Clustering: Basic Concepts

AW

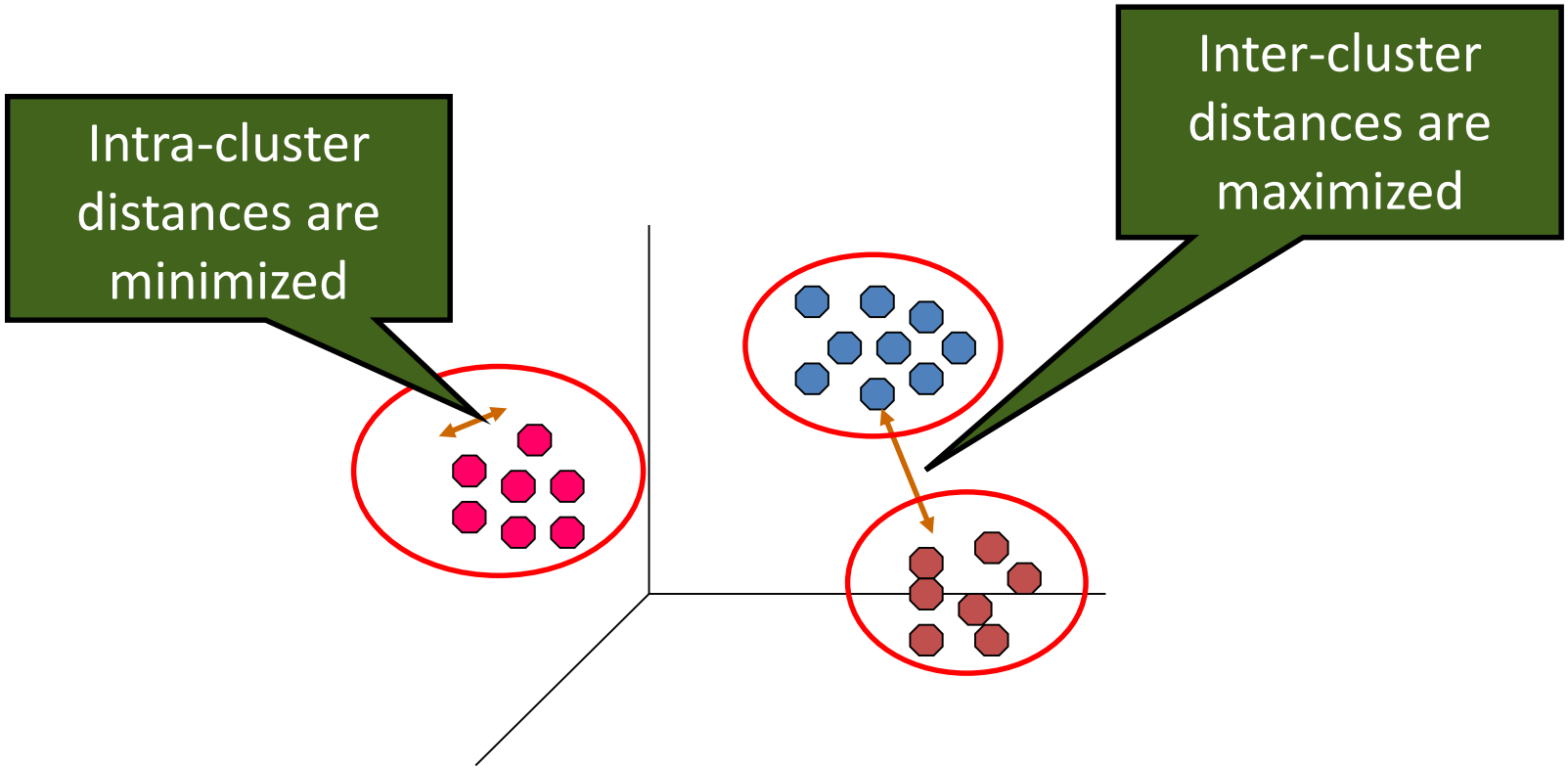
Lecture Overview

1. Cluster Analysis Defined

2. Types of Clustering

What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Basic Clustering Model

- Distance function: $d: X \times X \to \mathbb{R}^{\geq 0}$ non-negative function that is
  - symmetric $d(x, y) = d(y, x)$,
  - satisfies identity $d(x, y) = 0 \Leftrightarrow x \equiv y$ for all $x \in X$)
  - satisfies the triangle inequality, i.e. $d(x, y) \leq d(x, z) + d(z, y)$.
- A similarity function: s: $X \times X \to [0,1]$ that is symmetric and satisfies $s(x, x) = 1$

Input: a set of elements, X, supplied with one of the following:

- A distance function,
- Similarity function
- *Optional input:* some clustering algorithms expect the number of required clusters.

Output: a partition of the domain set X into subsets, i.e. $C = (C_1, \ldots, C_k)$ such that $X = \bigcup_{i=1}^{k} C_i$ and for all $i \neq j$, $C_i \cap C_j = \emptyset$.

Most Common Measures

Data set has real features: $X \subset \mathbb{R}^n$

- Minkowski distance: $\|\bar{x} - \bar{y}\|_l = \sqrt[l]{|x_1 - y_1|^l + \cdots + |x_n - y_n|^l}$.
  Special cases:
  - Euclidean: $\|\bar{x} - \bar{y}\| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$
  - Manhattan: $\log_{l \to \infty} \left(\sqrt[l]{|x_1 - y_1|^l + \cdots + |x_n - y_n|^l} \right) = \max_{i \in \{1,\ldots,n\}} |x_i - y_i|$
- Cosine Similarity $csine(\bar{x}, \bar{y}) = \cos\theta = \frac{\bar{x} \cdot \bar{y}}{\|\bar{x}\| \|\bar{y}\|}$
- Data set has Boolean features: $X \subset \mathbb{B}^{\boldsymbol{n}}$
  - Hamming distance: $\|\bar{x} - \bar{y}\| = \sum_{i=1}^{n} |x_i - y_i|$
  - Jaccard Similarity: $J(\bar{x}, \bar{y}) = \frac{\bar{x} \cdot \bar{y}}{n - (\mathbb{1} - \bar{x}) \cdot (\mathbb{1} - \bar{y})}$ where $\mathbb{1}$ is all-1's vector
- Data set is a set of finite strings in alphabet Σ: $X \subseteq \Sigma^*$
  - Edit (Levenstein) distance: $E(s_1, s_2)$ =number of replace, insert and delete operations necessary to convert s_1 into s_2

Cluster Analysis as Learning

Can be interpreted as unsupervised learning:

- Classes (groupings) for training data points are not known
- Number of classes may not be known either
- Class attribute is not known or does not exist

Then clustering is defining a map $f: X \to \{1, \ldots, k\}$ that assigns each point $x \in X$ its cluster value $y \in \{1, \ldots, k\}$

Refinement: partition of X into the different clusters is probabilistic, i.e. output is a function assigning to each domain point, $x \in X$, a vector $\big(Pr_1(x), \ldots, Pr_k(x)\big)$.

Notion of a Cluster can be Ambiguous

How many clusters?

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters?

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters?

Two Clusters?

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters?

Two Clusters?

Four Clusters?

Lecture Overview

1. Cluster Analysis Defined

2. Types of Clustering

Model Modifications

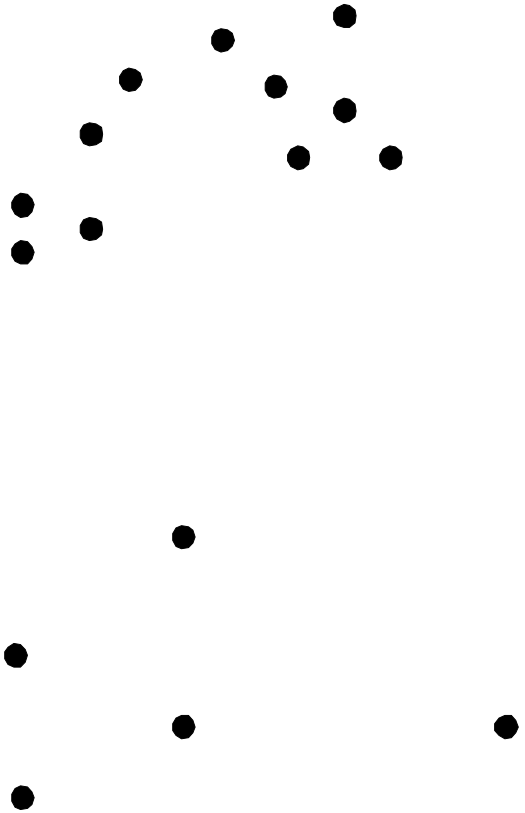
Basic Model: Partitional Clustering:

- Output: A partitioning of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

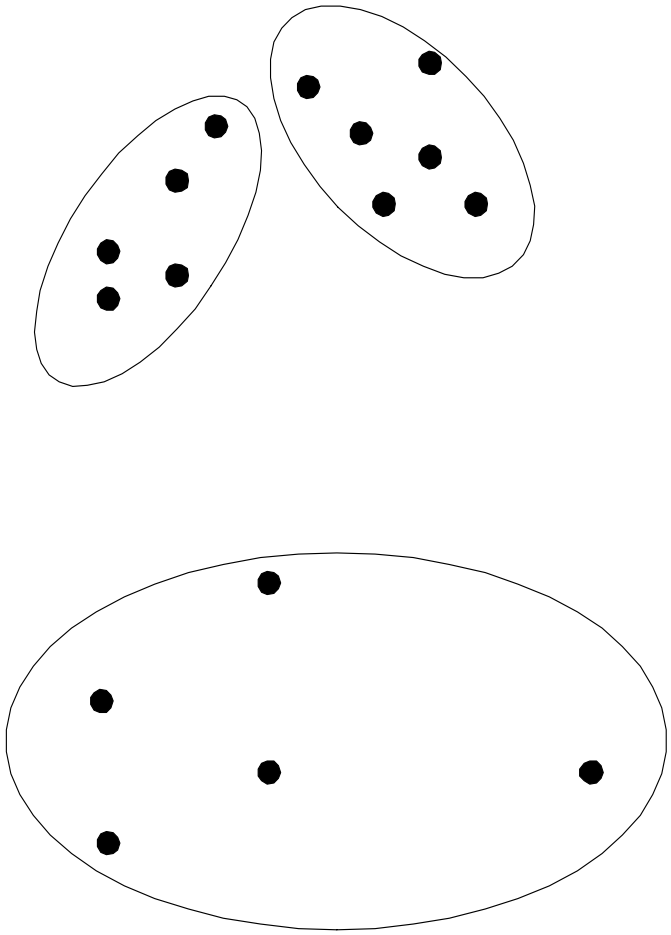
Model Modification: Hierarchical clustering:

- Output: A dendrorgam of clusters, i.e. a set of nested clusters organized as a hierarchical tree of domain subsets, having the singleton sets in its leaves, and the full domain as its root. Each ‘slice’ of a dendrogram is a partitional clustering

Partitional Clustering

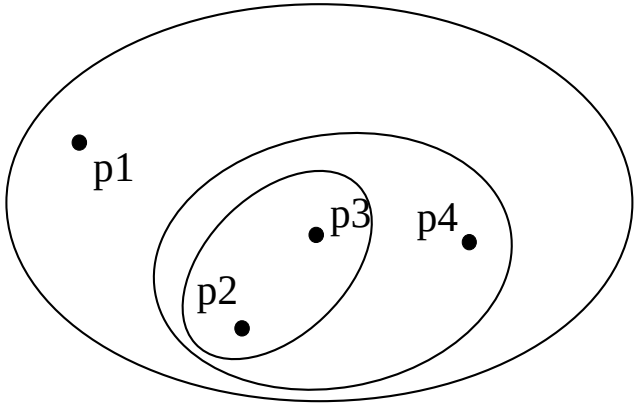


Original Points

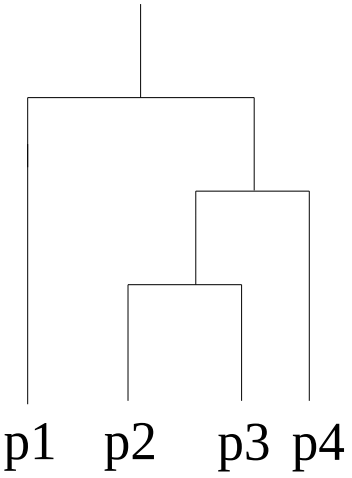


A Partitional Clustering

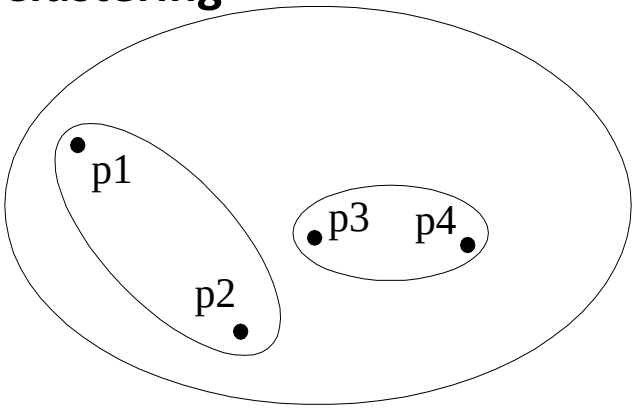
Hierarchical Clustering



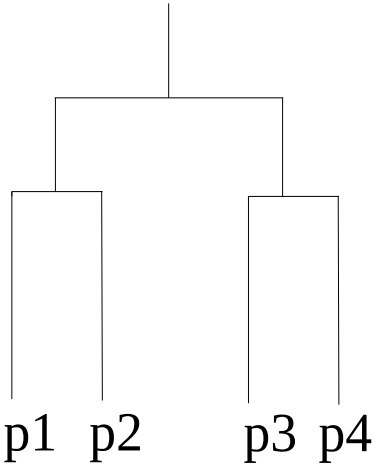
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



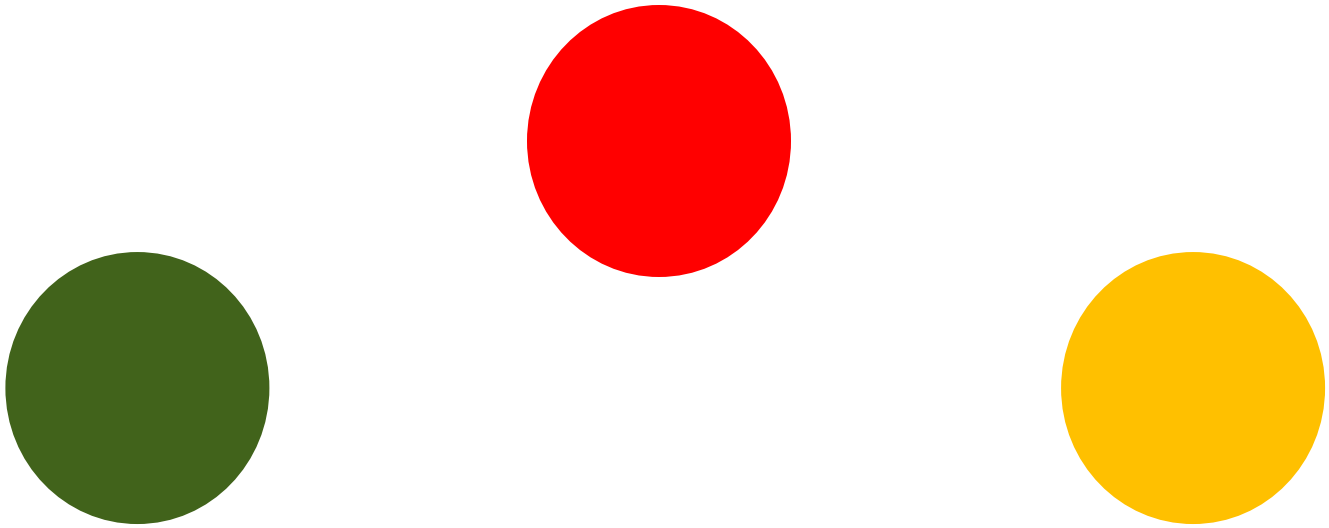
Non-traditional Dendrogram

More Model Modifications

- Basic model: partitioning is based on distance similarity function:
    - Well-separated clusters
    - Center-based clusters (or medoid based)
    - Contiguous clusters
- Modified Models:
    - Density-based clusters
    - Property or Conceptual
    - Described by an Objective Function

Well-Separated Cluusters

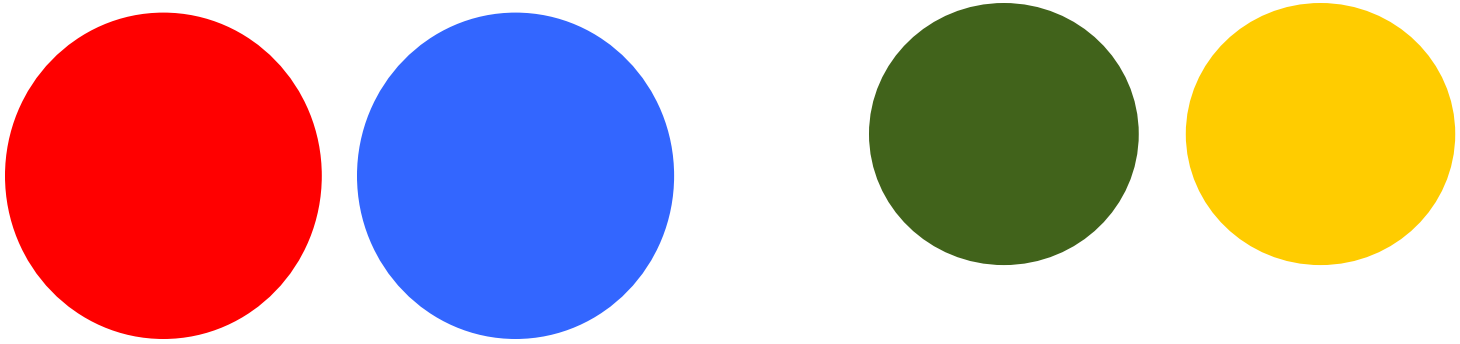
- Well-Separated Clusters:
  - a point belong to a cluster iff its distance (similarity) to ALL points in a cluster is smaller (bigger) than to any other data point



3 well-separated clusters

Center-Based Clusters

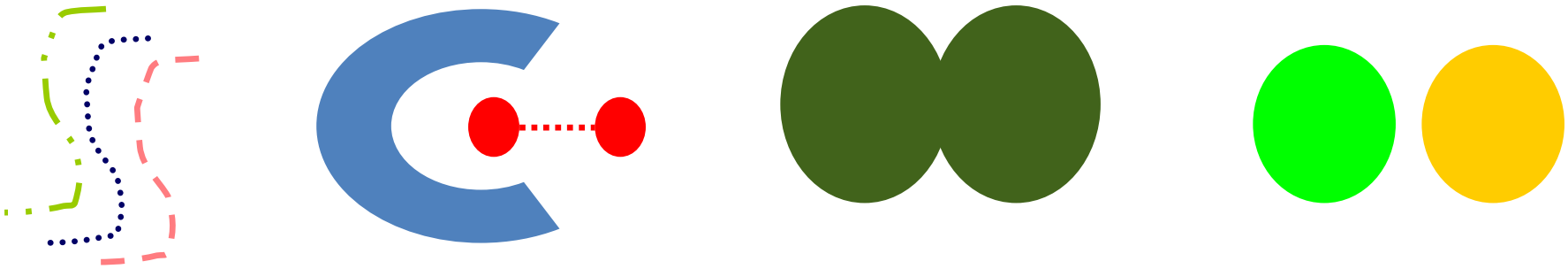
- Center-based
  - A cluster is a set of data points such that a data point in a cluster is closer (more similar) to the “leader” of a cluster, than to a leader of any other cluster
  - The leader of a cluster is often a centroid, the average of all the points in the cluster (may be a point of a domain space that is not a data point), or a medoid, the most “representative” data point of a cluster



4 center-based clusters

Contiguity-Based Clusters

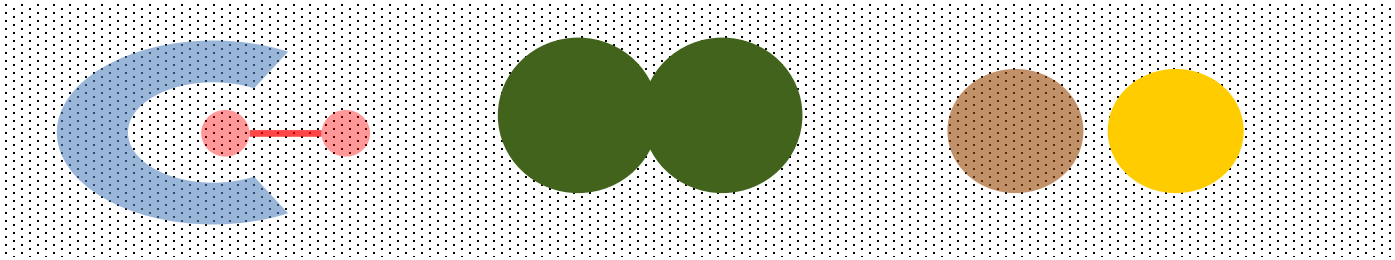
- Contiguous Cluster (Nearest neighbor or Transitive)
  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.



9 contiguous clusters

Density-Based Clusters

- 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.

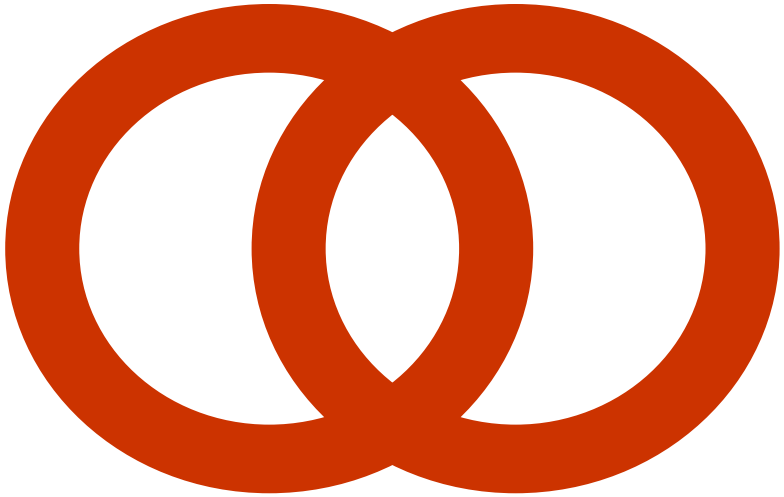


6 density-based clusters

Conceptual Clusters - Patterns

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

.



2 Overlapping Circles

Clusters by Objective Function

- Clusters Defined by an Objective Function
  - Clusters defined as a solution to optimization problem (e.g. maximize total between cluster distance).
  - Always have brute force solution: enumerate all possible ways of dividing the points into clusters and evaluate the `goodness' of each potential set of clusters by using the given objective function. (NP Hard)
  - Can have global or local objectives.
    - Hierarchical clustering typically have local objective functions (i.e. global objective is satisfied when some local objective is)
    - Partition-inducing algorithms typically have global objectives (greedy algorithms can only be approximate)

Reading

- TSKK 7.1