Clustering Analysis

Lesson 7 – Section 4

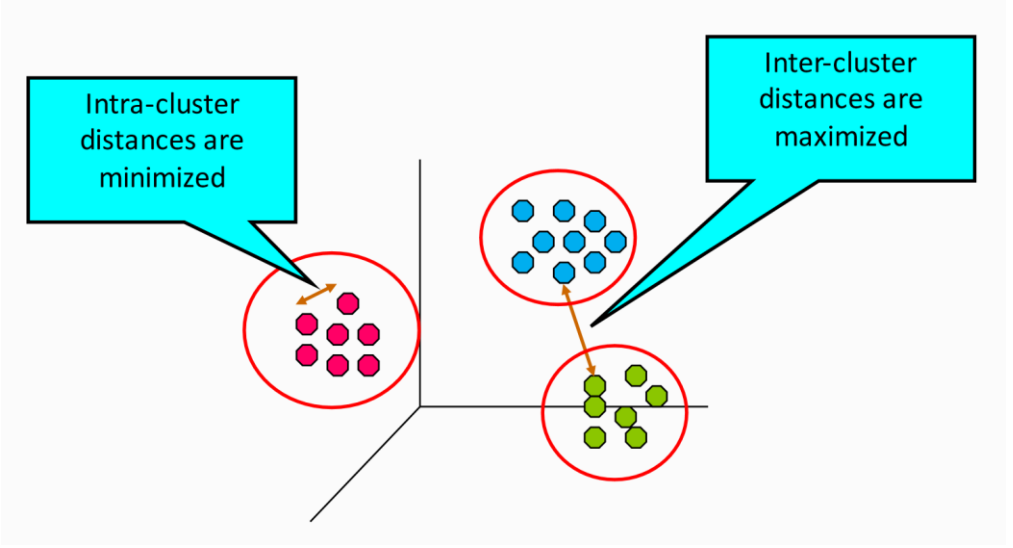
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Some Definitions

- Cluster: a group of observations that are similar with each other
- Clustering Analysis: Find groups of objects (observations), such that observations within each cluster are similar with each other, and observations in different clusters are dissimilar with each other
  - Intra-cluster **similarity** is higher than inter-cluster **similarity**
- **Unsupervised Machine Learning:** there is no labels telling you which observations should be in the same cluster. You need to determine the clustering pattern based on the features that describe the objects

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

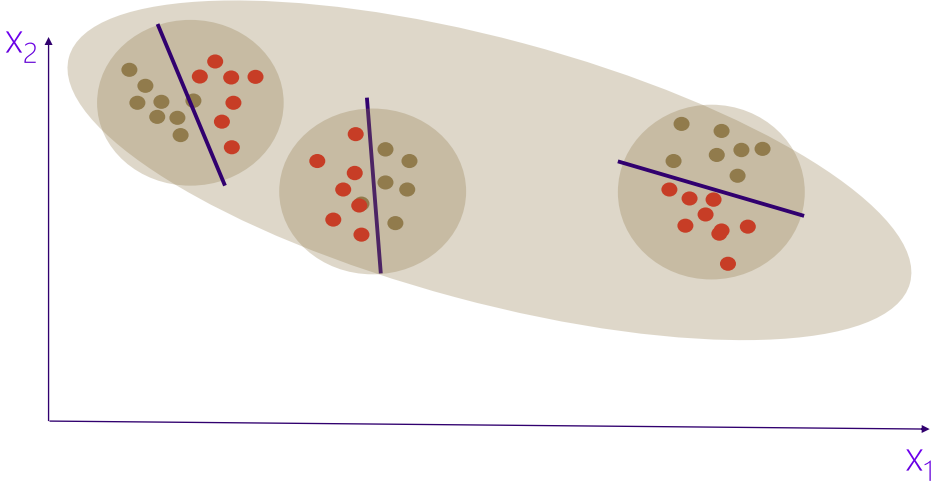
A Toy Example



PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Unsupervised Machine Learning

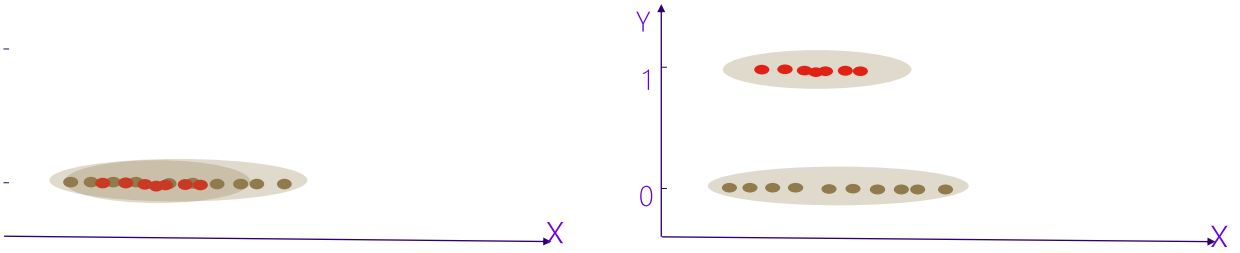
- Might be useful for supervised machine learning



PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

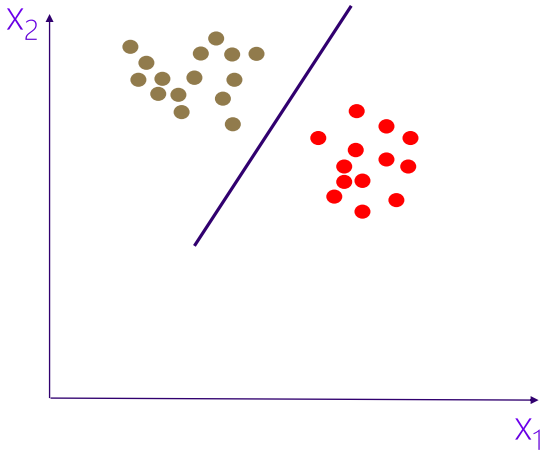
Clustering Analysis for Supervised Machine Learning (Classification) Tasks

- Do not include the label column in your clustering analysis
  - Since the observations are always clustered at the dimension of the label column
  - You will always see clustering pattern in this dimension



PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Can Be Used to Assess the Difficulty of your Classification Task



A dual assessment is the relevance of the data (feature) set with your classification task

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

How do we define "similarity"?

Goal: Group "similar" data

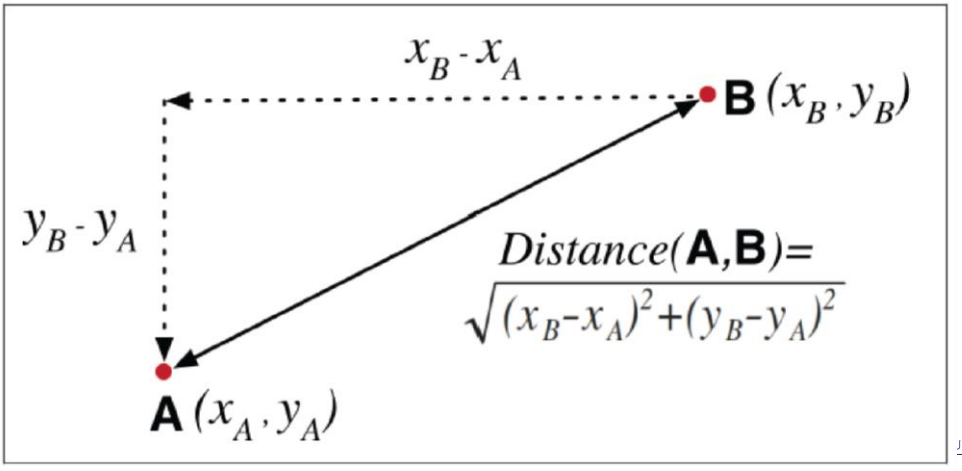
- Similarity measure more important than clustering algorithm

Depends on what to emphasize in the data:

- Data reduction
- "natural clusters"
- "useful" clusters
- Outlier detection
- Et cetera

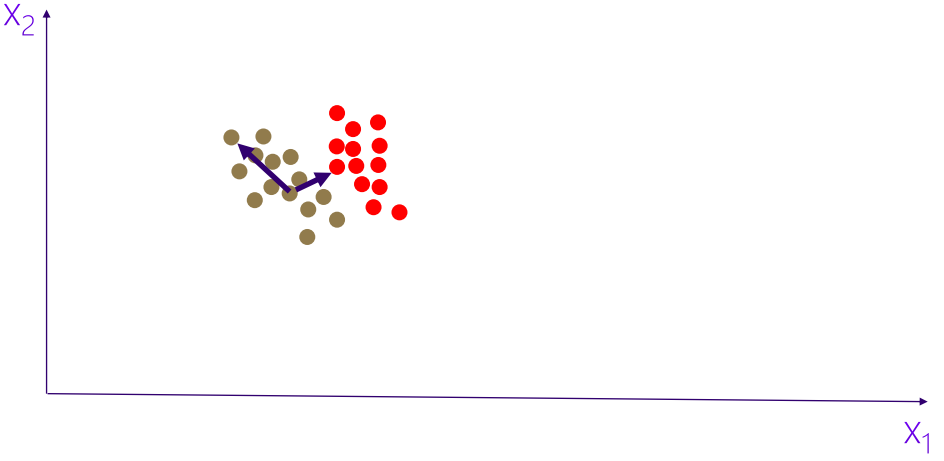
Similarity Measures

Euclidean Distance



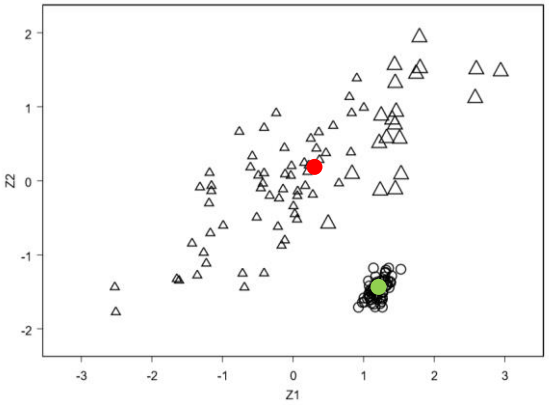
Problem of Euclidean Distance as Similarity Measurement

- It assumes that clusters are spheres



Mahalanobis Distance

$$D(\mathbf{x}_i, \mathbf{c}_k) = \{(\mathbf{x}_i - \mathbf{c}_k)' \mathbf{S}^{-1}_k (\mathbf{x}_i - \mathbf{c}_k)\}^{1/2}$$



Other Similarity Measures

Manhattan Distance

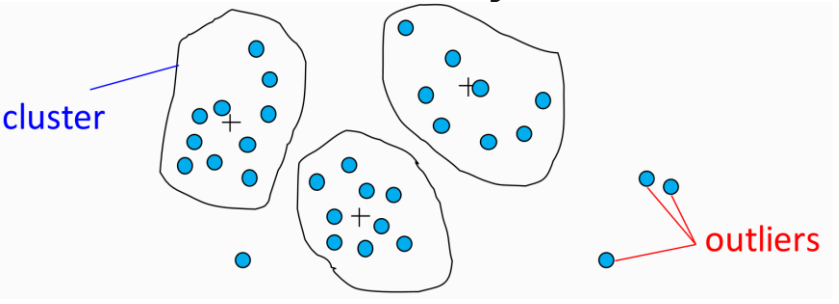
$$d_{\text{Manhattan}}(\mathbf{X},\mathbf{Y}) = \| \mathbf{X} - \mathbf{Y} \|_1 = | x_1 - y_1 | + | x_2 - y_2 | + \cdots$$

Jaccard Distance

$$d_{\text{Jaccard}}(X, Y) = 1 - \frac{| X \cap Y |}{| X \cup Y |}$$

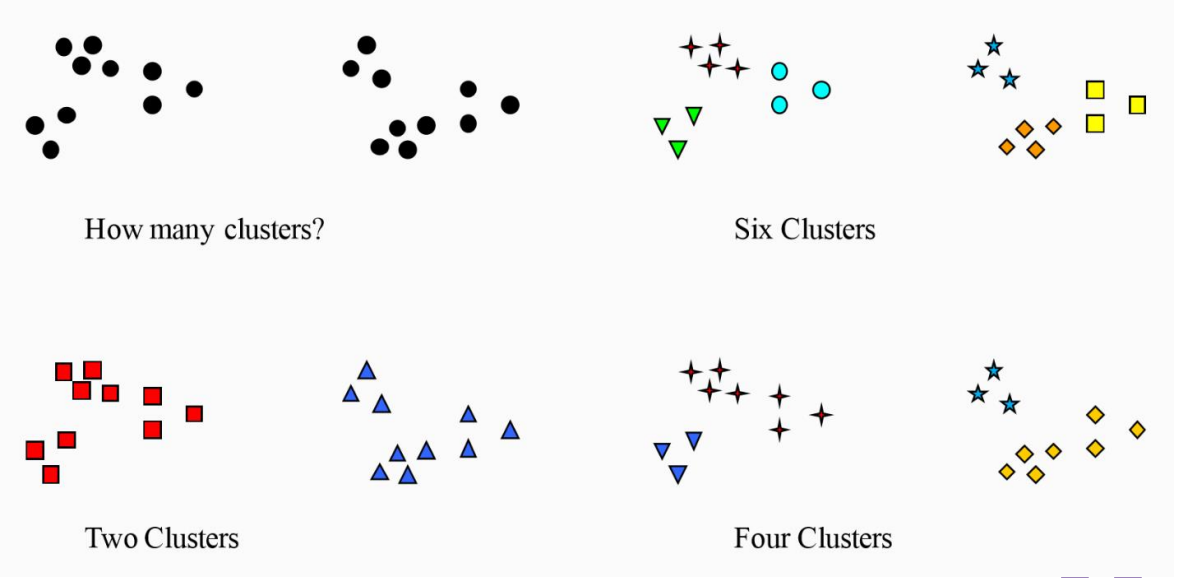
Outliers

Outliers are objects that **do not belong to any cluster** or form clusters of very small cardinality



In some applications we are interested in discovering outliers, not clusters (**outlier analysis**)

Notion of a Cluster can be Ambiguous



Types of Clusterings

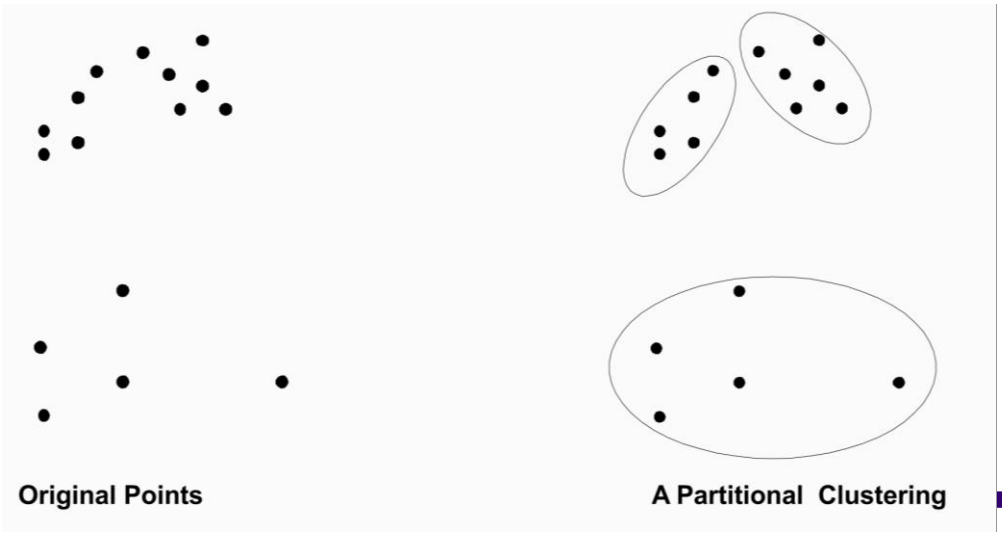
Partitional Clustering

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

Hierarchical Clustering

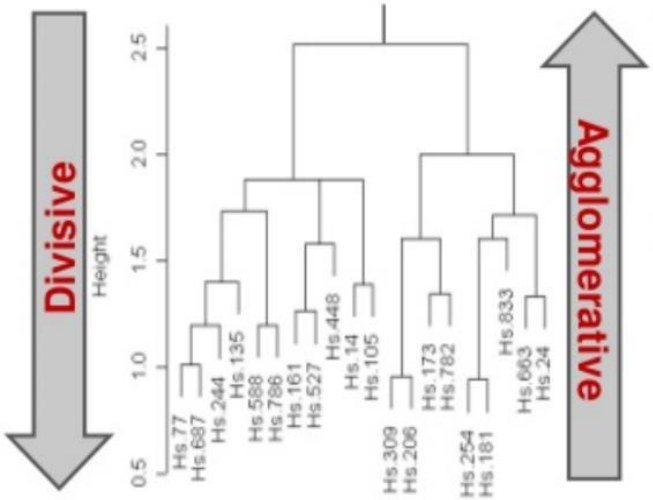
- A set of nested clusters organized as a hierarchical tree

Partitional Clustering



W

Hierarchical Clustering Tree, Dendogram



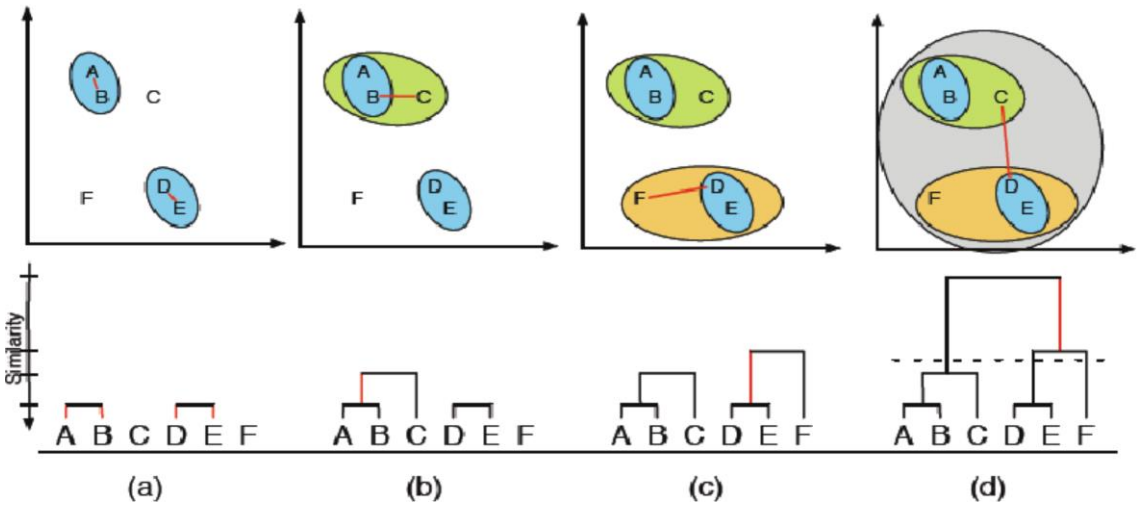
The hierarchy of clustering is given as a clustering tree or dendrogram

- leaves of the tree represent the individual objects
- internal nodes of the tree represent the clusters

Two main types of hierarchical clustering

- agglomerative (bottom-up)
  - place each object in its own cluster (a singleton)
  - merge in each step the two most similar clusters until there is only one cluster left or the termination condition is satisfied
- divisive (top-down)
  - start with one big cluster containing all the objects
  - divide the most distinctive cluster into smaller clusters and proceed until there are *n* clusters or the termination condition is satisfied

Example of Agglomerative Hierarchical Clustering



PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

How to Determine Which Two Clusters to Merge in Agglomerative Hierarchical Clustering Analysis?

- Each cluster (before merge) might have multiple objects
- We need a measure to describe the similarity between two clusters

- Single linkage tends to create bigger clusters than complete linkage.
- Centroid method is the most popular linkage in use
- Trial and fail to see which measure gives you the most desired clusters

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Summary

>Basic concepts in clustering analysis

>Similarity measurements

>Hierarchical clustering analysis