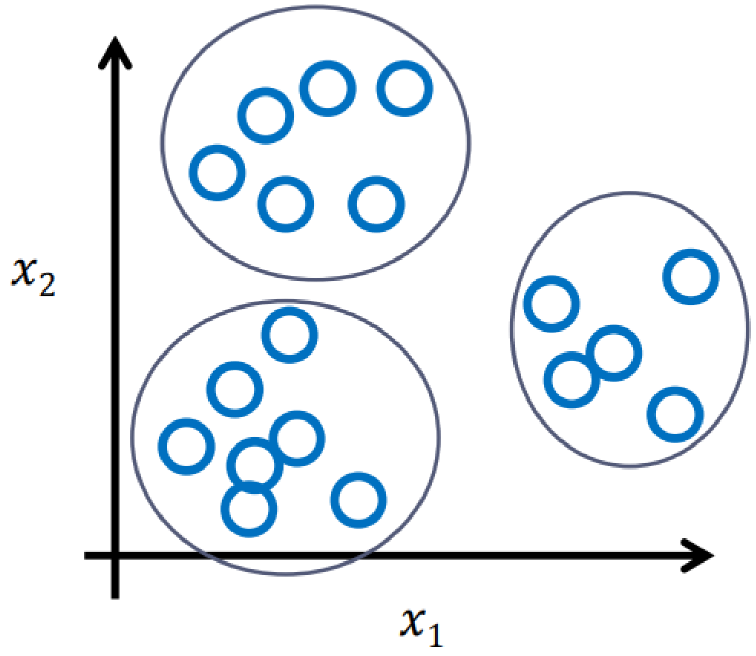
Clustering

Outline

- Clustering Definition
- Clustering main approaches
  - Partitional (flat)
  - Hierarchical

Definition

- We have a set of unlabeled data points and we intend to find groups of similar objects (based on the observed features)
  - high intra-cluster similarity
  - low inter-cluster similarity



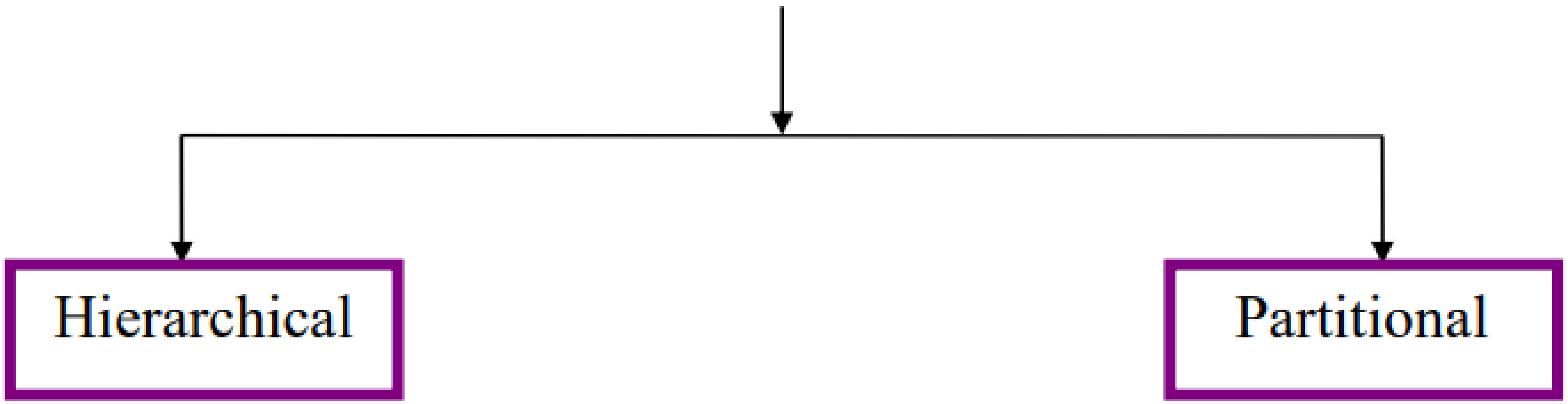
Clustering Purpose

- Preprocessing stage to index, compress, or reduce the data
- Representing high-dimensional data in a low-dimensional space

Clustering Applications

- Information retrieval (search and browsing)
- Cluster users of social networks by interest (community detection).
- Bioinformatics
- Market segmentation

Categorization of Clustering Algorithms



Partitional Algorithms

- Objective based clustering
  - K-means
  - EM-style algorithm for clustering for mixture of Gaussians

Partitional Clustering

$\mathcal{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$

$\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K\}$

- $\forall j, \mathcal{C}_j \neq \emptyset$
- $\bigcup_{j=1}^{K} \mathcal{C}_j = \mathcal{X}$
- $\forall i, j,\ \mathcal{C}_i \cap \mathcal{C}_j = \emptyset$

Objective Based Clustering

- k-median: find center pts **c**1, **c**2, … , **c**K to minimize
- k-means: find center pts **c**1, **c**2, … , **c**K to minimize

$$\sum_{i=1}^{N} \min_{j \in 1,\ldots,K} d^2(\boldsymbol{x}^{(i)}, \boldsymbol{c}_j)$$

- k-center: find partition to minimize the maxim radius

K means

- Input: a set $\boldsymbol{x}\,1, \ldots, \boldsymbol{x}\,N$ of data points (in a d-dim feature space) and an integer k
- Output: a set of K representatives $\boldsymbol{c}1, \boldsymbol{c}2, \ldots, \boldsymbol{c}K \in \mathbb{R}$ as the cluster representatives
- Objective: choose $\boldsymbol{c}1, \boldsymbol{c}2, \ldots, \boldsymbol{c}K$ to minimize:

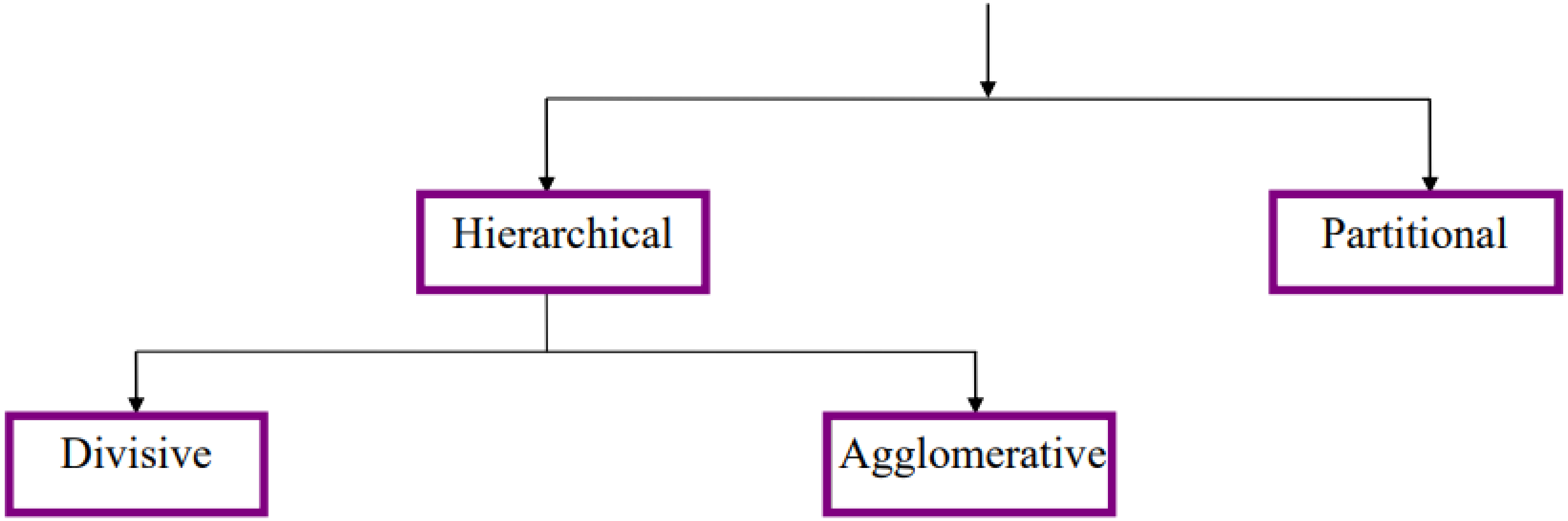
$$\sum_{i=1}^{N} \min_{j \in 1,\ldots,K} d^2(\boldsymbol{x}^{(i)}, \boldsymbol{c}_j)$$

Advantages and disadvantages

- Strength
  - It is a simple method
  - Relatively efficient: $O(tKNd)$, where t is the number of iterations
- Weakness
  - Need to specify K, the number of clusters, in advance
  - Works for numerical data.What about categorical data?

Hierarchical Clustering

- Hierarchical Clustering: Clusters contain sub-clusters and subclusters themselves can have sub-sub-clusters, and so on



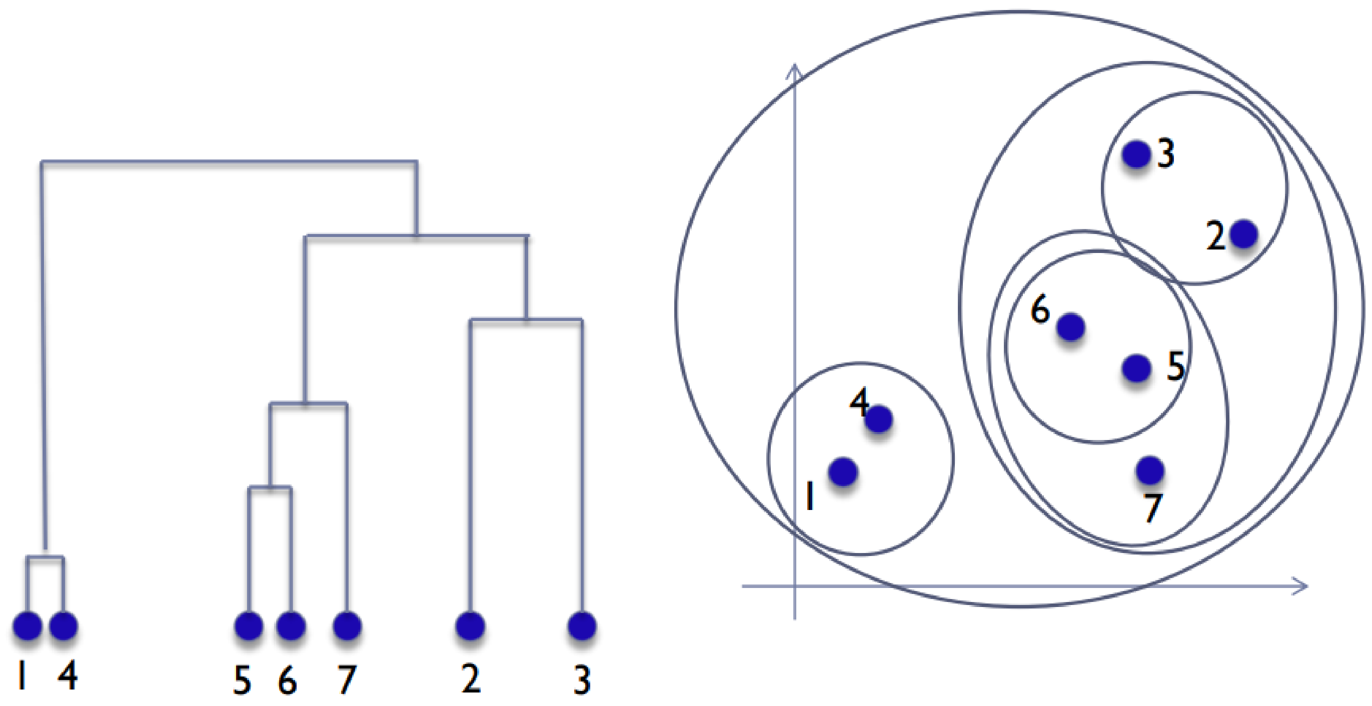
Hierarchical Clustering

- Agglomerative (bottom up):
  - Starts with each data in a separate cluster
  - Repeatedly joins the closest pair of clusters, until there is only one cluster

- Divisive (top down):
  - Starts with the whole data as a cluster
  - Repeatedly divide data in one of the clusters until there is only one data in each cluster

Example

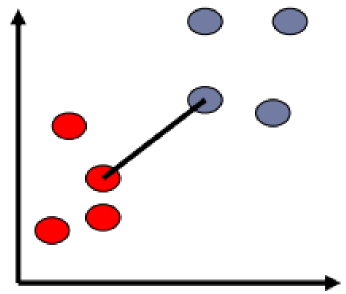
- Hierarchical Agglomerative Clustering (HAC)



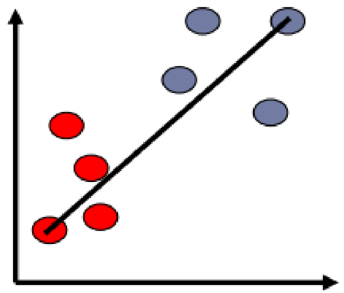
Distances between Cluster Pairs

- Single-link
- Complete-link
- Centroid
- Average-link

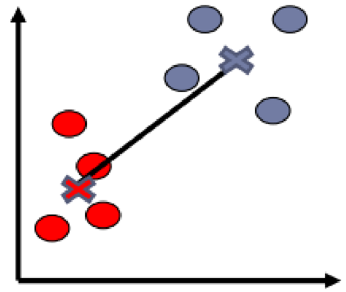
Distances between Cluster Pairs



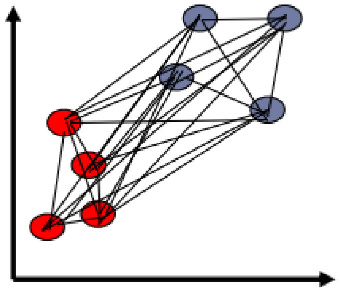
Single-link



Complete-link



Ward's



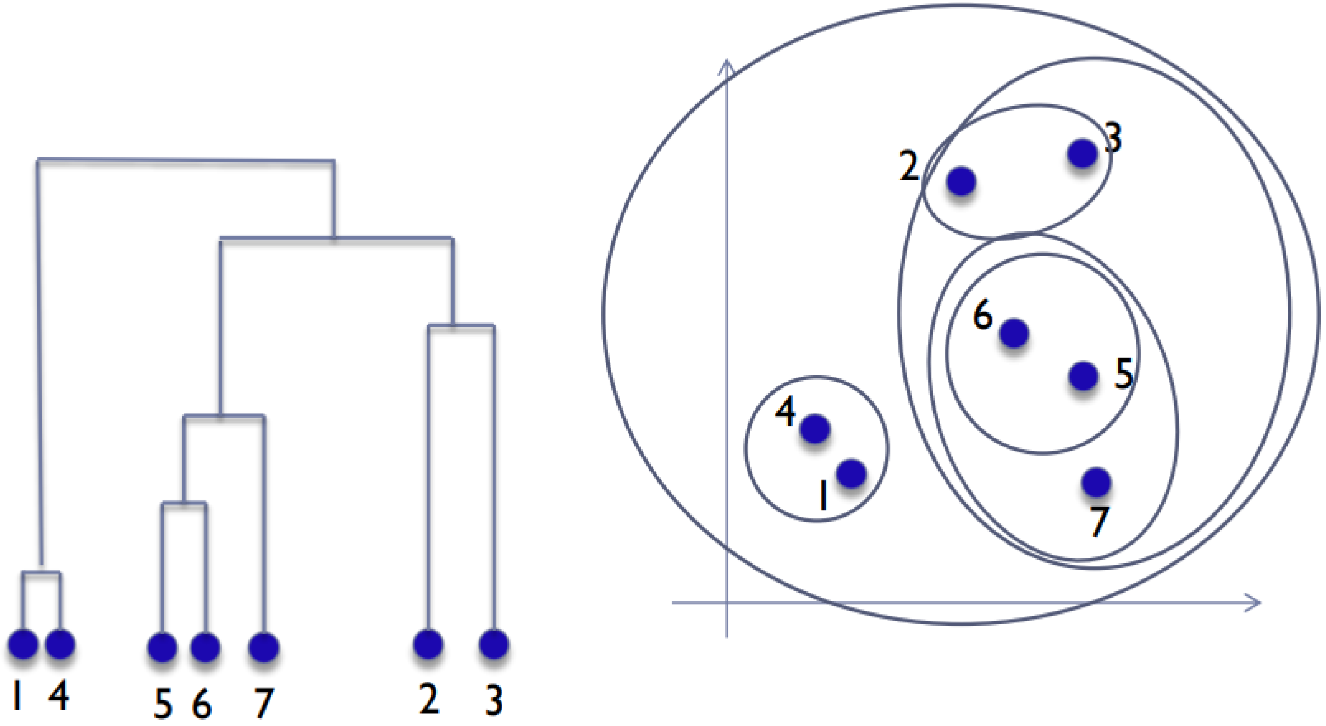
Average-link

Single Linkage

- The minimum of all pairwise distances between points in the two clusters:

$$dist_{SL}(\mathcal{C}_i, \mathcal{C}_j) = \min_{\boldsymbol{x} \in \mathcal{C}_i,\, \boldsymbol{x}' \in \mathcal{C}_j} dist(\boldsymbol{x}, \boldsymbol{x}')$$

Single-Link



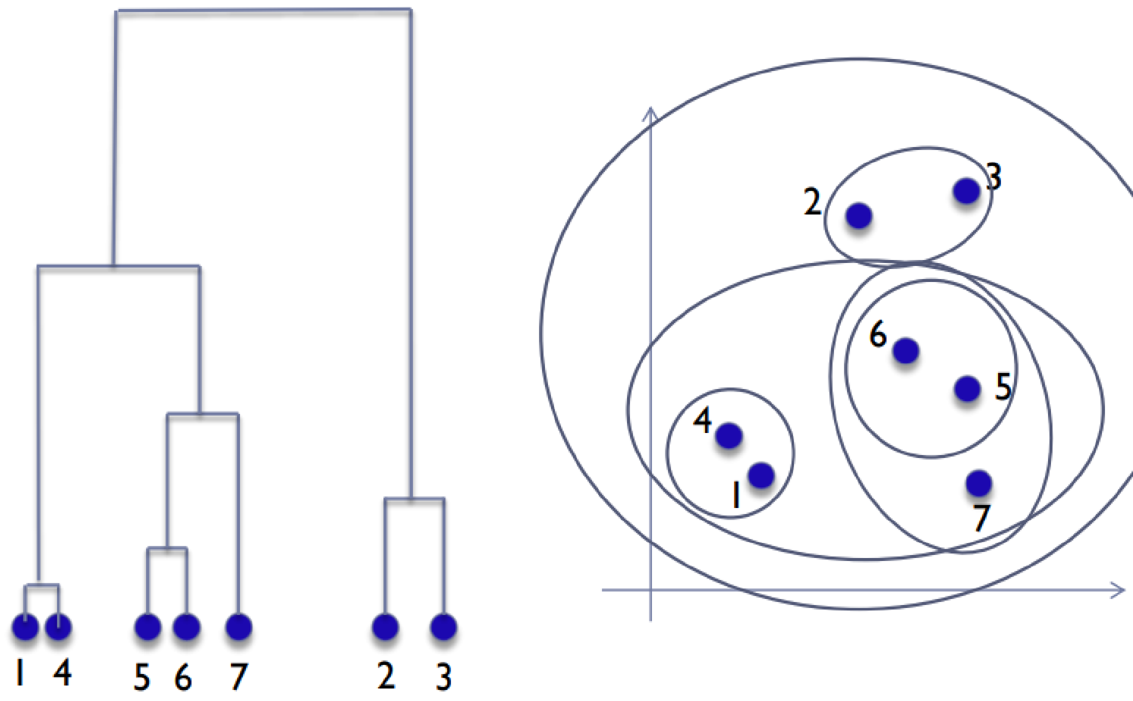
keep max bridge length as small as possible.

Complete Linkage

- The maximum of all pairwise distances between points in the two clusters:

$$dist_{CL}(\mathcal{C}_i, \mathcal{C}_j) = \max_{\boldsymbol{x} \in \mathcal{C}_i, \boldsymbol{x}' \in \mathcal{C}_j} dist(\boldsymbol{x}, \boldsymbol{x}')$$

Complete Link



Ward's method

- The distances between centers of the two clusters

$$dist_{Ward}(\mathcal{C}_i, \mathcal{C}_j) = \frac{|\mathcal{C}_i||\mathcal{C}_j|}{|\mathcal{C}_i| + |\mathcal{C}_j|} dist(\boldsymbol{c}_i, \boldsymbol{c}_j)$$

- Merge the two clusters such that the increase in k-means cost is as small as possible
- Works well in practice.

K-means vs Hierarchical

- Time cost:
  - K-means is usually fast while hierarchical methods do not scale well
- Human intuition
- Choosing of the number of clusters
  - There is no need to specify the number of clusters in advance for hierarchical methods