Data Mining and Warehousing

CPIT 440
15th Lecture
Dr. Reem Alotaibi

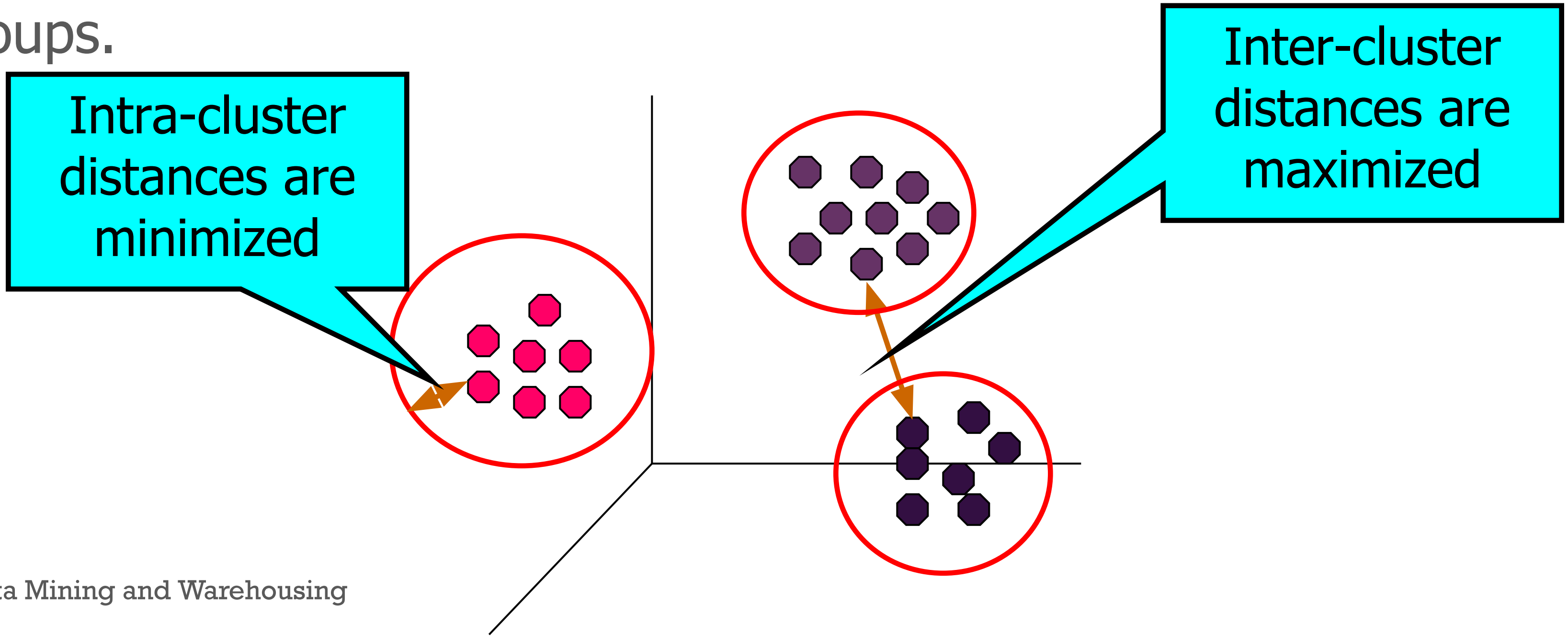
Outlines

- Cluster Analysis
- Basic Clustering Methods
  - Partitioning Methods
  - Hierarchical Methods
  - Density-Based Methods
  - Grid-Based Methods
- Evaluation of Clustering
- Summary

+ Cluster Analysis?

- Clustering is known as unsupervised learning because the class label information is not present.

- 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.



+ 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.

	Mean values	
	Age	Income
Cluster 1	35	$ 200,000
Cluster 2	50	$ 30,000
Cluster 3	23	$ 50,000
Cluster 4	36	$ 400,000

+ Applications of Clustering

- Image Recognition
  - Cluster images based on their visual content.
- Web
  - Cluster groups of users based on their access patterns on webpages.
  - Cluster webpages based on their content.
- Biology
  - Cluster similar proteins together (similarity wrt chemical structure and/or functionality etc).
- Many more…

+ 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.

+ Requirements for Cluster Analysis

- Scalability.
- Ability to deal with different types of attributes:
  - numeric, binary, nominal, etc.
- Discovery of clusters with arbitrary shape.
- Allow input parameters:
  - i.e. *k* number of cluster.
- Ability to deal with noisy data.

Requirements for Cluster Analysis

- Incremental clustering and insensitivity to input order.
- Capability of clustering high-dimensionality data.
- Constraint-based clustering.
- Interpretability and usability.

+ Considerations for Cluster Analysis

- **Partitioning criteria**
  - Single level vs. hierarchical partitioning.
- **Separation of clusters**
  - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class).

+ Considerations for Cluster Analysis

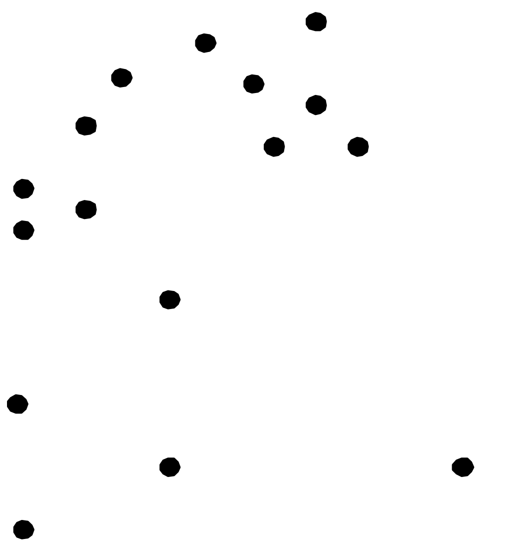
- **Similarity measure**
  - Distance-based (e.g., Euclidean distance).
- **Clustering space**
  - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering).

Outlines

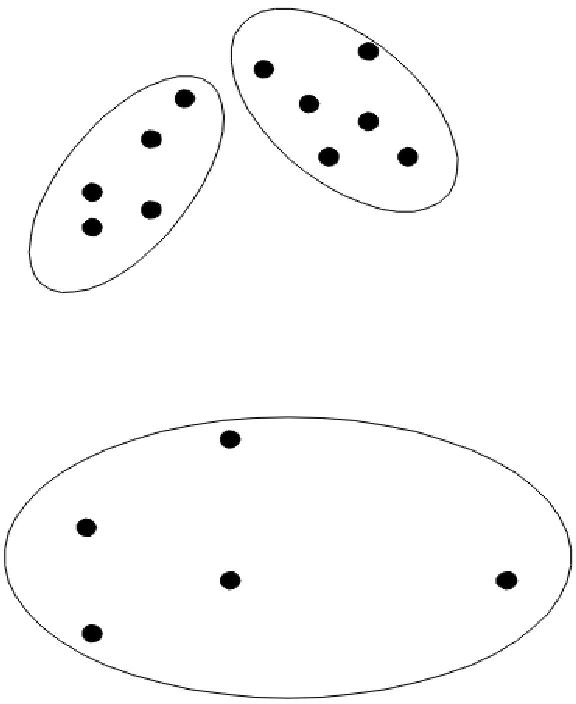
- Cluster Analysis
- Basic Clustering Methods
  - Partitioning Methods
  - Hierarchical Methods
  - Density-Based Methods
  - Grid-Based Methods
- Evaluation of Clustering
- Summary

+ Basic Clustering Methods

- Partitioning Methods: A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.
  - *K-means and k-medoids*.



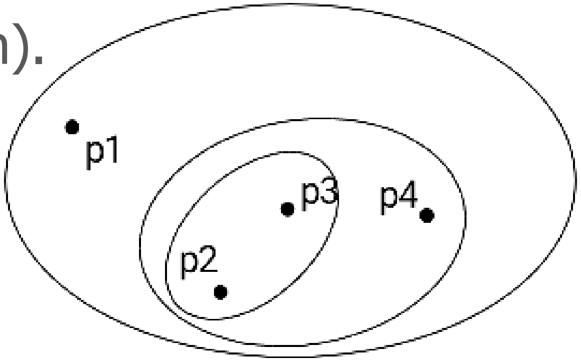
Original Points



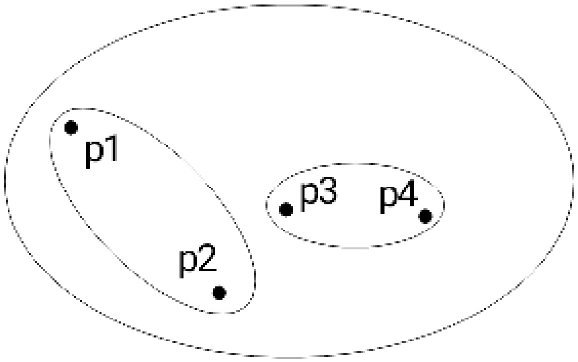
A Partitional Clustering

+ Basic Clustering Methods

- Hierarchical Methods: A set of nested clusters organized as a hierarchical tree.
  - *Agglomerative* (bottom-up approach).
  - *Divisive* (top-down approach).



Traditional Hierarchical Clustering



Non-traditional Hierarchical Clustering

+ Basic Clustering Methods

- Density-based methods: A set of objects divided into multiple exclusive clusters, or a hierarchy of clusters based on the notion of *density*.

+ Basic Clustering Methods

- Grid-based methods: They quantize the object space into a finite number of cells that form a grid structure.

Basic Clustering Methods

Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

+ K-Means

- K-means algorithm is the most well-known and commonly used partitioning methods.

- The basic algorithm is very simple:
  1. Initial set of clusters randomly chosen (called *centroids*).
  2. Each point is assigned to the cluster with the closest centroid.
  3. The *cluster mean* is the mean value of all points within the cluster.
  4. Iteratively, items are moved among sets of clusters until the desired set is reached.
  5. The iterations continue until the assignment is stable.

+ K-Means Algorithm

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- k: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects from D as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

+ K-Means Example

- **Given:**
  - Data={2,4,10,12,3,20,30,11,25}
  - Let k=2
- **Algorithm:**
  1. Randomly assign means: $m_1=3, m_2=4$
  2. $K_1=\{2,3\}$, $K_2=\{4,10,12,20,30,11,25\}$, $m_1=2.5, m_2=16$
  3. $K_1=\{2,3,4\}, K_2=\{10,12,20,30,11,25\}$, $m_1=3, m_2=18$
  4. $K_1=\{2,3,4,10\}, K_2=\{12,20,30,11,25\}$, $m_1=4.75, m_2=19.6$
  5. $K_1=\{2,3,4,10,11,12\}, K_2=\{20,30,25\}$, $m_1=7, m_2=25$
  6. Stop as the clusters with these means are the same.

+ Evaluating K-means Clusters

- Most common measure is Sum of Squared Error (SSE):
  - For each point, the error is the distance to the nearest cluster.
  - To get SSE, we square these errors and sum them:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - ***x*** is a data point in cluster C_i and $\boldsymbol{m_i}$ is the representative point for cluster C_i.

Discussion on K-means Algorithm

- Need to specify K.
- Finds a local optimum.
- Converges often quickly (but not always)
- The choice of initial points can have large influence
  - Clusters of different densities
  - Clusters of different sizes
- Outliers can also cause a problem ☐ k-medoids

+ K-Medoids

- It is called Partitioning Around Medoids (PAM)
- Handles outliers well.
- Ordering of input does not impact results.
- The algorithm is similar to k-means:
  - Initial set of k medoids randomly chosen.
  - Each cluster represented by one item, called the medoid.

K-medoids Algorithm

Algorithm: ***k*-medoids.** PAM, a *k*-medoids algorithm for partitioning based on medoid or central objects.

Input:

- k: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects in D as the initial representative objects or seeds;
(2) **repeat**
(3) assign each remaining object to the cluster with the nearest representative object;
(4) randomly select a nonrepresentative object, $\boldsymbol{o}_{random}$;
(5) compute the total cost, S, of swapping representative object, $\boldsymbol{o}_j$, with $\boldsymbol{o}_{random}$;
(6) **if** $S < 0$ **then** swap $\boldsymbol{o}_j$ with $\boldsymbol{o}_{random}$ to form the new set of k representative objects;
(7) **until** no change;

Summary

- K-means and k-medoids are very similar algorithms.
- K-medoids uses the medoids while k-means uses the centroids.

Medoids	Centroids
• Medoid is the median of the data points within the clusters. • Medoids are actual data points. • Robust to outliers.	• Centroid is the mean of the data points within the clusters. • Centroids are not actual data points. • Sensitive to outliers.