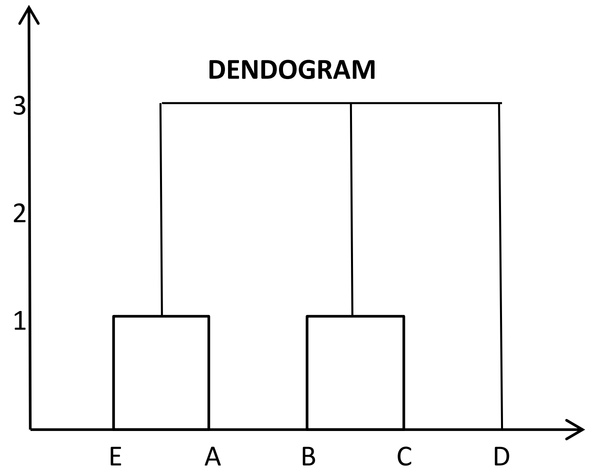
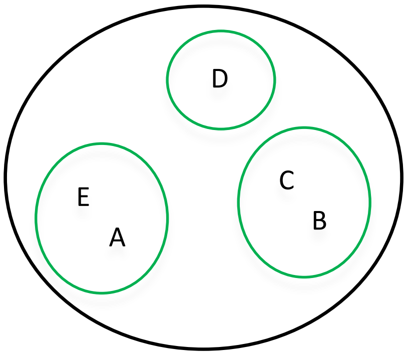
Agglomerative Clustering





Bottom to Up

Question: *Perform Agglomerative Algorithm on the following data and plot a dendogram using single link approach. The given data indicates the distance between elements.*

Item	**E**	**A**	**C**	**B**	**D**
E	0	1	2	2	3
A	1	0	2	5	3
C	2	2	0	1	6
B	2	5	1	0	3
D	3	3	6	3	0

Proximity Matrix with Original Data

Item	**E**	**A**	**C**	**B**	**D**
E	0				
A	1	0			
C	2	2	0		
B	2	5	1	0	
D	3	3	6	3	0

Proximity Matrix
After ignoring the data above diagonal

Item	E	A	C	B	D
E	0				
A	1	0			
C	2	2	0		
B	2	5	1	0	
D	3	3	6	3	0

Pair with minimum value
Here, two cells have minimum value 1
Take any one
I have taken the pair (E,A)

Item	(E, A)	C	B	D
(E, A)	0			
C	2	0		
B	2	1	0	
D	3	6	3	0

Pair (E, A)
(1)
((E, A)->C) = min[(E, C),(A, C)]=min[2, 2]=2

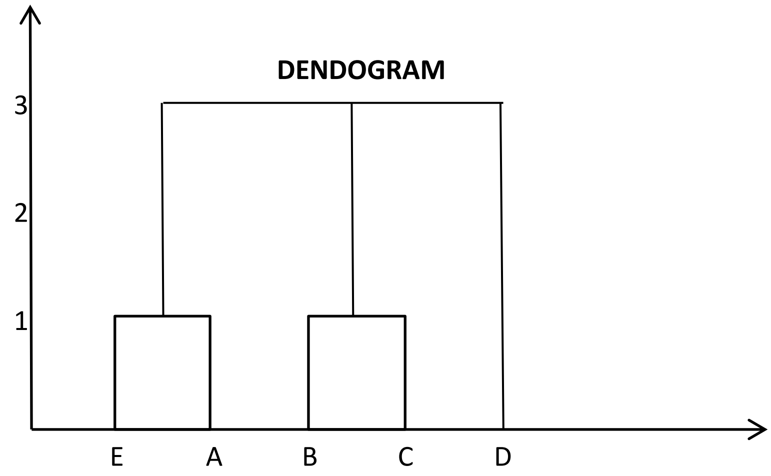
Item	(E, A)	(B, C)	D
(E, A)	0		
(B, C)	2	0	
D	3	3	0

Pair (B, C)
(1)
((B, C)->(E, A)) = min[(B, E),(B, A),(C, E)(C, A)]=min[2, 5, 2, 2]=2
((B, C)->D) = min[(B, D),(C, D)]=min[3, 6]=3

Item	((E, A) (B, C))	D
((E, A) (B, C))	0	
D	3	0

Pair (((E, A) (B, C)),D)
(2)
((E, A) (B, C))->D = min[(E, D), (A, D),(B, D),(C, D)]
=min[3, 3, 3, 6]=3

Dendogram: A tree like diagram that records the sequences of merges or splits. Merge is used in Agglomerative clustering and Split is used in Divisive Clustering. Agglomerative clustering is Bottom-Up while Divisive clustering is Top-Down.



Bottom to Up

In this example, I have taken minimum distance between two elements. Also, there are other options available. They are:

1. MIN
2. MAX
3. Group AVERAGE
4. Distance between Centroids etc.

Agglomerative Clustering Algorithm

- More popular hierarchical clustering technique
- Basic algorithm is straightforward
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. Repeat
  4. Merge the two clusters
  5. Update the proximity matrix
  6. Until only a single cluster remains
- Key operation is the computation of the proximity of two clusters
  - ✓ Different approaches to defining the distance between clusters distinguish the different algorithms

Summery

- **Agglomerative clustering**
  - → Choose a cluster distance / dissimilarity scoring method
  - → Successively merge closest pair clusters
  - → "Dendrogram" shows sequence of merge & distances
  - → Complexity: O(m^2 log m)
- **"Clustering for understanding data matrix**
  - → Build clusters on rows (data) and columns (features)
  - → Reorder data & features to expose behavior across groups
- **Agglomerative clusters depend on dissimilarity**
  - → Choice determines characteristics of "found" clusters