HAC ALGORITHM

SYRACUSE UNIVERSITY
School of Information Studies

K-MEANS

Process

Problem

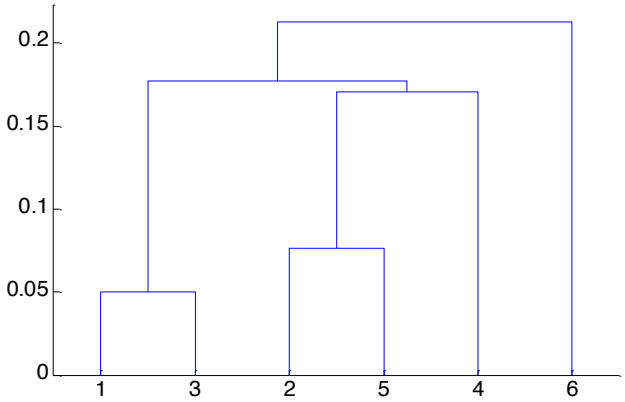
- Fixed number of clusters
- Initial choice of centroid

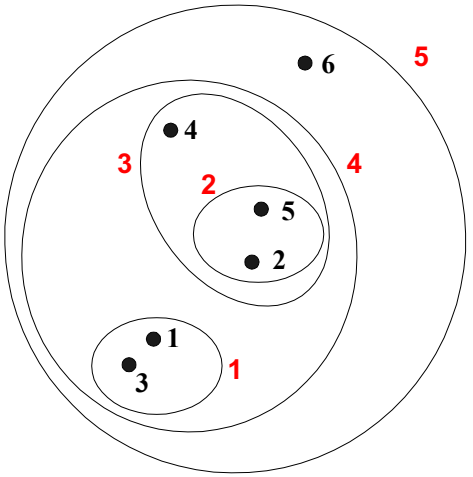
HIERARCHICAL CLUSTERING

Produces a set of nested clusters organized as a hierarchical tree

Can be visualized as a dendrogram:

A treelike diagram that records the sequences of merges or splits

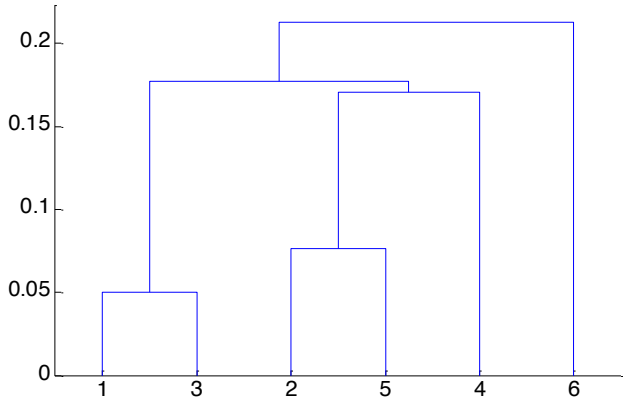




DENDROGRAM: SHOWS HOW THE CLUSTERS ARE MERGED

Decompose data objects into a several levels of nested partitioning (**tree** of clusters), called a **dendrogram**.

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



STRENGTHS OF HIERARCHICAL CLUSTERING

Do not have to assume any particular number of clusters.

- Any desired number of clusters can be obtained by "cutting" the dendrogram at the proper level.

They may correspond to meaningful taxonomies.

AGGLOMERATIVE CLUSTERING ALGORITHM

More popular hierarchical clustering technique:

Basic algorithm is straightforward.

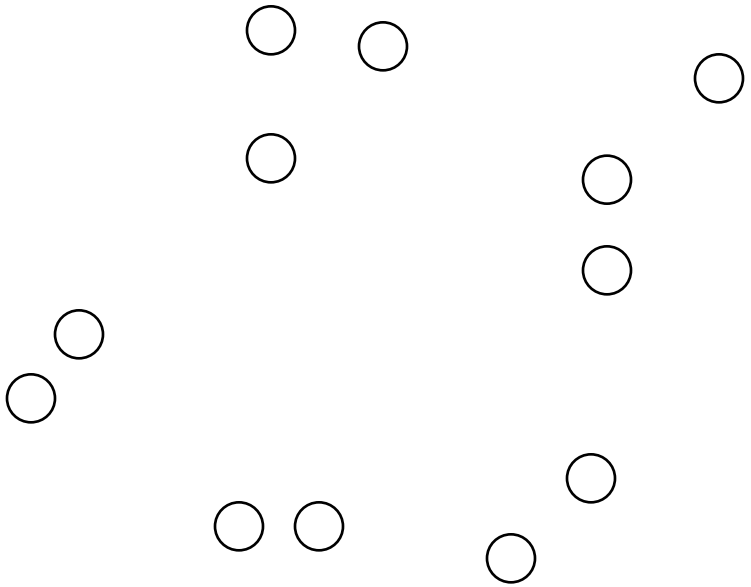
1. Let each data point be a cluster.
2. Compute the distance matrix.
3. **Repeat.**
4. Merge the two closest clusters.
5. Update the distance matrix …
6. **… until** only a single cluster remains.

Key operation is the computation of the distance of two clusters.

Different approaches to defining the distance between clusters distinguish the different algorithms.

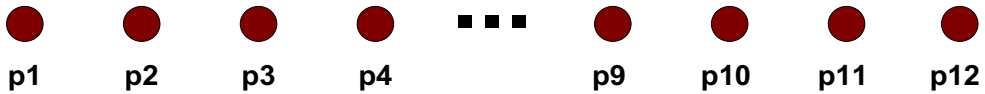
STARTING SITUATION

Start with clusters of individual points and a distance matrix.

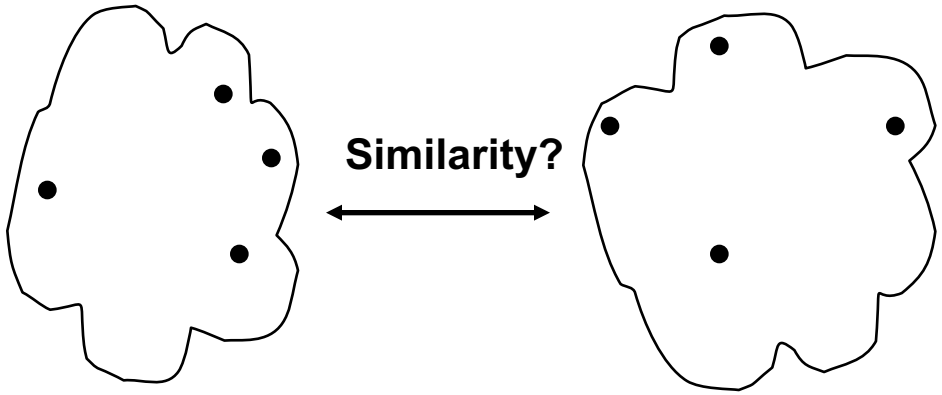


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Distance Matrix



HOW TO DEFINE INTERCLUSTER DISTANCE



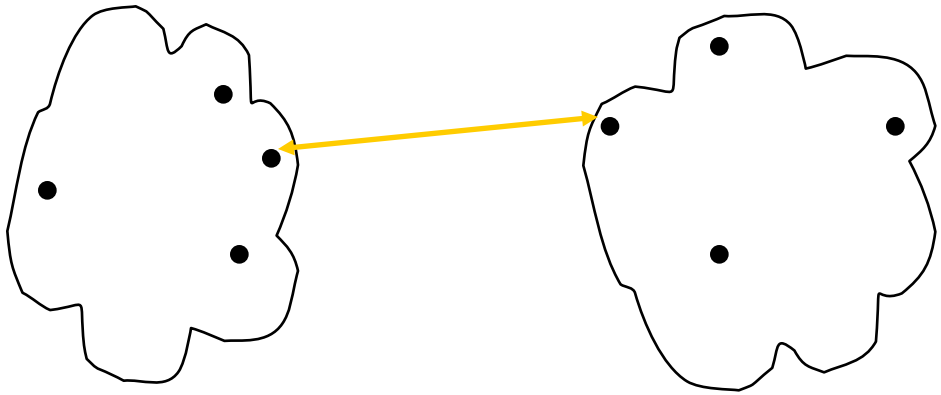
Single linkage

Complete linkage

Average linkage

Centroid linkage

HOW TO DEFINE INTERCLUSTER DISTANCE



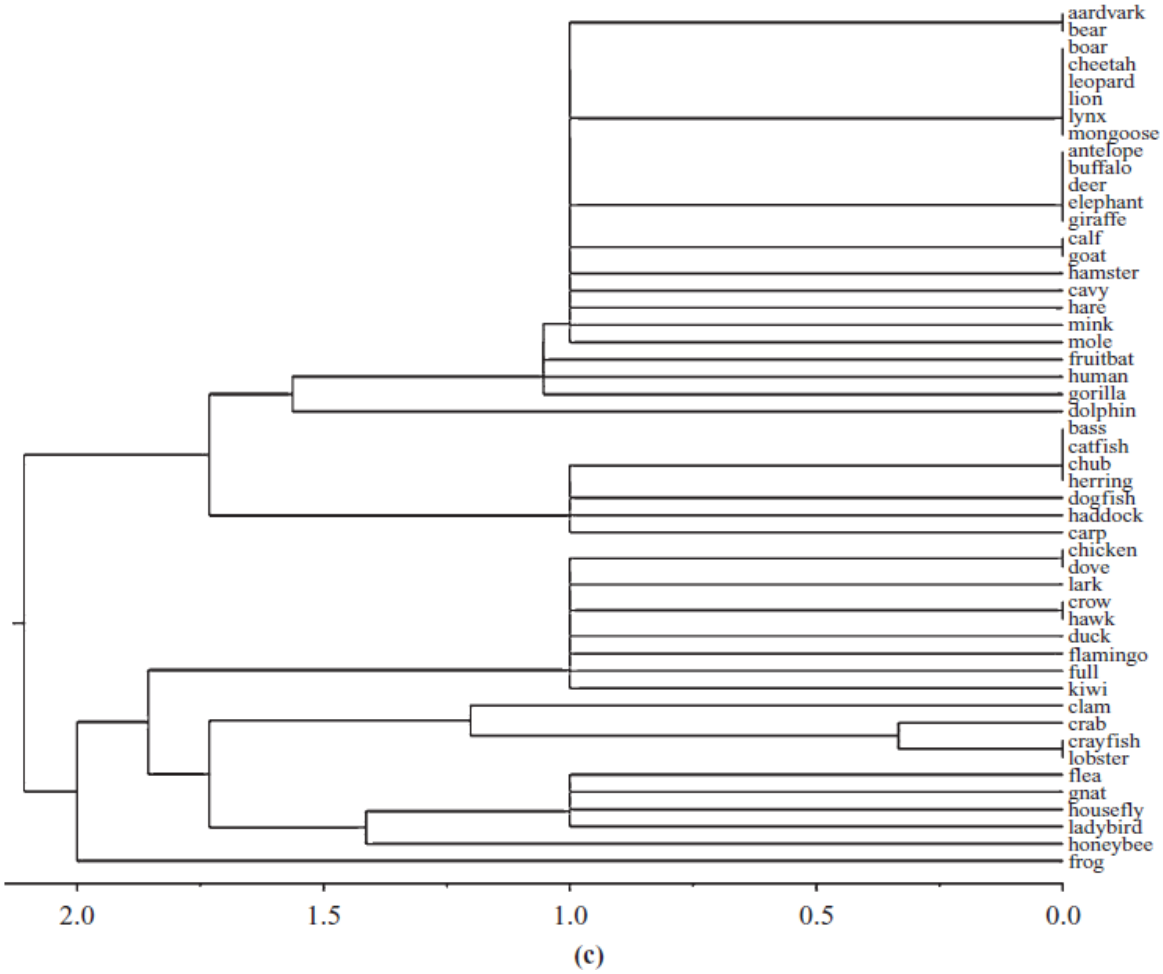
Single linkage:

- Minimum distance between two clusters
- The distance between a pair of closest members

Pros and cons:

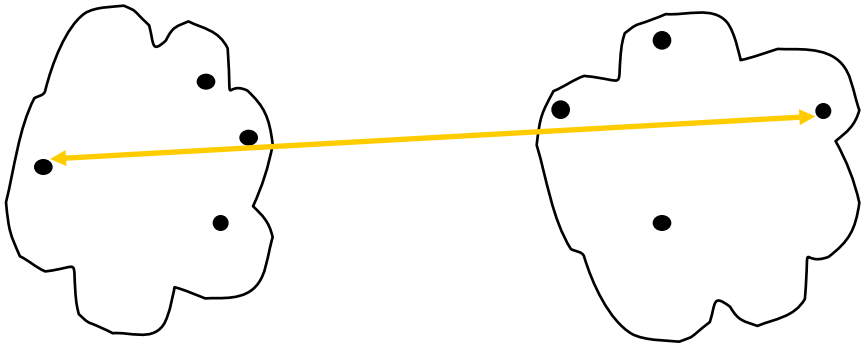
- Depends only on the distance ordering
- Sensitive to outliers
- Clusters with large diameters

SINGLE LINKAGE



(c)

HOW TO DEFINE INTERCLUSTER DISTANCE (CONT.)



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Distance Matrix

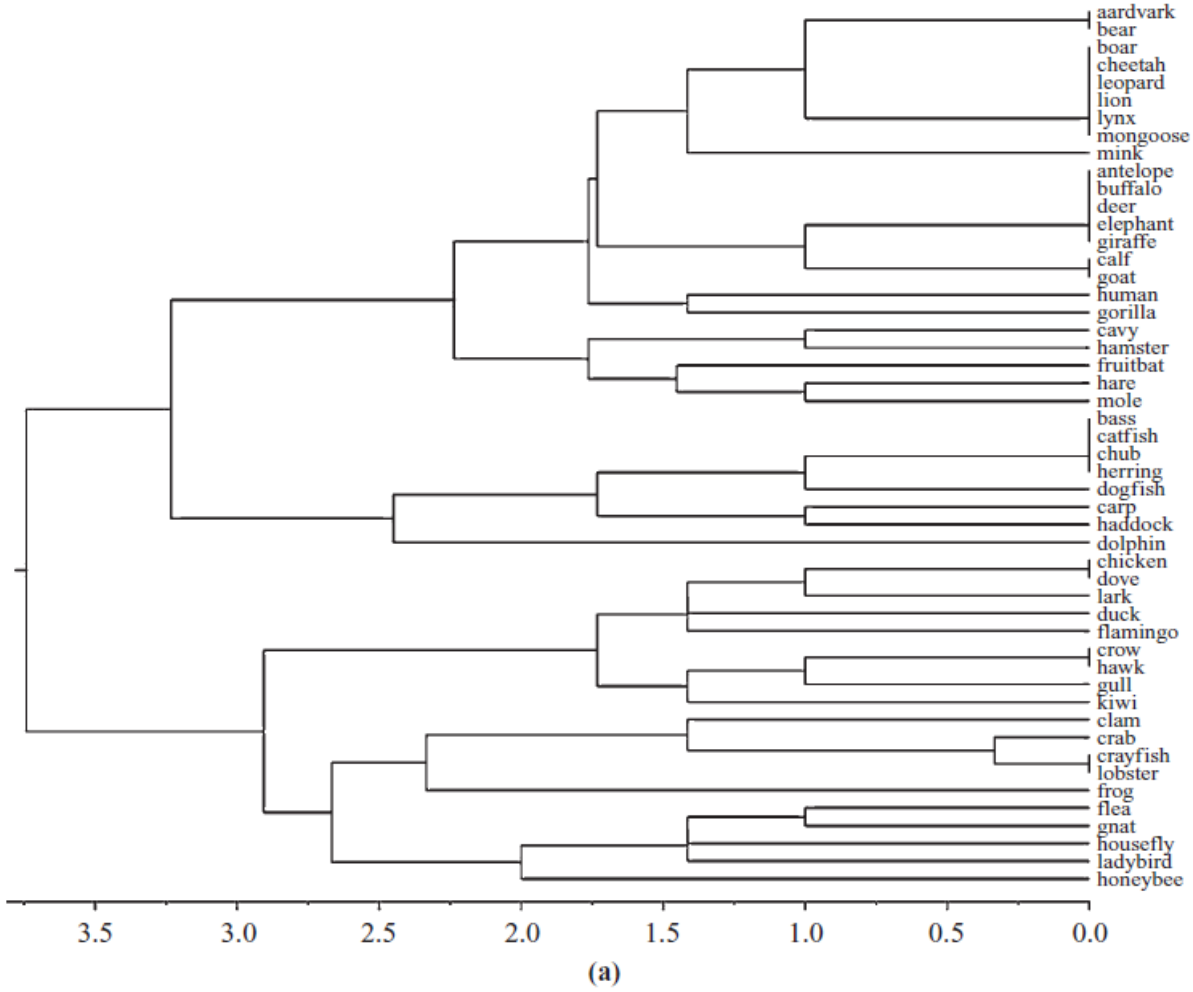
Complete linkage:

- Maximum distance between clusters
- The distance between a pair of farthest members

Pros and cons:

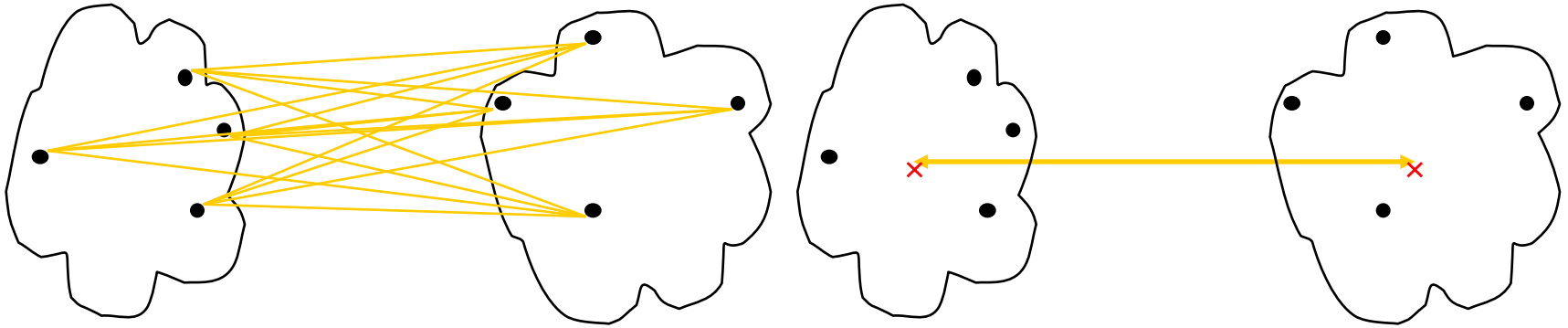
- Depends only on the distance ordering
- Sensitive to outliers
- Clusters with small diameters

COMPLETE LINKAGE



(a)

HOW TO DEFINE INTERCLUSTER DISTANCE (CONT.)



To overcome sensitivity to outliers –

Average linkage: The average distance between each pair of members of the two clusters

Centroid linkage: Distance between two centroids

HIERARCHICAL CLUSTERING: PROBLEMS AND LIMITATIONS

Once a decision is made to combine two clusters, it cannot be undone.

No objective function is directly minimized.

Different linkage calculations have problems with one or more of the following:

- Sensitivity to noise and outliers
- Difficulty handling different-sized clusters and convex shapes
- Breaking large clusters