ML
LECTURE-18

BY
Dr. Ramesh Kumar Thakur
Assistant Professor (II)
School Of Computer Engineering

Divisive hierarchical clustering

- ❖ The **divisive method starts at the top and at each level recursively split one of the existing clusters at that level into two new clusters.**

- ❖ If there are N observations in the dataset, there the divisive method also will produce $N - 1$ levels in the hierarchy.

- ❖ The **split is chosen to produce two new groups with the largest "between-group dissimilarity".**

- ❖ For example, the divisive method is shown in Figure 13.11. Each nonterminal node has two daughter nodes.

- ❖ The two daughters represent the two groups resulting from the split of the parent.

Divisive hierarchical clustering

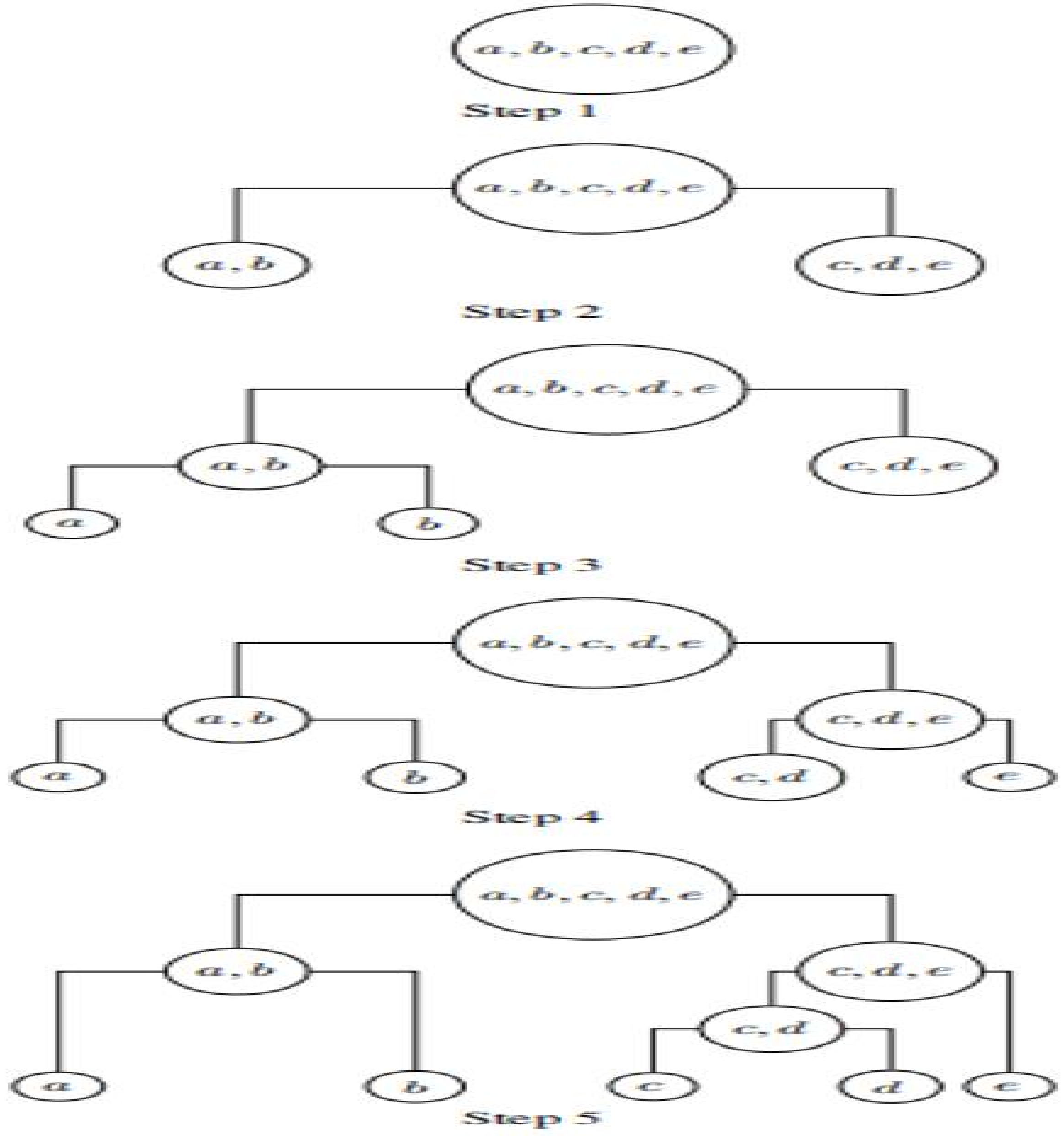


Figure 13.11: Hierarchical clustering using divisive method

Algorithm for divisive hierarchical clustering

- Divisive clustering algorithms begin with the entire data set as a single cluster, and recursively divide one of the existing clusters into two daughter clusters at each iteration in a top-down fashion.

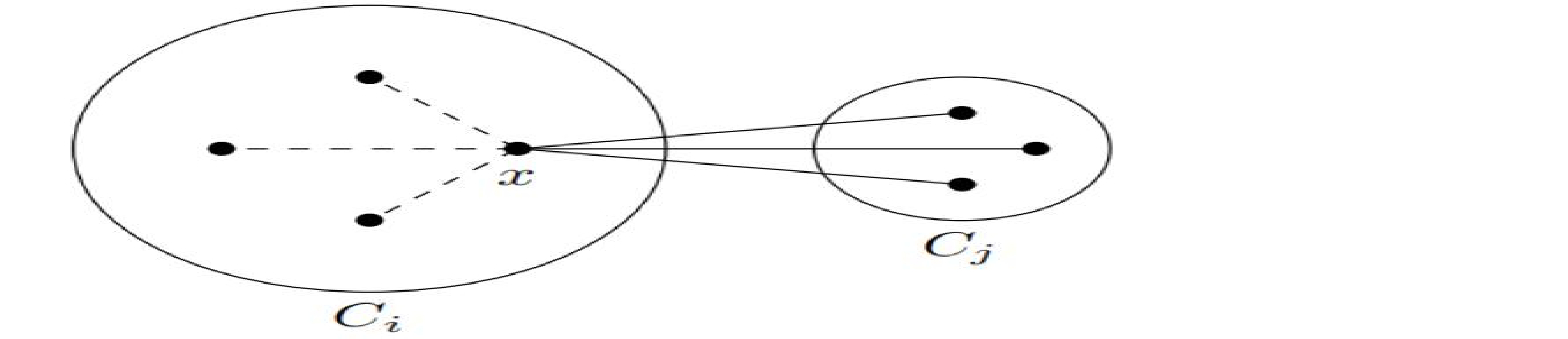
- To apply this procedure, we need a separate algorithm to divide a given dataset into two clusters.

- The divisive algorithm may be implemented by using the k-means algorithm with $k = 2$ to perform the splits at each iteration.

- However, it would not necessarily produce a splitting sequence that possesses the monotonicity property required for dendrogram representation.

DIANA (DIvisive ANAlysis)

- DIANA is a divisive hierarchical clustering technique. Here is an outline of the algorithm.
- Step 1. Suppose that cluster Cl is going to be split into clusters Ci and Cj .
- Step 2. Let Ci = Cl and Cj = ∅ .
- Step 3. For each object x ∈ Ci:
- (a) For the first iteration, compute the average distance of x to all other objects.
- (b) For the remaining iterations, compute
- Dx = average {d(x, y) : y ∈ Ci} − average{d(x, y) : y ∈ Cj}.



D_x = (average of dashed lines) − (average of solid lines)

DIANA (DIvisive ANAlysis)

- Step 4. (a) For the first iteration, move the object with the maximum average distance to Cj .
- (b) For the remaining iterations, find an object x in Ci for which Dx is the largest. If Dx > 0 then move x to Cj .
- Step 5. Repeat Steps 3(b) and 4(b) until all differences Dx are negative. Then Cl is split into Ci and Cj .
- Step 6. Select the smaller cluster with the largest diameter. (The diameter of a cluster is the largest dissimilarity between any two of its objects.) Then divide this cluster, following Steps 1-5.
- Step 7. Repeat Step 6 until all clusters contain only a single object.

Example of DIANA (DIvisive ANAlysis)

- **Problem :** Given the dataset {a, b, c, d, e} and the distance matrix in Table 13.4, construct a dendrogram by the divisive analysis algorithm.

	a	b	c	d	e
a	0	9	3	6	11
b	9	0	7	5	10
c	3	7	0	9	2
d	6	5	9	0	8
e	11	10	2	8	0

Table 13.4: Example for distance matrix

- **Solution:**
- 1. We have, initially
- Cl = {a, b, c, d, e}
- 2. We write
- Ci = Cl, Cj = ∅ .
- 3. Division into clusters

Example of DIANA (DIvisive ANAlysis)

- (a) Initial iteration
- Let us calculate the average dissimilarities of the objects in Ci with the other objects in Ci.
- Average dissimilarity of a $= \frac{1}{4}(d(a,b) + d(a,c) + d(a,e)) = \frac{1}{4}(9 + 3 + 6 + 11) = 7.25$

- Similarly we have :
- Average dissimilarity of b = 7.75
- Average dissimilarity of c = 5.25
- Average dissimilarity of d = 7.00
- Average dissimilarity of e = 7.75
- The highest average distance is 7.75 and there are two corresponding objects. We choose one of them, b, arbitrarily. We move b to Cj .

- We now have
- Ci = {a, c, d, e}, Cj = $\varnothing \cup \{b\} = \{b\}$

Example of DIANA (DIvisive ANAlysis)

- (b) Remaining iterations
- (i) 2-nd iteration.

$$D_a = \frac{1}{3}(d(a,c) + d(a,d) + d(a,e)) - \frac{1}{1}(d(a,b)) = \frac{20}{3} - 9 = -2.33$$
$$D_c = \frac{1}{3}(d(c,a) + d(c,d) + d(c,e)) - \frac{1}{1}(d(c,b)) = \frac{14}{3} - 7 = -2.33$$
$$D_d = \frac{1}{3}(d(d,a) + d(d,c) + d(d,e)) - \frac{1}{1}(d(c,b)) = \frac{23}{3} - 7 = 0.67$$
$$D_e = \frac{1}{3}(d(e,a) + d(e,c) + d(e,d)) - \frac{1}{1}(d(e,b)) = \frac{21}{3} - 7 = 0$$

- Dd is the largest and Dd > 0. So we move, d to Cj .
- We now have
- Ci = {a, c, e}, Cj = {b} ∪ {d} = {b, d}.
- (ii) 3-rd iteration

$$D_a = \frac{1}{2}(d(a,c) + d(a,e)) - \frac{1}{2}(d(a,b) + d(a,d)) = \frac{14}{2} - \frac{15}{2} = -0.5$$
$$D_c = \frac{1}{2}(d(c,a) + d(c,e)) - \frac{1}{2}(d(c,b) + d(c,d)) = \frac{5}{2} - \frac{16}{2} = -13.5$$
$$D_e = \frac{1}{2}(d(e,a) + d(e,c)) - \frac{1}{2}(d(e,b) + d(e,d)) = \frac{13}{2} - \frac{18}{2} = -2.5$$

- All are negative. So we stop and form the clusters Ci and Cj .

Example of DIANA (DIvisive ANAlysis)

- 4. To divide, Ci and Cj , we compute their diameters.

$$\begin{aligned}
\text{diameter}(C_i) &= \max\{d(a,c), d(a,e), d(c,e)\} \\
&= \max\{3, 11, 2\} \\
&= 11 \\
\text{diameter}(C_j) &= \max\{d(b,d)\} \\
&= 5
\end{aligned}$$

- The cluster with the largest diameter is Ci.
- So we now split Ci.
- We repeat the process by taking Cl = {a, c, e}.
- The remaining computations are left as an exercise to the students.

Example of AGNES (AGglomerative NESting)

- **Problem 1 :**Given the dataset {a, b, c, d, e} and the following distance matrix, construct a dendrogram by complete linkage hierarchical clustering using the agglomerative method.

	a	b	c	d	e
a	0	9	3	6	11
b	9	0	7	5	10
c	3	7	0	9	2
d	6	5	9	0	8
e	11	10	2	8	0

Table 13.4: Example for distance matrix

- **Solution:**
- The complete-linkage clustering uses the "maximum formula", that is, the following formula to compute the distance between two clusters A and B:
- $d(A, B) = \max\{d(x, y) : x \in A, y \in B\}$
- 1. Initial clustering (singleton sets)
- Dataset : {a, b, c, d, e}.
- C1: {a}, {b}, {c}, {d}, {e}.

Example of AGNES (AGglomerative NESting)

- 2. The following table gives the distances between the various clusters in C1:

	$\{a\}$	$\{b\}$	$\{c\}$	$\{d\}$	$\{e\}$
$\{a\}$	0	9	3	6	11
$\{b\}$	9	0	7	5	10
$\{c\}$	3	7	0	9	**2**
$\{d\}$	6	5	9	0	8
$\{e\}$	11	10	**2**	8	0

- In the above table, the minimum distance is the distance between the clusters {c} and {e}.
- Also
- d({c}, {e}) = 2.
- We merge {c} and {e} to form the cluster {c, e}.
- The new set of clusters C2: {a}, {b}, {d}, {c, e}.

Example of AGNES (AGglomerative NESting)

- 3. Let us compute the distance of {c, e} from other clusters.
- d({c, e}, {a}) = max{d(c, a), d(e, a)} = max{3, 11} = 11.
- d({c, e}, {b}) = max{d(c, b), d(e, b)} = max{7, 10} = 10.
- d({c, e}, {d}) = max{d(c, d), d(e, d)} = max{9, 8} = 9.
- The following table gives the distances between the various clusters in C2.

	$\{a\}$	$\{b\}$	$\{d\}$	$\{c,e\}$
$\{a\}$	0	9	6	11
$\{b\}$	9	0	**5**	10
$\{d\}$	6	**5**	0	9
$\{c,e\}$	11	10	9	0

- In the above table, the minimum distance is the distance between the clusters {b} and {d}.
- Also
- d({b}, {d}) = 5.
- We merge {b} and {d} to form the cluster {b, d}.
- The new set of clusters C3: {a}, {b, d}, {c, e}.

Example of AGNES (AGglomerative NESting)

- 4. Let us compute the distance of {b, d} from other clusters.
- d({b, d}, {a}) = max{d(b, a), d(d, a)} = max{9, 6} = 9.
- d({b, d}, {c, e}) = max{d(b, c), d(b, e), d(d, c), d(d, e)} = max{7, 10, 9, 8} = 10.
- The following table gives the distances between the various clusters in C3.

	$\{a\}$	$\{b,d\}$	$\{c,e\}$
$\{a\}$	0	**9**	11
$\{b,d\}$	**9**	0	10
$\{c,e\}$	11	10	0

- In the above table, the minimum distance is the distance between the clusters {a} and {b, d}.
- Also
- d({a}, {b, d}) = 9.
- We merge {a} and {b, d} to form the cluster {a, b, d}.
- The new set of clusters C4: {a, b, d}, {c, e}

Example of AGNES (AGglomerative NESting)

- 5. Only two clusters are left. We merge them form a single cluster containing all data points. We have
- d({a, b, d}, {c, e}) = max{d(a, c), d(a, e), d(b, c), d(b, e), d(d, c), d(d, e)}

  = max{3, 11, 7, 10, 9, 8}

  = 11
- 6. Figure 13.14 shows the dendrogram of the hierarchical clustering.

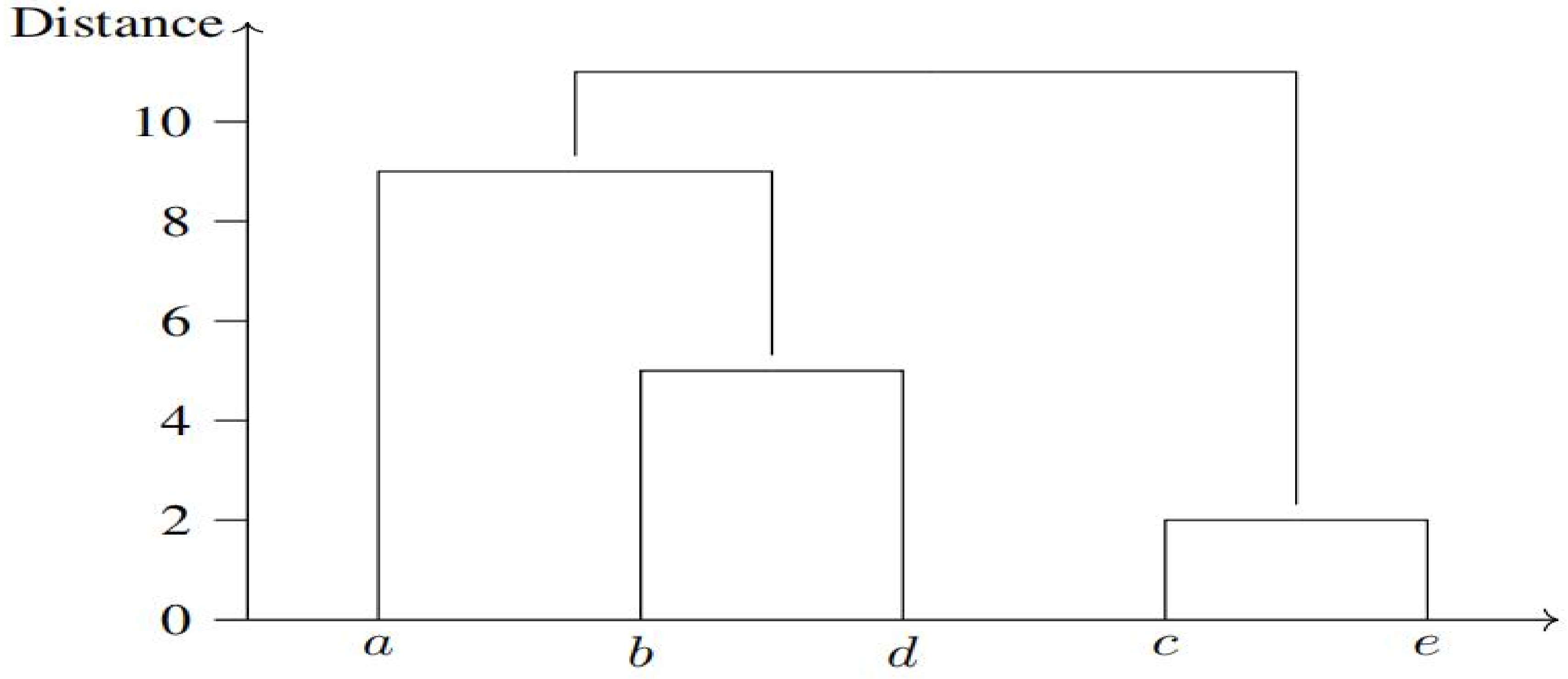


Figure 13.14: Dendrogram for the data given in Table 13.4 (complete linkage clustering)

Example of AGNES (AGglomerative NESting)

- **Problem 2 :**Given the dataset {a, b, c, d, e} and the distance matrix given in Table 13.4, construct a dendrogram by single-linkage hierarchical clustering using the agglomerative method.

	a	b	c	d	e
a	0	9	3	6	11
b	9	0	7	5	10
c	3	7	0	9	2
d	6	5	9	0	8
e	11	10	2	8	0

Table 13.4: Example for distance matrix

- **Solution:**
- The complete-linkage clustering uses the "maximum formula", that is, the following formula to compute the distance between two clusters A and B:
- $d(A, B) = \min\{d(x, y) : x \in A, y \in B\}$
- 1.Initial clustering (singleton sets)
- Dataset : {a, b, c, d, e}.
- C1: {a}, {b}, {c}, {d}, {e}.

Example of AGNES (AGglomerative NESting)

- 2. The following table gives the distances between the various clusters in C1:

	$\{a\}$	$\{b\}$	$\{c\}$	$\{d\}$	$\{e\}$
$\{a\}$	0	9	3	6	11
$\{b\}$	9	0	7	5	10
$\{c\}$	3	7	0	9	**2**
$\{d\}$	6	5	9	0	8
$\{e\}$	11	10	**2**	8	0

- In the above table, the minimum distance is the distance between the clusters {c} and {e}.
- Also
- d({c}, {e}) = 2.
- We merge {c} and {e} to form the cluster {c, e}.
- The new set of clusters C2: {a}, {b}, {d}, {c, e}.

Example of AGNES (AGglomerative NESting)

- 3. Let us compute the distance of {c, e} from other clusters.
- d({c, e}, {a}) = min{d(c, a), d(e, a)} = max{3, 11} = 3.
- d({c, e}, {b}) = min{d(c, b), d(e, b)} = max{7, 10} = 7.
- d({c, e}, {d}) = min{d(c, d), d(e, d)} = max{9, 8} = 8.
- The following table gives the distances between the various clusters in C2.

	$\{a\}$	$\{b\}$	$\{d\}$	$\{c, e\}$
$\{a\}$	0	9	6	**3**
$\{b\}$	9	0	5	7
$\{d\}$	6	5	0	8
$\{c, e\}$	**3**	7	8	0

- In the above table, the minimum distance is the distance between the clusters {a} and {c, e}.
- Also
- d({a}, {c, e}) = 3.
- We merge {a} and {c, e} to form the cluster {a, c, e}.
- The new set of clusters C3: {a, c, e}, {b}, {d}.

Example of AGNES (AGglomerative NESting)

- 4. Let us compute the distance of {a, c, e} from other clusters.
- d({a, c, e}, {b}) = min{d(a, b), d(c, b), d(e, b)} = {9, 7, 10} = 7
- d({a, c, e}, {d}) = min{d(a, d), d(c, d), d(e, d)} = {6, 9, 8} = 6
- The following table gives the distances between the various clusters in C3.

	$\{a, c, e\}$	$\{b\}$	$\{d\}$
$\{a, c, e\}$	0	7	6
$\{b\}$	7	0	**5**
$\{d\}$	6	**5**	0

- In the above table, the minimum distance is between {b} and {d}. Also
- d({b}, {d}) = 5.
- We merge {b} and {d} to form the cluster {b, d}.
- The new set of clusters C4: {a, c, e}, {b, d}

Example of AGNES (AGglomerative NESting)

- 5. Only two clusters are left. We merge them form a single cluster containing all data points. We have
- $d(\{a, c, e\}, \{b, d\}) = \min\{d(a, b), d(a, d), d(c, b), d(c, d), d(e, b), d(e, d)\}$

$$= \min\{9, 6, 7, 9, 10, 8\}$$

$$= 6$$

- 6. Figure 13.15 shows the dendrogram of the hierarchical clustering.

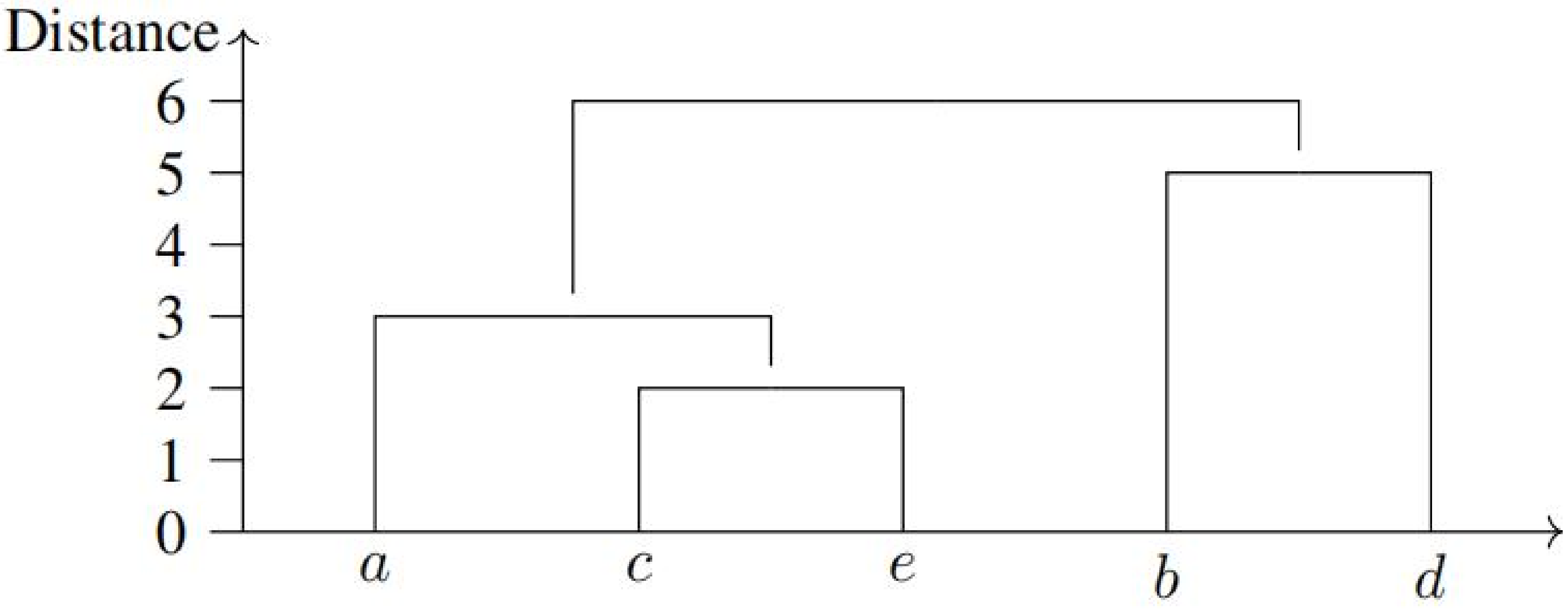
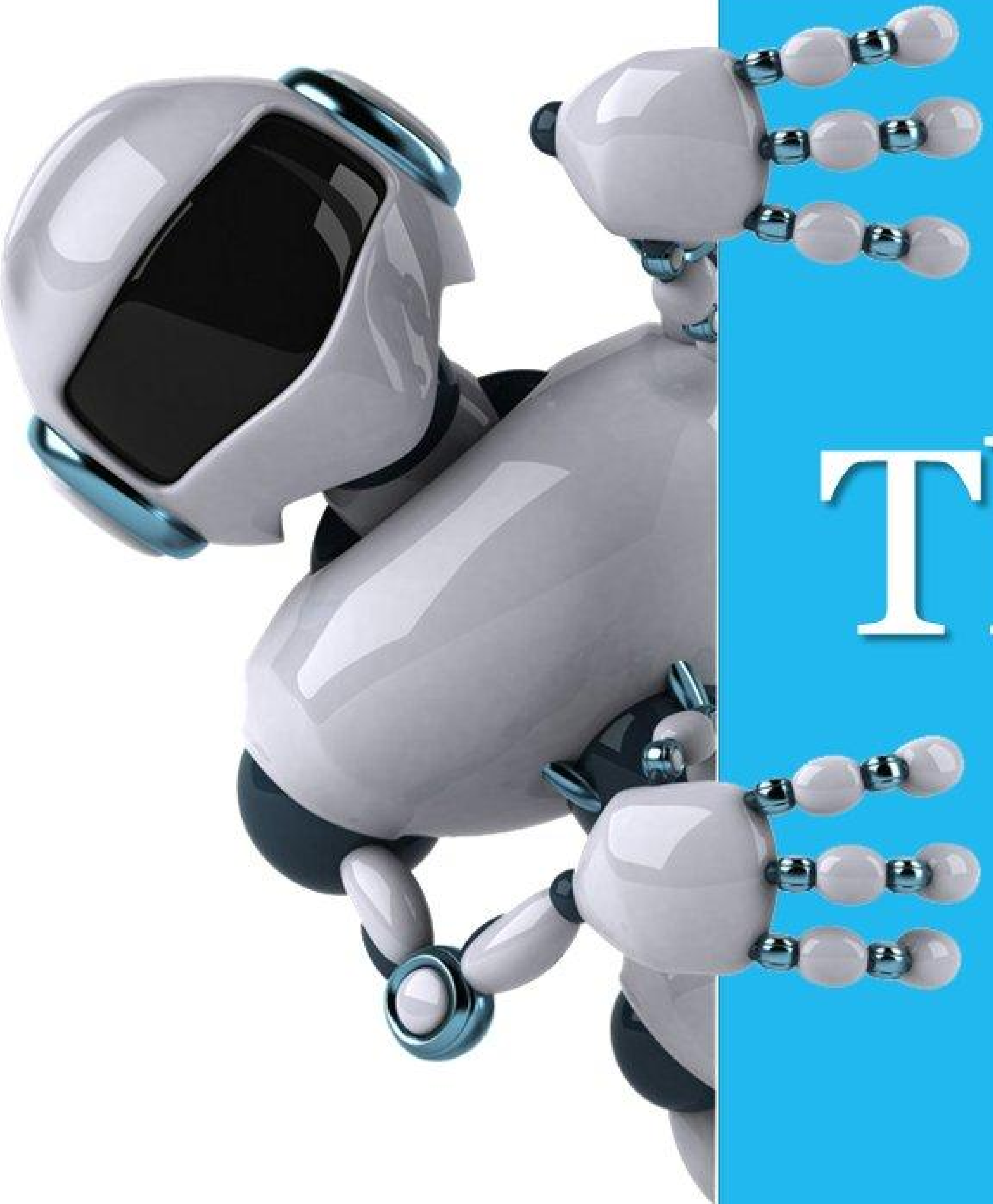


Figure 13.15: Dendrogram for the data given in Table 13.4 (single linkage clustering)



Thank you