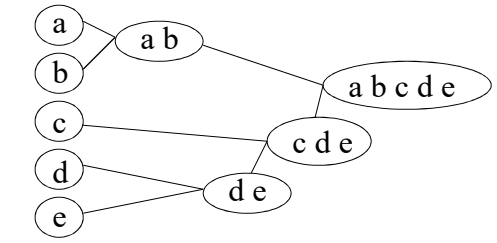
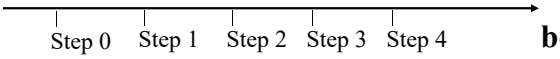
Hierarchical Clustering

- **Agglomerative approach**





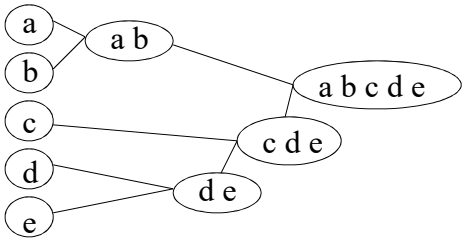
bottom-up

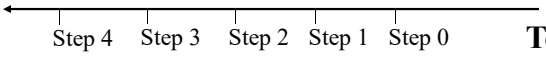
Initialization:
Each object is a cluster
Iteration:
Merge two clusters which are most similar to each other;
Until all objects are merged into a single cluster

1

Hierarchical Clustering

- **Divisive Approaches**





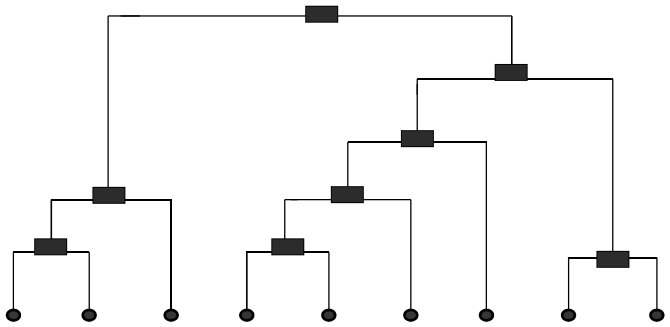
Top-down

Initialization:
All objects stay in one cluster
Iteration:
Select a cluster and split it into two sub clusters
Until each leaf cluster contains only one object

2

Dendrogram

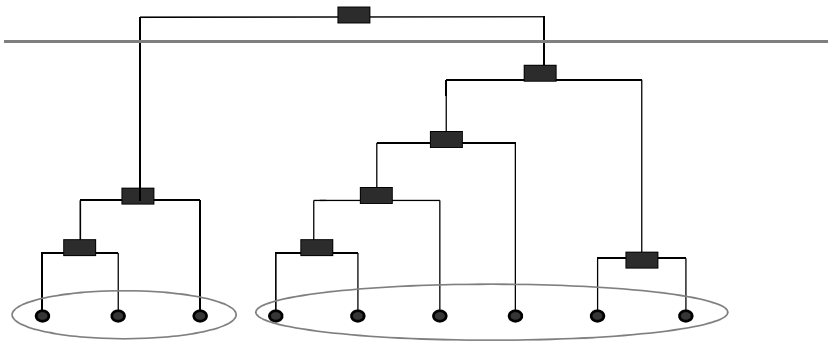
- A binary tree that shows how clusters are merged/split hierarchically
- Each node on the tree is a cluster; each leaf node is a singleton cluster



3

Dendrogram

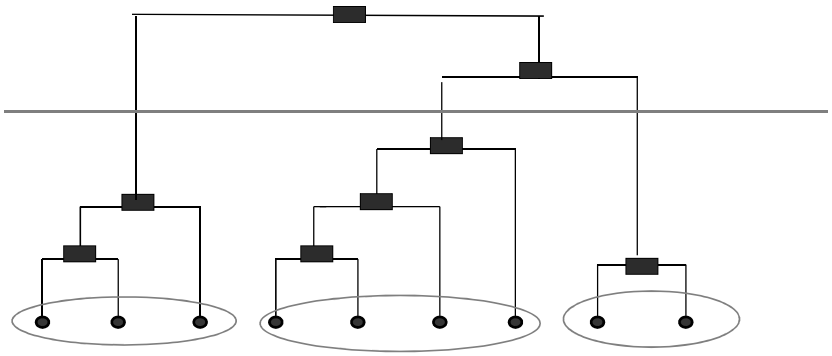
- A clustering of the data objects is obtained by cutting the *dendrogram* at the desired level, then each connected component forms a cluster



4

Dendrogram

- **A clustering of the data objects is obtained by cutting the *dendrogram* at the desired level, then each connected component forms a cluster**

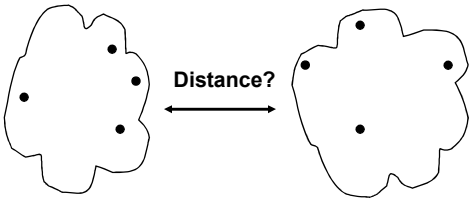


5

How to Merge Clusters?

- **How to measure the distance between clusters?**

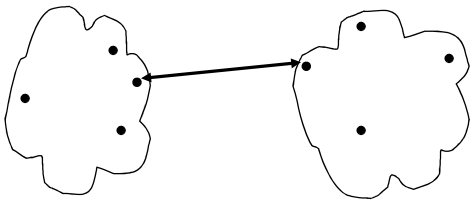
- **Single-link**
- **Complete-link**
- **Average-link**
- **Centroid distance**



Hint: *Distance between clusters* is usually defined on the basis of ***distance between objects.***

6

How to Define Inter-Cluster Distance



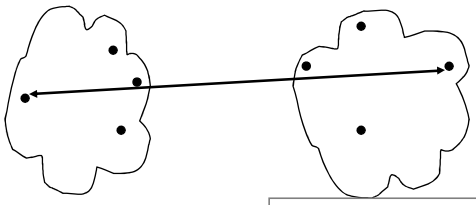
- **Single-link**
- **Complete-link**
- **Average-link**
- **Centroid distance**

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, q \in C_j} d(p, q)$$

The distance between two clusters is represented by the distance of the *closest pair of data objects* belonging to different clusters.

7

How to Define Inter-Cluster Distance



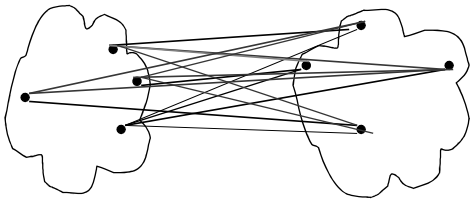
- **Single-link**
- **Complete-link**
- **Average-link**
- **Centroid distance**

$$d_{\min}(C_i, C_j) = \max_{p \in C_i, q \in C_j} d(p, q)$$

The distance between two clusters is represented by the distance of the *farthest pair of data objects* belonging to different clusters.

8

How to Define Inter-Cluster Distance



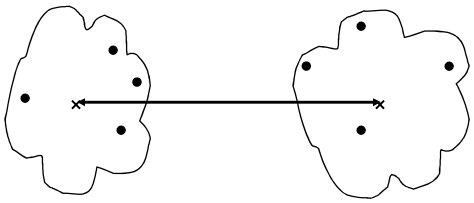
- **Single-link**
- **Complete-link**
- **Average-link**
- **Centroid distance**

$$d_{\min}(C_i, C_j) = \underset{p \in C_i, q \in C_j}{avg} d(p, q)$$

The distance between two clusters is represented by the *average* distance of *all pairs of data objects* belonging to different clusters.

9

How to Define Inter-Cluster Distance



m_i,m_j are the means of C_i, C_j,

- **Single-link**
- **Complete-link**
- **Average-link**
- **Centroid distance**

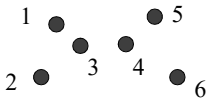
$$d_{mean}(C_i, C_j) = d(m_i, m_j)$$

The distance between two clusters is represented by the distance between *the means of the cluters.*

10

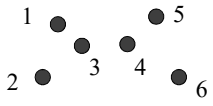
An Example of the Agglomerative Hierarchical Clustering Algorithm

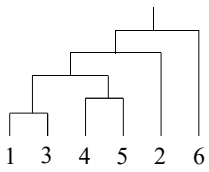
- For the following data set, we will get different clustering results with the single-link and complete-link algorithms.



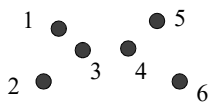
11

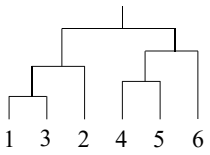
Result of the Single-Link algorithm





Result of the Complete-Link algorithm

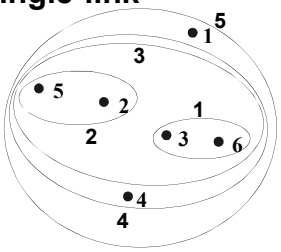




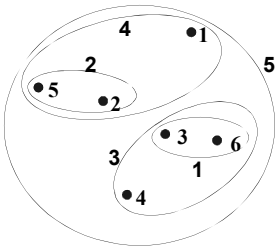
12

Hierarchical Clustering: Comparison

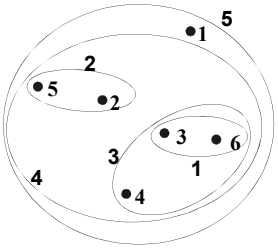
Single-link



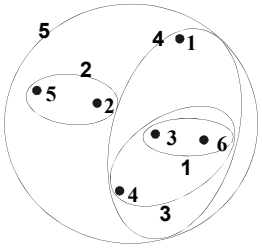
Complete-link



Average-link



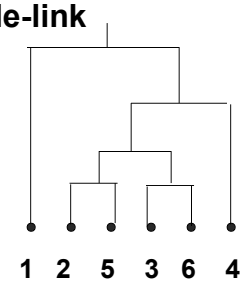
Centroid distance



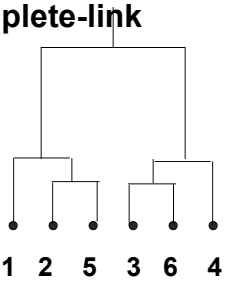
13

Compare Dendrograms

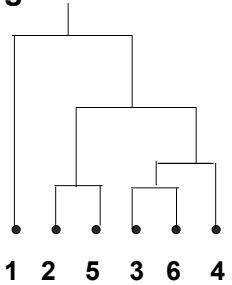
Single-link



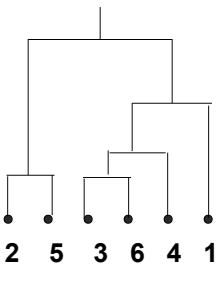
Complete-link



Average-link



Centroid distance



14