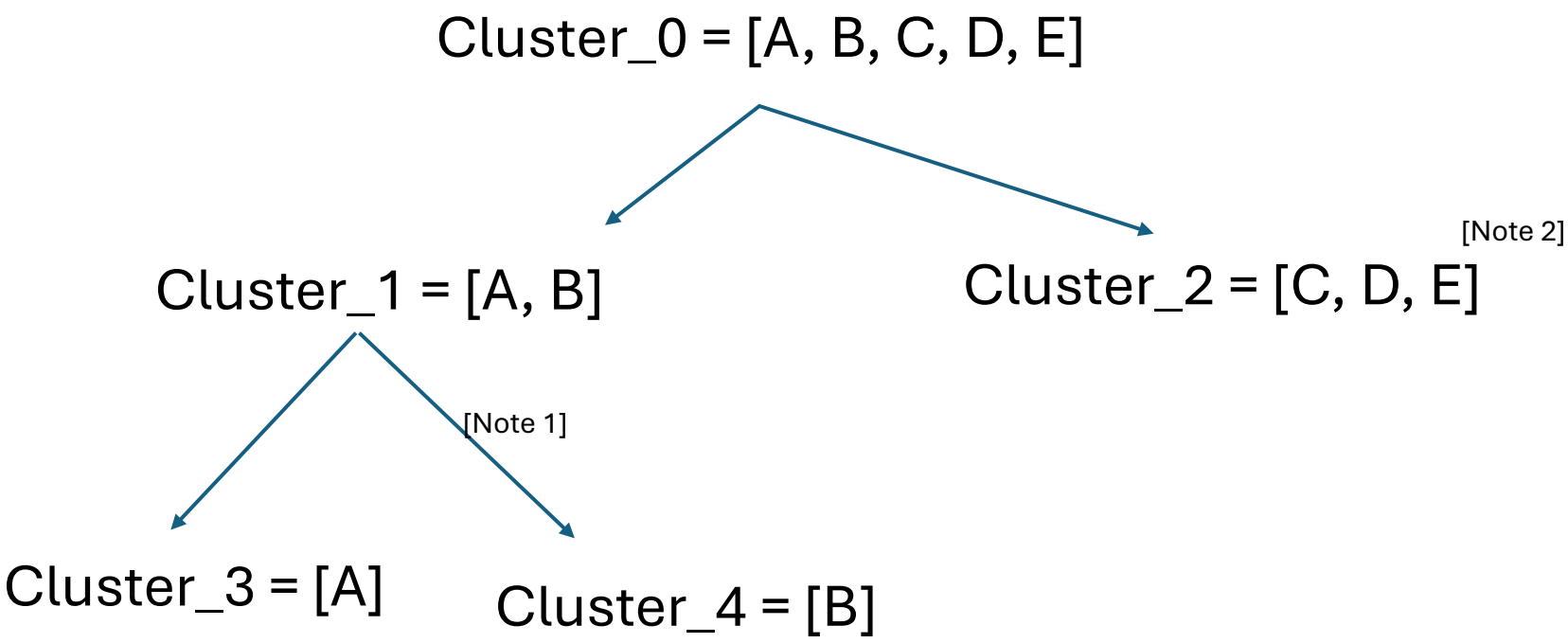
- **Cluster_2 (C, D, E):**
  Mean x = (4 + 5 + 6)/3 = 5
  Mean y = (5 + 4 + 5)/3 ≈ 4.67
  → New Centroid_2 = (5, 4.67)

Point	Dist to (1.5,1.5)	Dist to (5, 4.67)	Assign to
A	$\sqrt{[0.5^2 + 0.5^2]} = 0.71$	$\sqrt{[(4^2 + 2.67^2)]} \approx 4.81$	C1
B	$\sqrt{[0.5^2 + 0.5^2]} = 0.71$	$\sqrt{[(3^2 + 3.67^2)]} \approx 4.76$	C1
C	$\sqrt{[(2.5^2 + 3.5^2)]} \approx 4.30$	$\sqrt{[(1^2 + 0.33^2)]} \approx 1.05$	C2
D	$\sqrt{[(3.5^2 + 2.5^2)]} \approx 4.30$	$\sqrt{[(0^2 + 0.67^2)]} \approx 0.67$	C2
E	$\sqrt{[(4.5^2 + 3.5^2)]} \approx 5.70$	$\sqrt{[(1^2 + 0.33^2)]} \approx 1.05$	C2

No change → we stop here.

Step 3: Assign Cluster

Divide cluster based on K-Means assignments



[Note 1] Cluster 1 only consists of 2 members, so we can split them without calculation
[Note 2] Cluster 2 has three members, so we still need to split them.

Step 4: Calculate ED for Remaining Cluster

Recalculate K-means for the remaining cluster, i.e., Cluster 2

- Centroid_5 = C = (4, 5)
- Centroid_6 = E = (6, 5)

Point	Dist to (4,5)	Dist to (6,5)	Assign to
C(4,5)	0	2	C5
D(5,4)	$\sqrt{[(1^2 + 1^2)]} = 1.41$	$\sqrt{[(1^2 + 1^2)]} = 1.41$	Tie → pick randomly, e.g., C6
E(6,5)	2	0	C6

Step 5: Ensure Convergence

Recalculate the Centroid and Recalculate ED again to ensure convergence results.

- **Cluster_5 (C):**
  Mean = (4, 5)

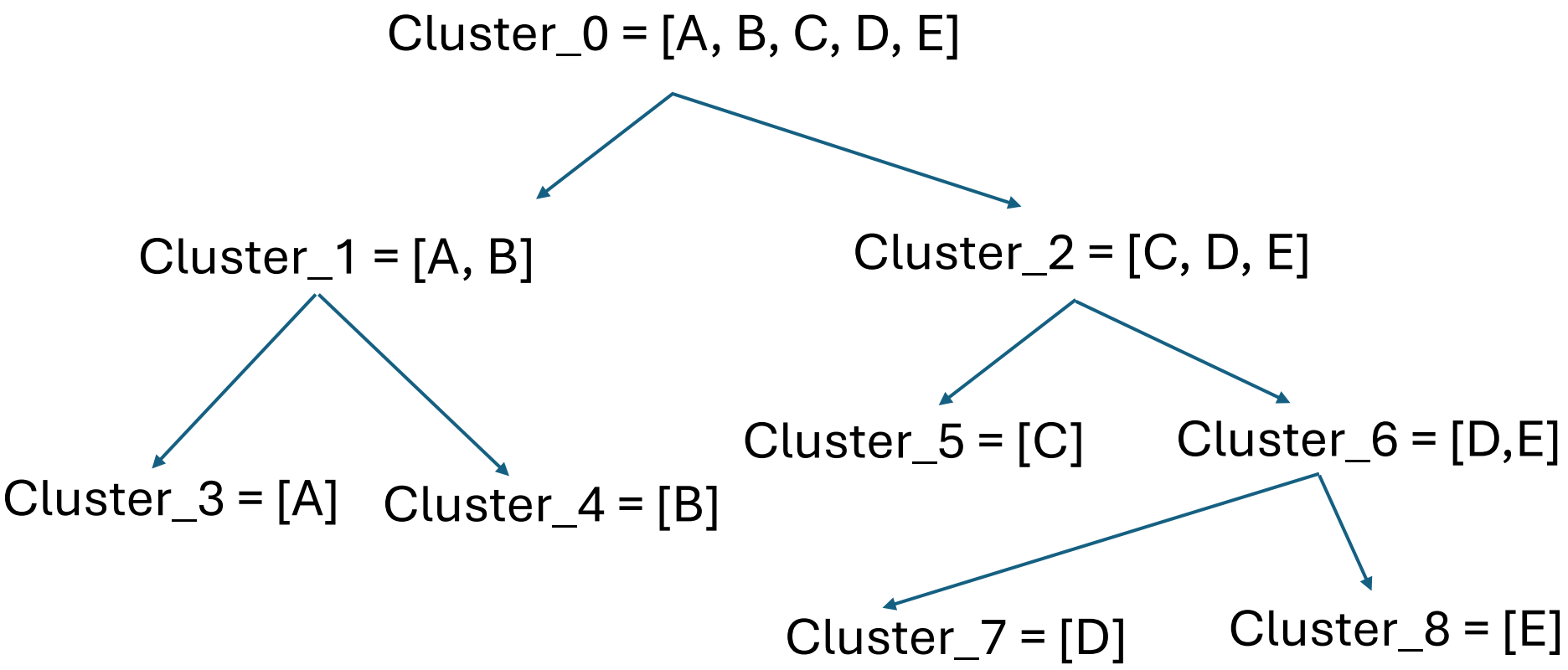
- **Cluster_6 (D, E):**
  Mean x = (5 + 6)/2 = 5.5
  Mean y = (4 + 5)/2 = 4.5
  → New Centroid_2 = (5.5, 4.5)

Point	Dist to (4,5)	Dist to (5.5,4.5)	Assign to
C	0	$\sqrt{[(1.5^2 + 0.5^2)]} \approx 1.58$	C5
D	$\sqrt{[(1^2 + 1^2)]} = 1.41$	$\sqrt{[(0.5^2 + 0.5^2)]} \approx 0.71$	C6
E	2	$\sqrt{[(0.5^2 + 0.5^2)]} \approx 0.71$	C6

No change → we stop here.

Step 6: Divide Cluster Again

Divide cluster based on K-Means assignments



[Note 3] Cluster 6 only consists of 2 members, so we can split them without calculation.

Example 1

You are analyzing emails to group them by similarity. Each email is rated based on two features, i.e., the number of links and the number of ALL-CAPS words. The data is:

Email	Links	ALL-CAPS Words
E1	1	2
E2	2	1
E3	7	8
E4	8	7

Use divisive clustering with K-Means (k=2) to split the emails into two clusters.

Example 2

Erwin and Widitra are clustering emails using K-Means (k=2) based on the number of links and ALL-CAPS words. But then, Kenjie and Rohit came and accidentally deleted a specific number of E4. Here's the remaining dataset:

Email	Links	ALL-CAPS Words	Assigned Cluster
E1	1	8	Cluster A
E2	2	7	Cluster A
E3	7	3	Cluster B
E4			Cluster B

If the centroids are fixed at (1,2) and (7,8). What is the possible value for E4 links and ALL-CAPS Words?