بسمه تعالی

Data Mining: Clustering

Dr O. Pournik *MD, MPH, MSc, PhD*

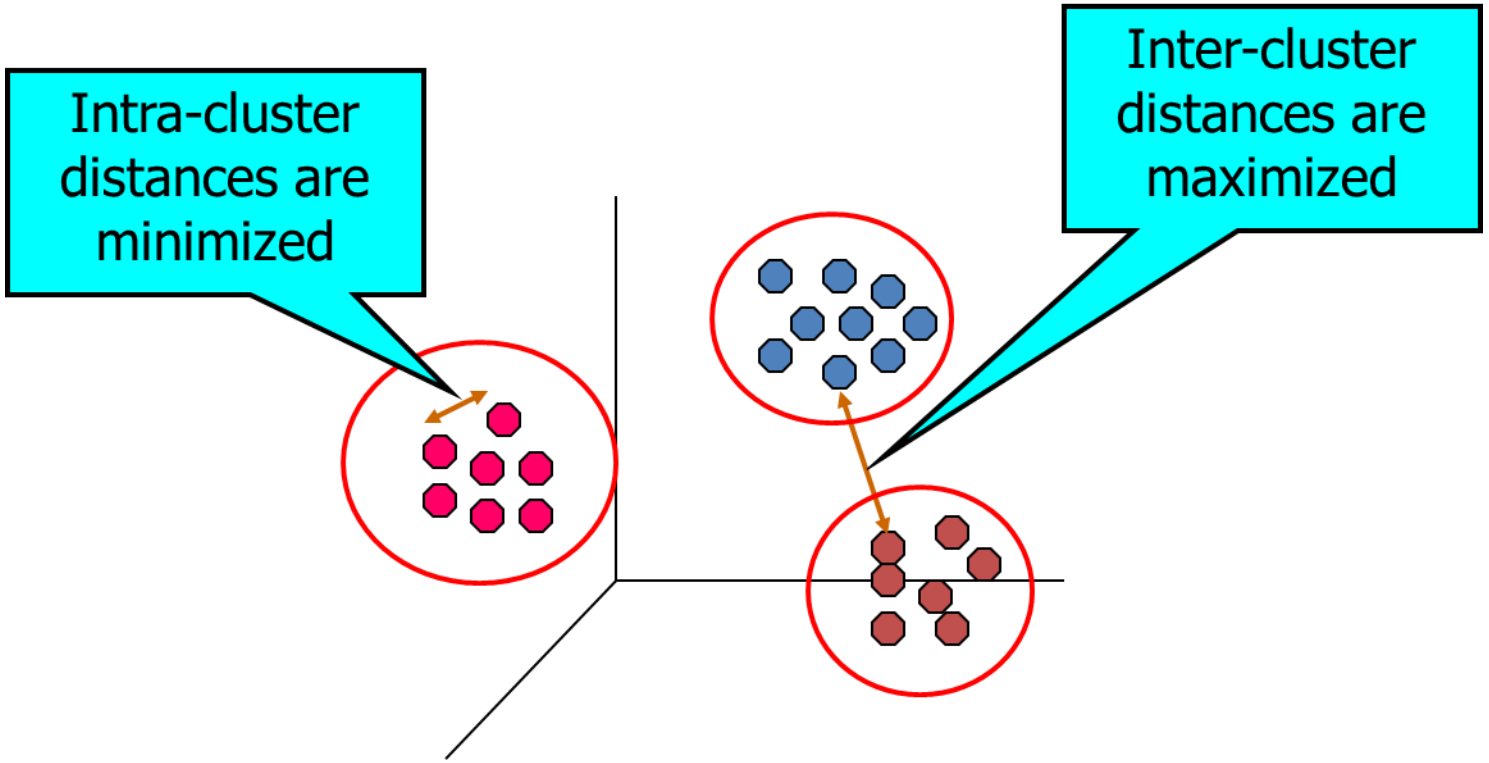
pournik@gmail.com

Clustering Definition

- ***Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that:***
  - ***Data points in one cluster are more similar to one another.***
  - ***Data points in separate clusters are less similar to one another.***
- ***Similarity Measures:***
  - ***Euclidean Distance if attributes are continuous.***
  - ***Other Problem-specific Measures.***

Clustering Definition

- Euclidean Distance Based Clustering in 3-D space.



Applications of Cluster Analysis

- ***Understanding***
  - ***Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations***

- ***Summarization***
  - ***Reduce the size of large data sets***

What is not Cluster Analysis?

- ***Supervised classification***
  - ***Have class label information.***
- ***Simple segmentation***
  - ***Dividing students into different registration groups alphabetically, by last name.***
- ***Results of a query***
  - ***Groupings are a result of an external specification.***
- ***Graph partitioning***
  - ***Some mutual relevance and synergy, but areas are not identical.***

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

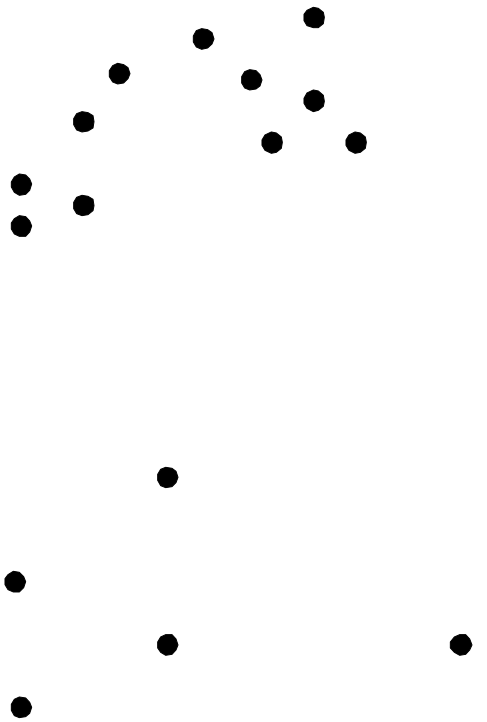
Four Clusters

Types of Clusterings

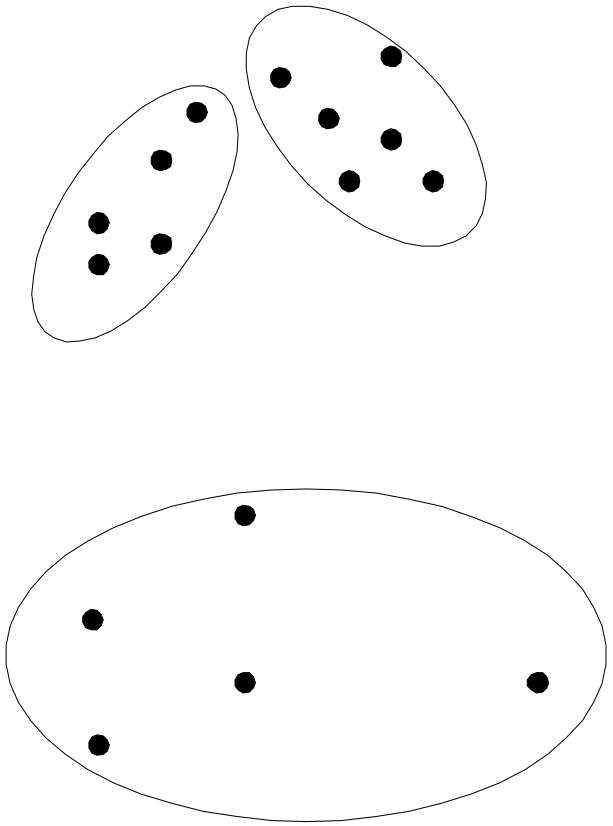
A clustering is a set of clusters:

- ***Important distinction between hierarchical and partitional sets of clusters.***
  - ***Partitional Clustering***
    - ***A division 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

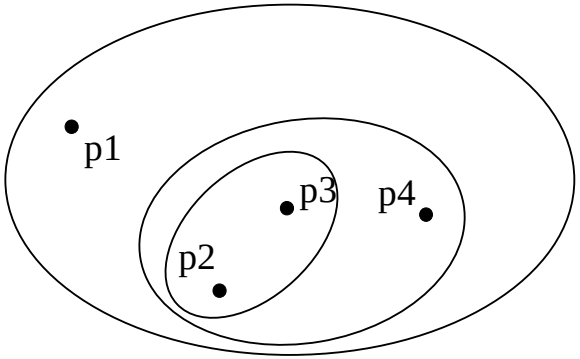


Original Points

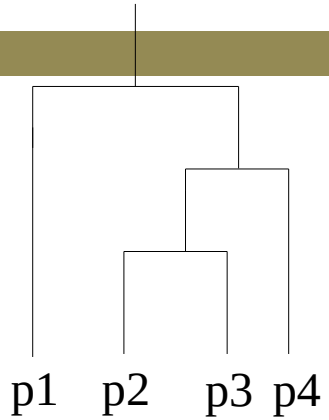


A Partitional Clustering

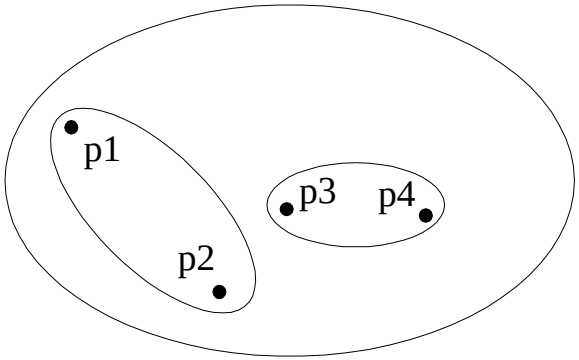
Hierarchical Clustering



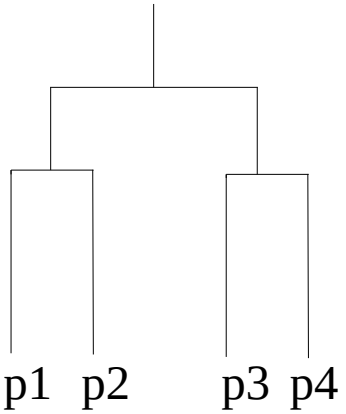
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Types of Clusters: Objective Function

- ***Clusters Defined by an Objective Function***
  - ***Finds clusters that minimize or maximize an objective function.***
  - ***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 algorithms typically have local objectives***
    - ***Partitional algorithms typically have global objectives***
  - ***A variation of the global objective function approach is to fit the data to a parameterized model.***
    - ***Parameters for the model are determined from the data.***
    - ***Mixture models assume that the data is a 'mixture' of a number of statistical distributions.***

Types of Clusters: Objective Function

- *Map the clustering problem to a different domain and solve a related problem in that domain*
  - *Proximity matrix defines a weighted graph, where the nodes are the points being clustered, and the weighted edges represent the proximities between points.*
  - *Clustering is equivalent to breaking the graph into connected components, one for each cluster.*
  - *Want to minimize the edge weight between clusters and maximize the edge weight within clusters.*

Clustering: Application 1

Market Segmentation:

- ***Goal:*** *subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.*
- ***Approach:***
  - *Collect different attributes of customers based on their geographical and lifestyle related information.*
  - *Find clusters of similar customers.*
  - *Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.*

Clustering: Application 2

Document Clustering:

- ***Goal:*** *To find groups of documents that are similar to each other based on the important terms appearing in them.*
- ***Approach:*** *To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.*
- ***Gain****: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.*

Illustrating Document Clustering

Clustering Points: 3204 Articles of Los Angeles Times.

Similarity Measure: How many words are common in these documents (after some word filtering).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metropolitan	943	746
Sports	738	573
Entertainment	354	278

Clustering: Application 3

- **Clustering of Stock Data**
  - ***Observe Stock Movements every day.***
  - ***Clustering points: Stock-{UP/DOWN}***
  - ***Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.***
  - ***We used clustering (News) with regards to association rules to quantify a similarity measure.***

Any Questions?