Cluster analysis - basics

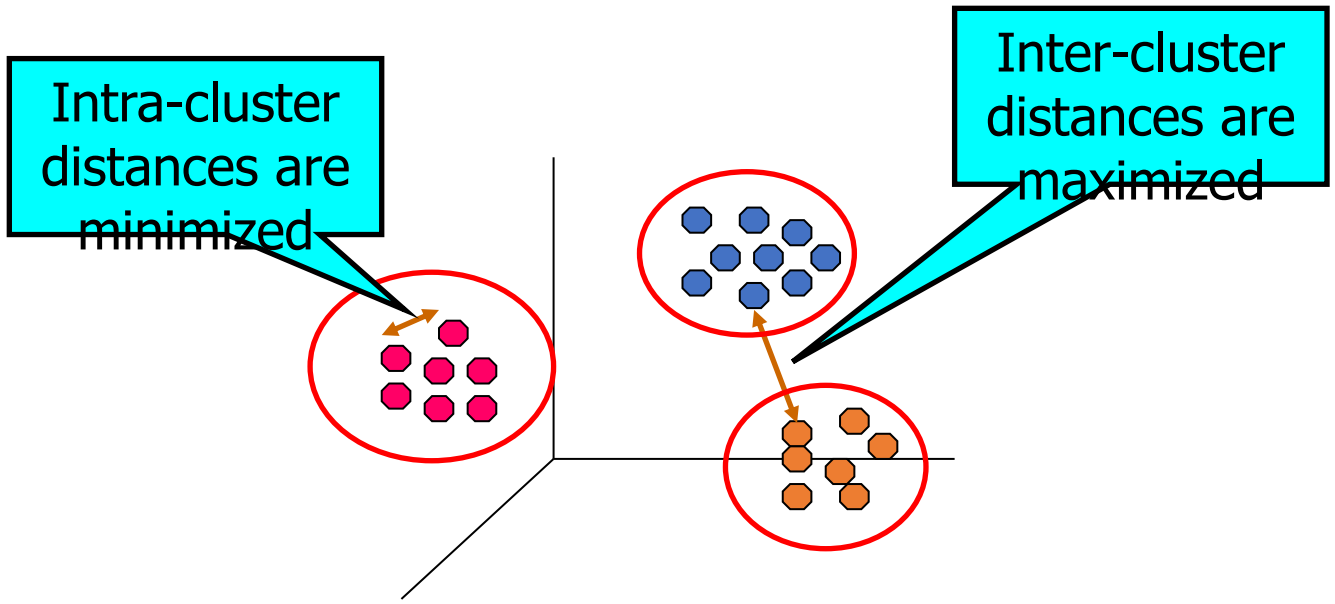
Dr. Huiping Cao

Introduction – what is cluster analysis

- **Unsupervised** learning: no predefined classes
- Given a collection of data objects, find the **natural grouping** in data so that items in the same cluster are more similar to each other than to those from different clusters.
- It allows us to discover **hidden structures** in data.

What is good clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity: similar to one another within the same cluster
  - low inter-class similarity: dissimilar to the objects in other clusters



Applications of Cluster Analysis

- **Typical applications**
  - As a stand-alone tool to get insight into data distribution.
  - As a preprocessing step for other algorithms
- **Insight: understanding**
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
  - Segment customers into a small number of groups for marketing activities
- **Insight: summarization**
  - Reduce the size of large data sets

What is not Cluster Analysis?

- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query
  - Groupings are a result of an external specification
  - Clustering is a grouping of objects based on the data
- Supervised classification
  - Have class label information

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

Four Clusters

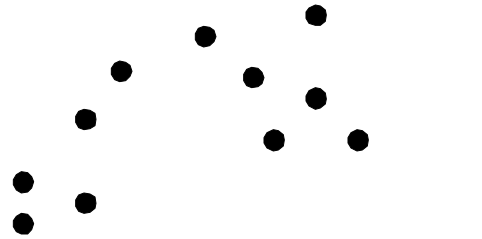
What is good clustering?

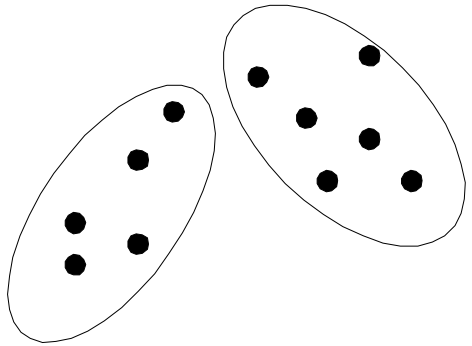
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns

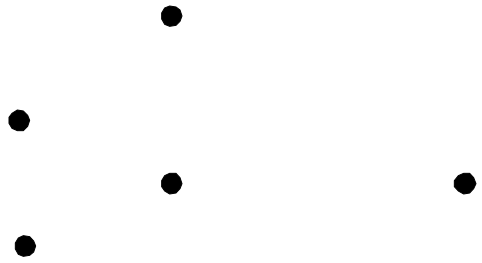
Types of Clusterings

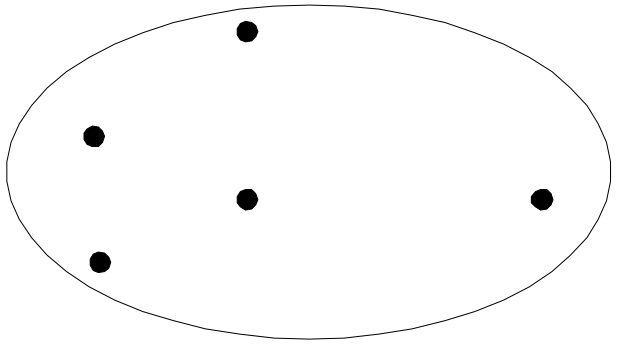
- A clustering is a set of clusters
- Important distinction between hierarchical and partitional sets of clusters
- Partitional Clustering
  - A division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree

Partitional Clustering





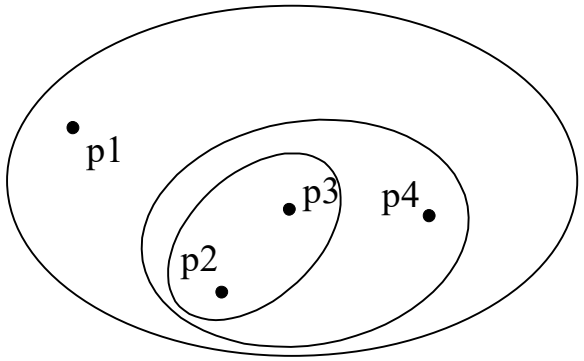




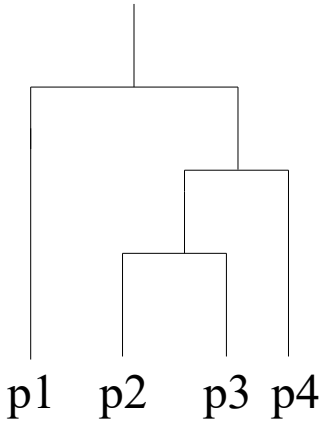
Original Points

A Partitional Clustering

Hierarchical Clustering



Hierarchical Clustering



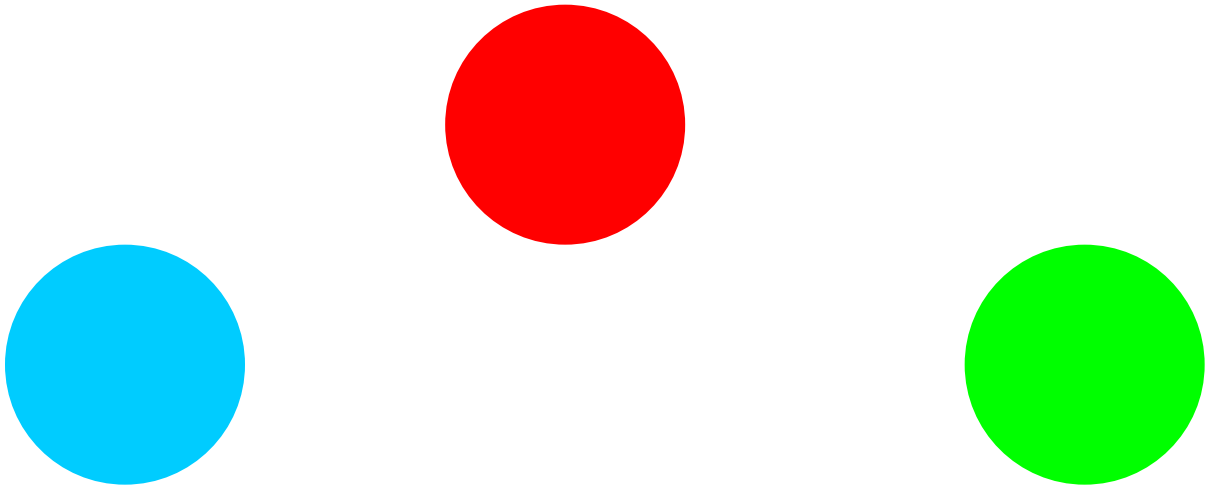
Dendrogram

Types of Clusters

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual

Types of Clusters: Well-Separated

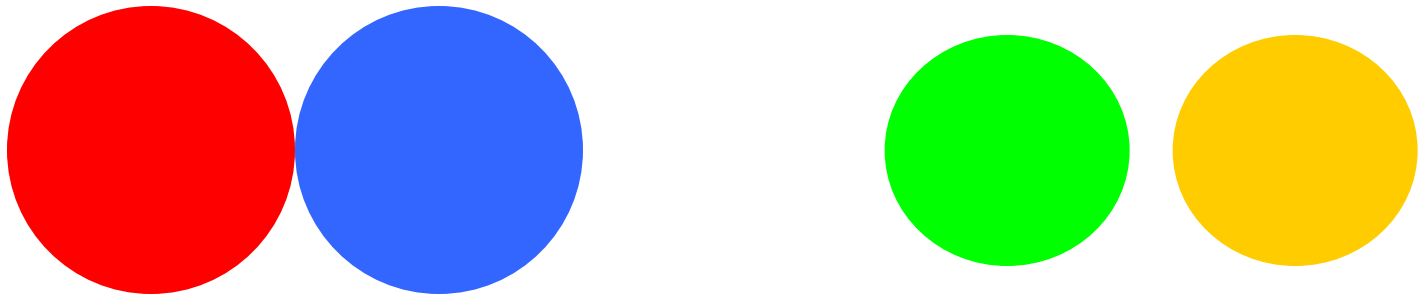
- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

Types of Clusters: Center-Based

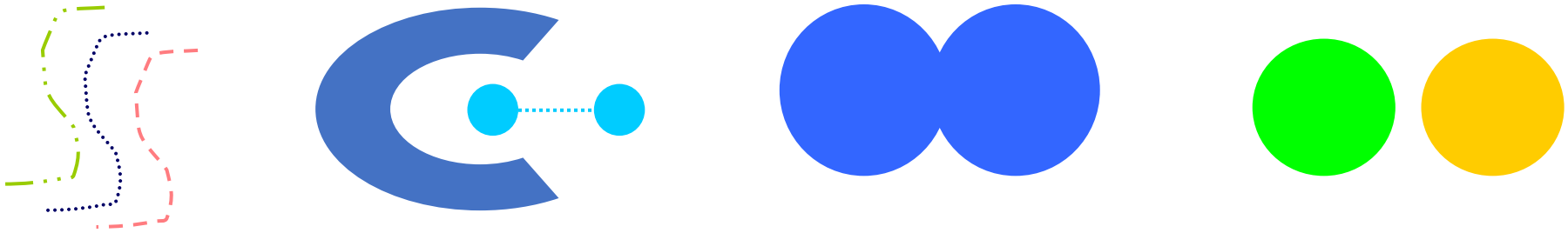
- Center-based
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster
  - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most “representative” point of a cluster



4 center-based clusters

Types of Clusters: Contiguity-Based

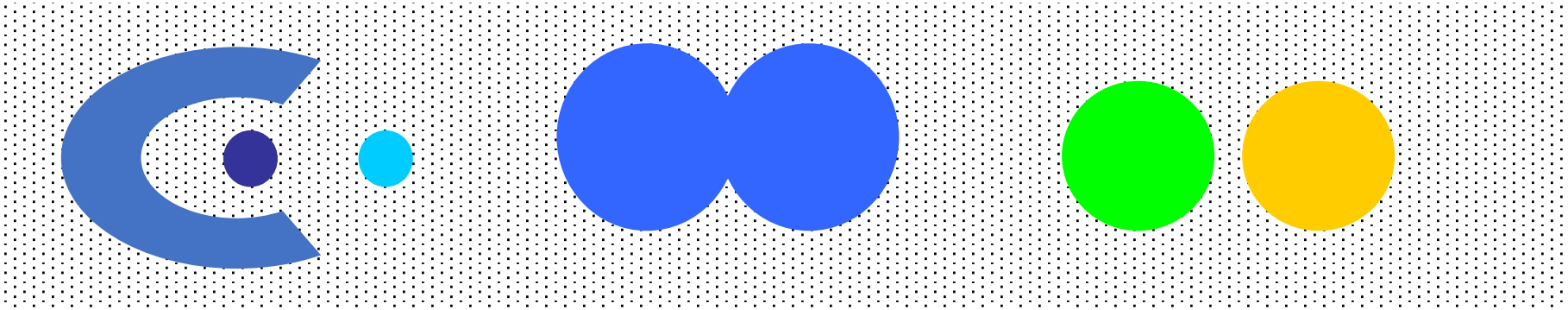
- Contiguous Cluster (Nearest neighbor or Transitive)
  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.



8 contiguous clusters

Types of Clusters: Density-Based

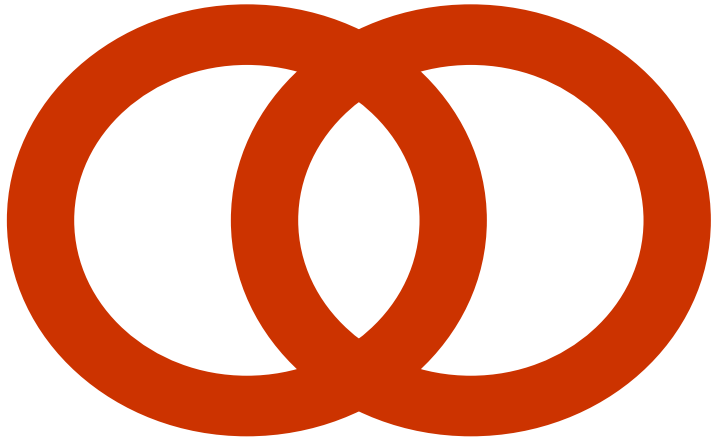
- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

Types of Clusters: Conceptual Clusters

- Shared Property or Conceptual Clusters
  - Finds clusters that share some common property or represent a particular concept.



2 Overlapping Circles

Clustering Algorithms

- K-means and its variants
- Hierarchical clustering
- Density-based clustering

References

- Chapter 11, Sebastian Raschka and Vahid Mirjalili: Python Machine Learning (Machine learning and deep learning with Python, scikit-learn, and TensorFlow), 3rd Edition.
- Chapter 7, Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar: Introduction to Data Mining, 2nd Edition.