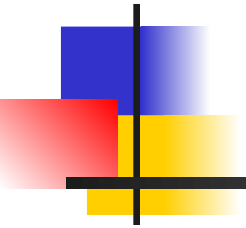
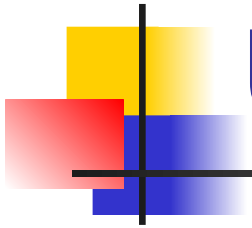


Cluster Analysis

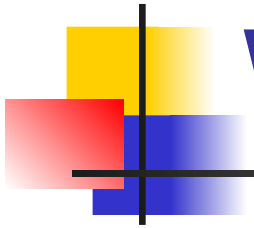
Dr. Bernard Chen

Troy University

Supervised learning VS Unsupervised learning



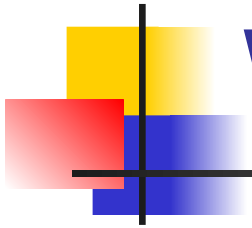
- If the class label of each training tuple is provided, this process is known as **supervised learning**
- It contrasts with **unsupervised learning** (or clustering), in which the class label of each training tuple is unknown



What is Cluster Analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters

- Cluster analysis
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters



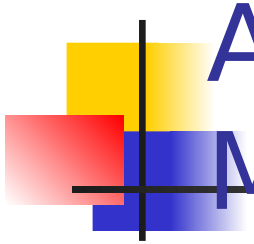
What is Cluster Analysis?

- Clustering analysis is an important human activity
- Early in childhood, we learn how to distinguish between cats and dogs

- Unsupervised learning: no predefined classes

- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

Clustering: Rich Applications and Multidisciplinary Efforts



- Pattern Recognition
- Spatial Data Analysis
  - Create thematic maps in GIS by clustering feature spaces
  - Detect spatial clusters or for other spatial mining tasks
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

Quality: 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)

- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns

Major Clustering Approaches

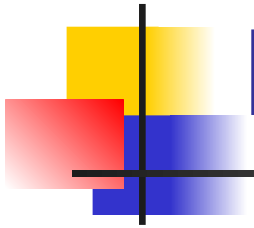
- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical methods: k-means, k-medoids, CLARANS
- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Hierarchical, Diana, Agnes, BIRCH, ROCK, CAMELEON
- Density-based approach:
  - Based on connectivity and density functions
  - Typical methods: DBSACN, OPTICS, DenClue

Partitioning Algorithms: Basic Concept

- Partitioning method: Construct a partition of a database **D** of **n** objects into a set of **k** clusters, s.t., min sum of squared distance

$$\sum_{m=1}^{k} \sum_{t_{mi} \in Km} (C_m - t_{mi})^2$$

Partitioning Algorithms: Basic Concept

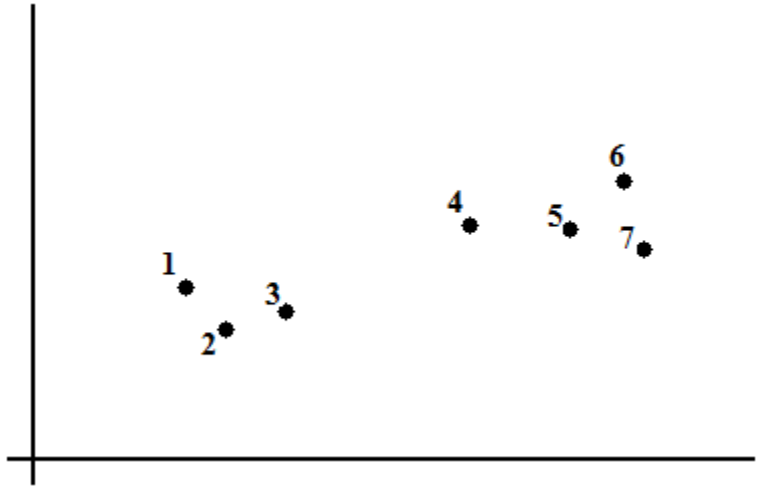


- Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: *k-means* and *k-medoids* algorithms
    - *k-means* (MacQueen'67): Each cluster is represented by the center of the cluster
    - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

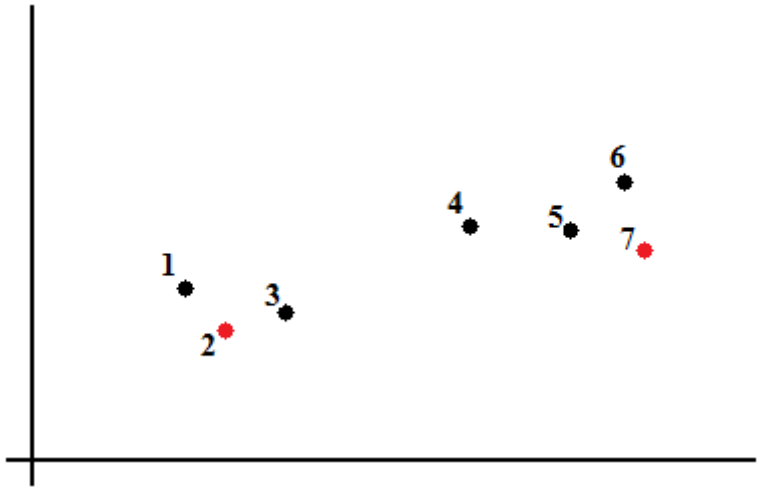
The *K-Means* Clustering Method

- Given *k*, the *k-means* algorithm is implemented in four steps:
  1. Partition objects into *k* nonempty subsets
  2. Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., *mean point*, of the cluster)
  3. Assign each object to the cluster with the nearest seed point
  4. Go back to Step 2, stop when no more new assignment

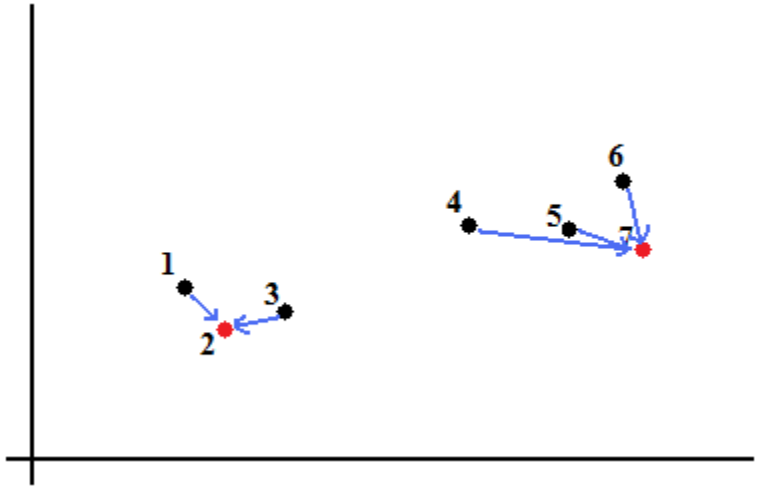
K-means Clustering



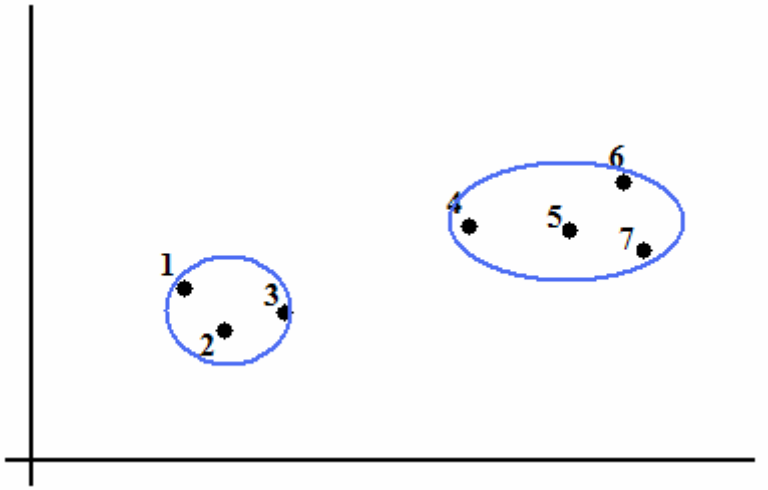
K-means Clustering



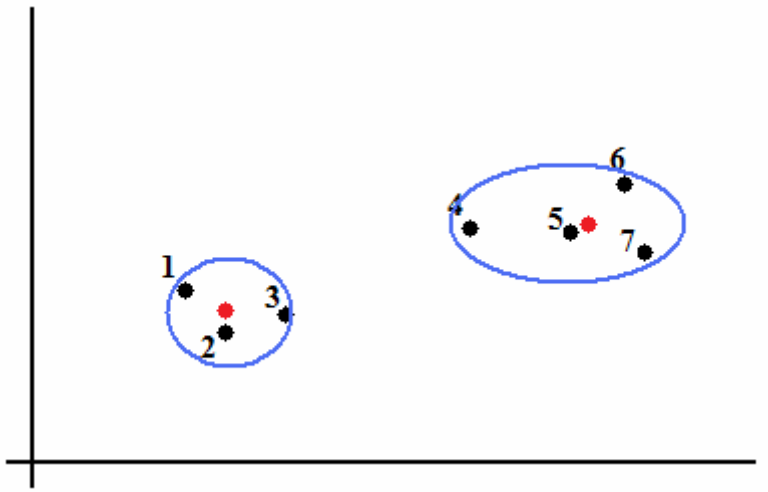
K-means Clustering



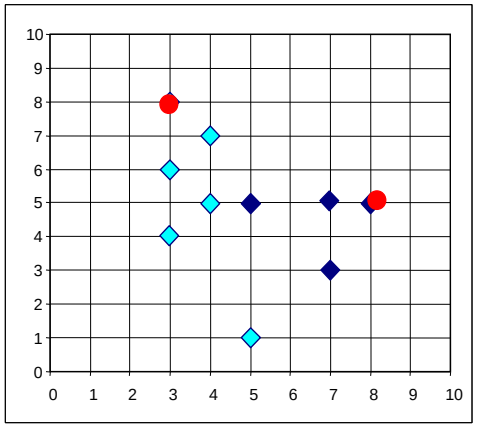
K-means Clustering



K-means Clustering



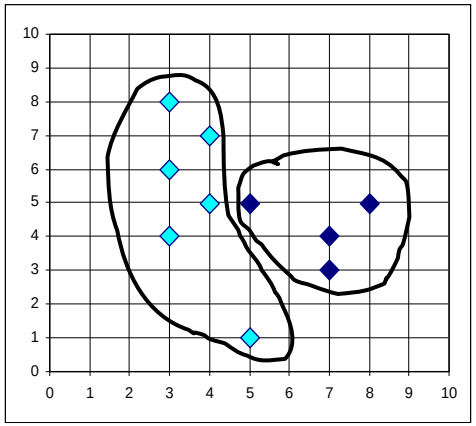
The *K-Means* Clustering Method



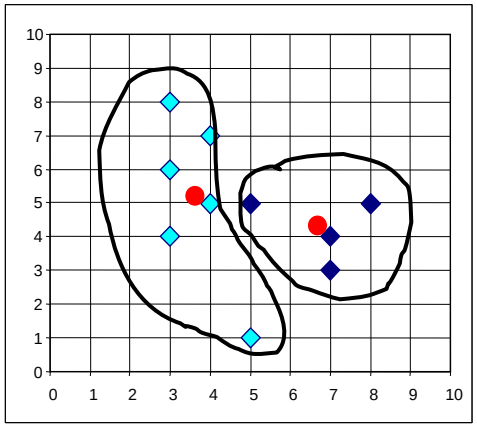
K=2

Arbitrarily choose K object as initial cluster center

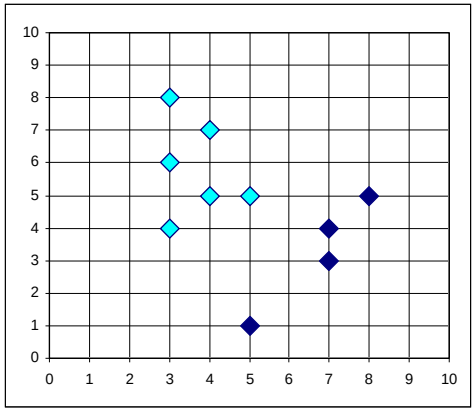
Assign each objects to most similar center



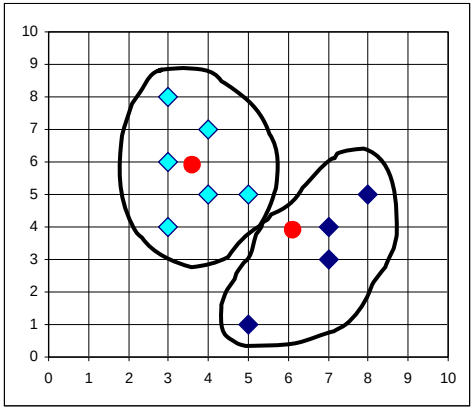
Update the cluster means



reassign



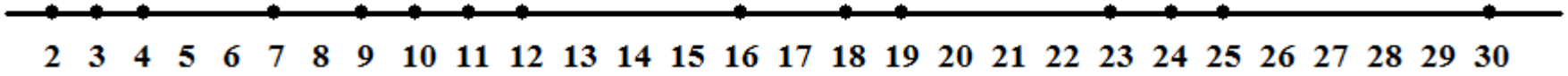
Update the cluster means



reassign

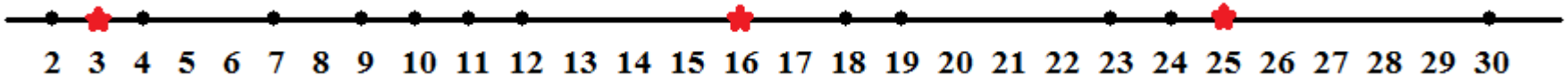
Example

- Run K-means clustering with 3 clusters (initial centroids: 3, 16, 25) for at least 2 iterations



Example

- Run K-means clustering with 3 clusters (initial centroids: 3, 16, 25) for at least 2 iterations



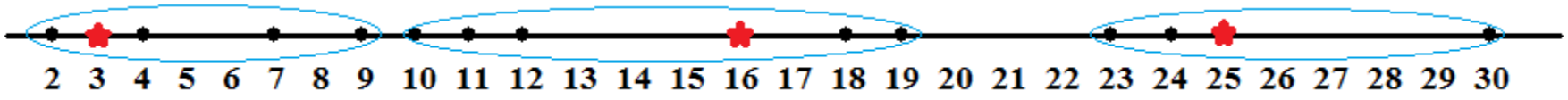
Example

- Centroids:

3 – 2 3 4 7 9

16 – 10 11 12 16 18 19

25 – 23 24 25 30



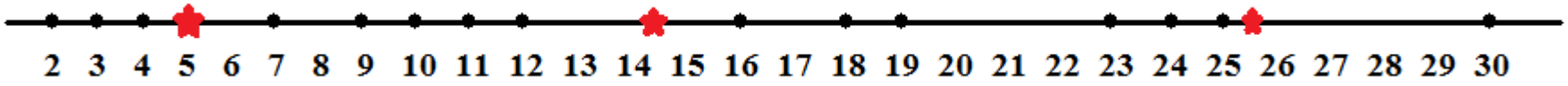
Example

- Centroids:

3 – 2 3 4 7 9 new centroid: 5

16 – 10 11 12 16 18 19 new centroid: 14.33

25 – 23 24 25 30 new centroid: 25.5



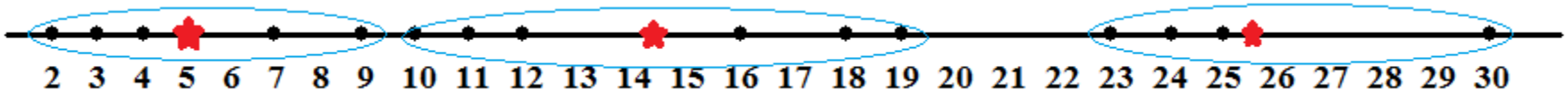
Example

- Centroids:

5 – 2 3 4 7 9

14.33 – 10 11 12 16 18 19

25.5 – 23 24 25 30



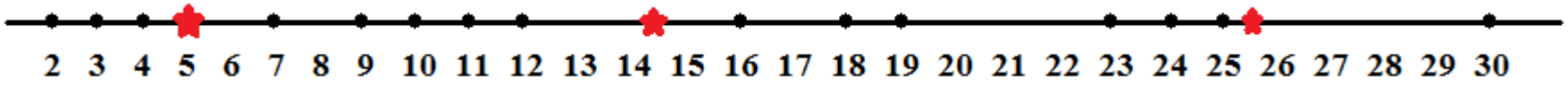
Example

- Centroids:

5 – 2 3 4 7 9 new centroid: 5

14.33 – 10 11 12 16 18 19 new centroid: 14.33

25.5 – 23 24 25 30 new centroid: 25.5



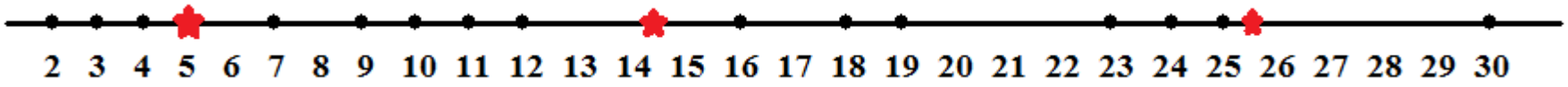
Example

- Centroids:

5 – 2 3 4 7 9 new centroid: 5

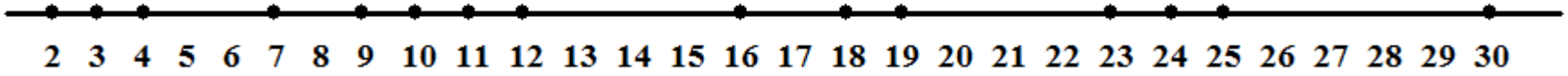
14.33 – 10 11 12 16 18 19 new centroid: 14.33

25.5 – 23 24 25 30 new centroid: 25.5

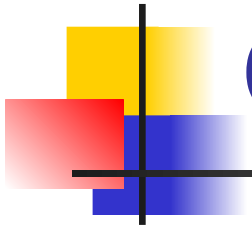


Class Assignment

- Run K-means clustering with 3 clusters (initial centroids: 3, 12, 19) for at least 2 iterations



Look deeper into Clustering



- Three more things can be discussed in clustering:
  - Distance measure
  - Distance between clusters
  - Centroid location

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + ... + |x_{i_p} - x_{j_p}|^q)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}|$$

Similarity and Dissimilarity Between Objects (Cont.)

- *If q = 2, d* is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2)}$$

- Also, one can use weighted distance, parametric Pearson correlation, or other disimilarity measures