Hierarchical clustering

- Records are sequentially grouped to create clusters, based on distances between records and distances between clusters.
- Hierarchical clustering also produces a useful graphical display of the clustering process and results, called a dendrogram.

Strengths of Hierarchical Clustering

- No assumptions on the number of clusters
- Any desired number of clusters can be obtained by 'cutting'

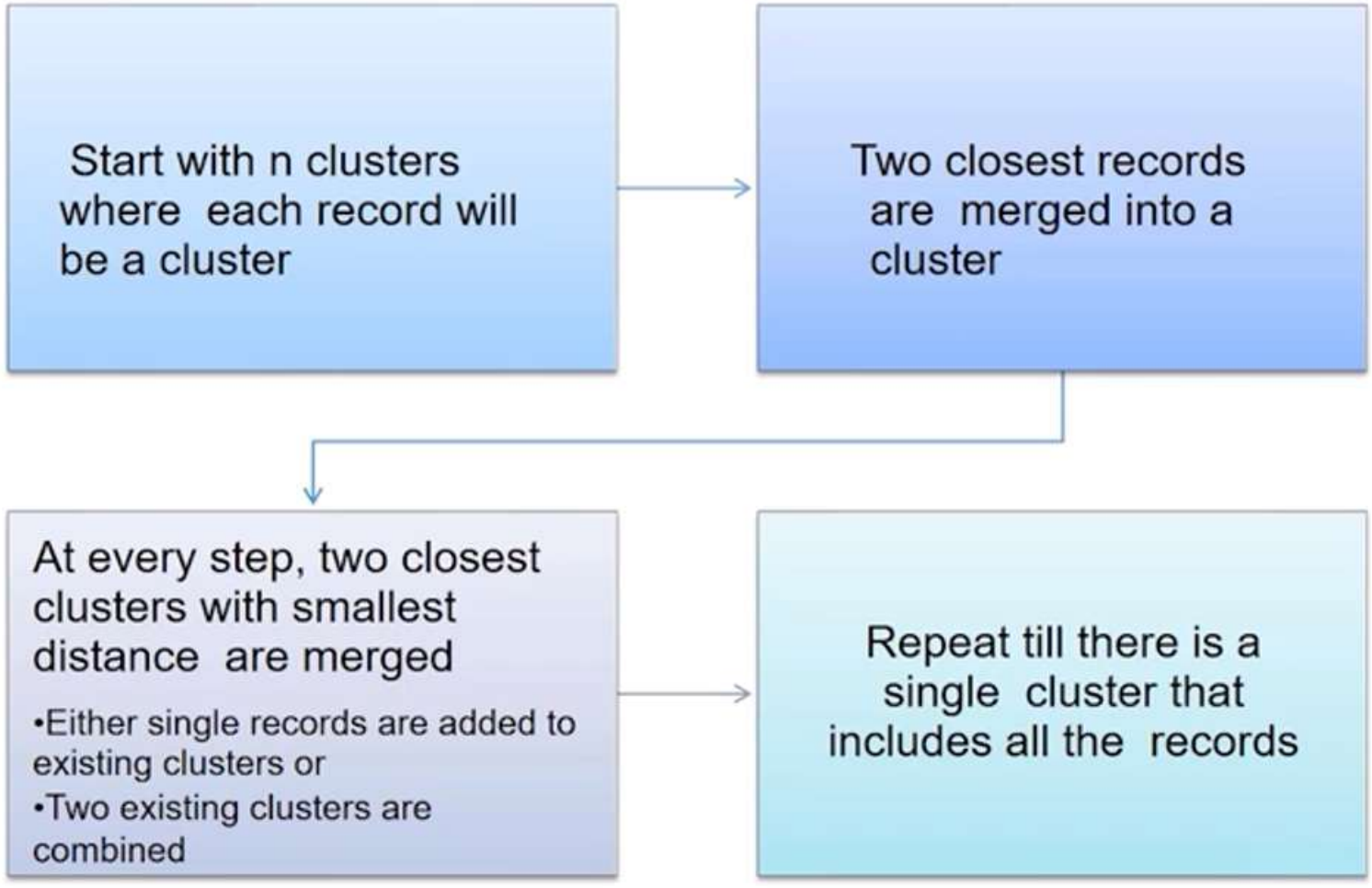
the dendrogram at the proper level

- Hierarchical clustering may correspond to meaningful taxonomies

Disadvantages of Hierarchical clustering

- Time complexity: not suitable for larger data sets.
- Very sensitive to outliers

Hierarchical clustering -Steps



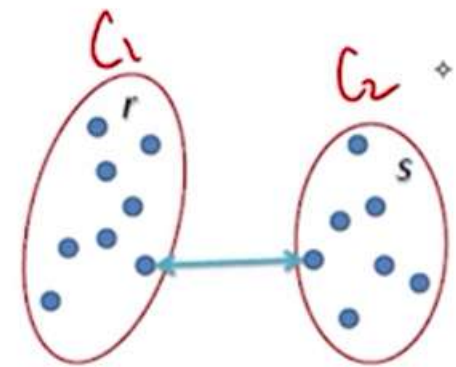
Hierarchical clustering – Distance between clusters

Linkage types

- Single linkage
- Complete linkage
- Average linkage
- Centroid linkage
- Ward's Method

Single Linkage

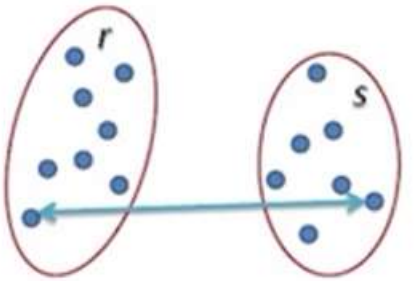
- Distance between two clusters is defined as the shortest distance between two points in each cluster.



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Complete linkage

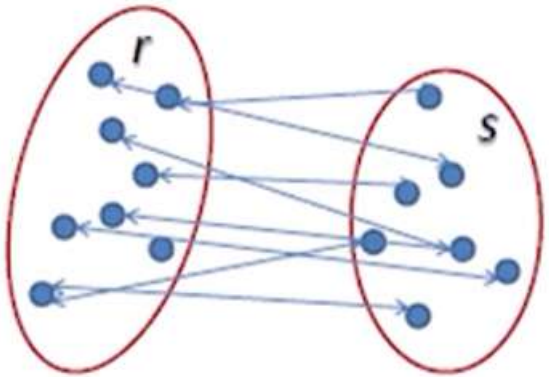
- Distance between two clusters is defined as the longest distance between two points in each cluster.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Average linkage

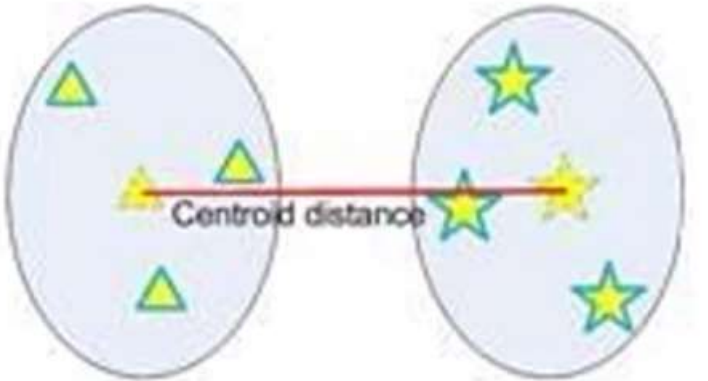
- Distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

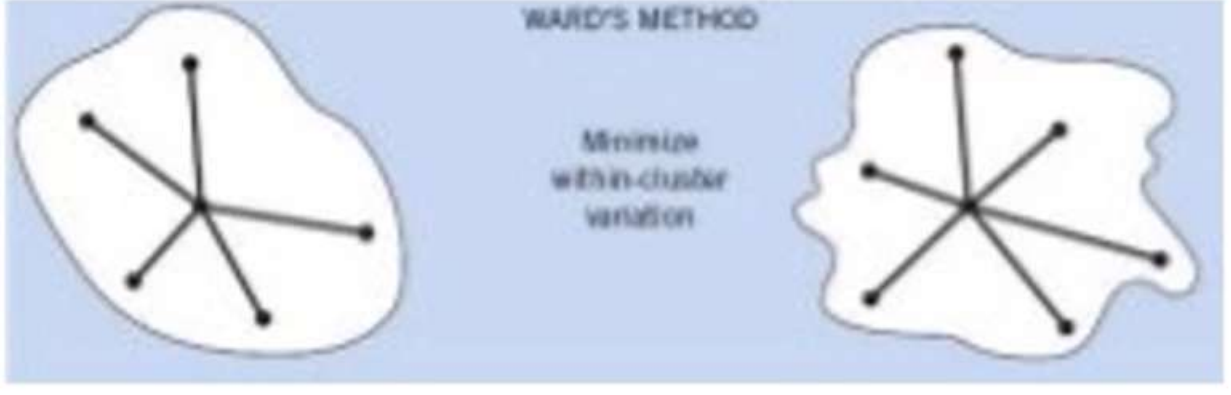
Centroid linkage

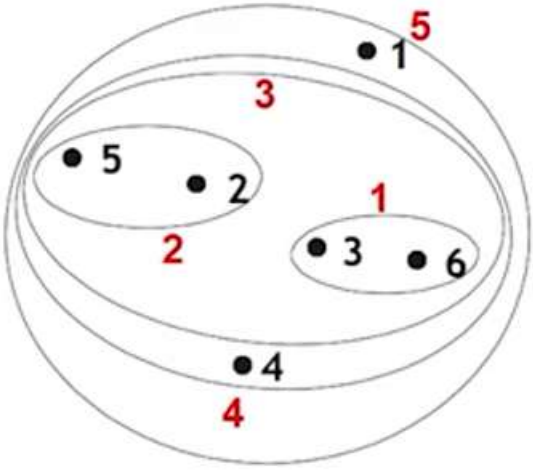
- Based on centroid distance. clusters are represented by their mean values for each variable, which forms a vector of means.
- Distance between 2 clusters is distance between the 2 vectors



Ward's linkage

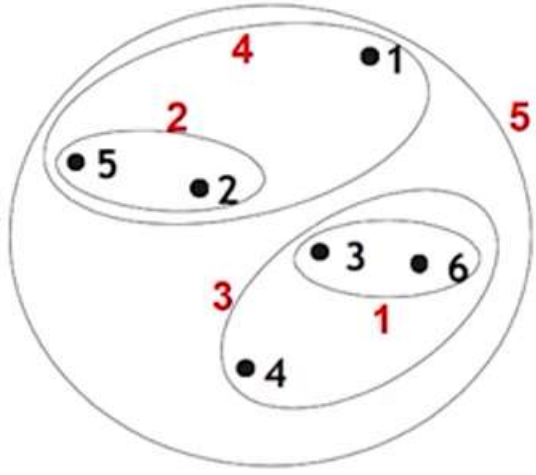
- Similar to group average and centroid distance
- joins records and clusters together progressively to produce larger and larger clusters, but operates slightly differently from the general approach.

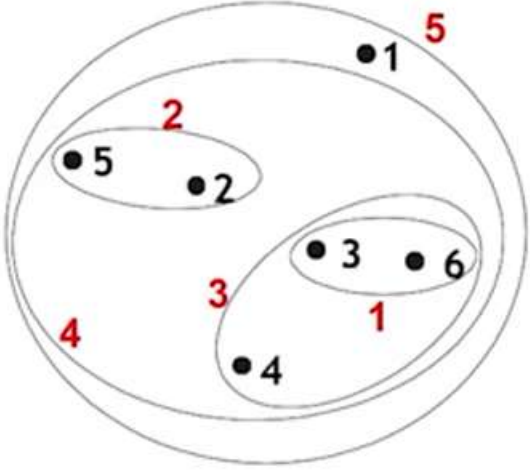




MIN

MAX





Group Average

Dendrograms

- A *dendrogram* is a treelike diagram that summarizes the process of clustering
- On the *x*-axis are the records
- Similar records are joined by lines whose vertical length reflects the distance between the records
- the greater the difference in height, the more dissimilarity
- By choosing a cutoff distance on the *y*-axis, a set of clusters is created

Dendrograms