Kmeans Clustering

Nazerfard, Ehsan
nazerfard@aut.ac.ir

Figure © towardsdatascience.com/

Cluster Analysis

- Cluster analysis or clustering: 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.

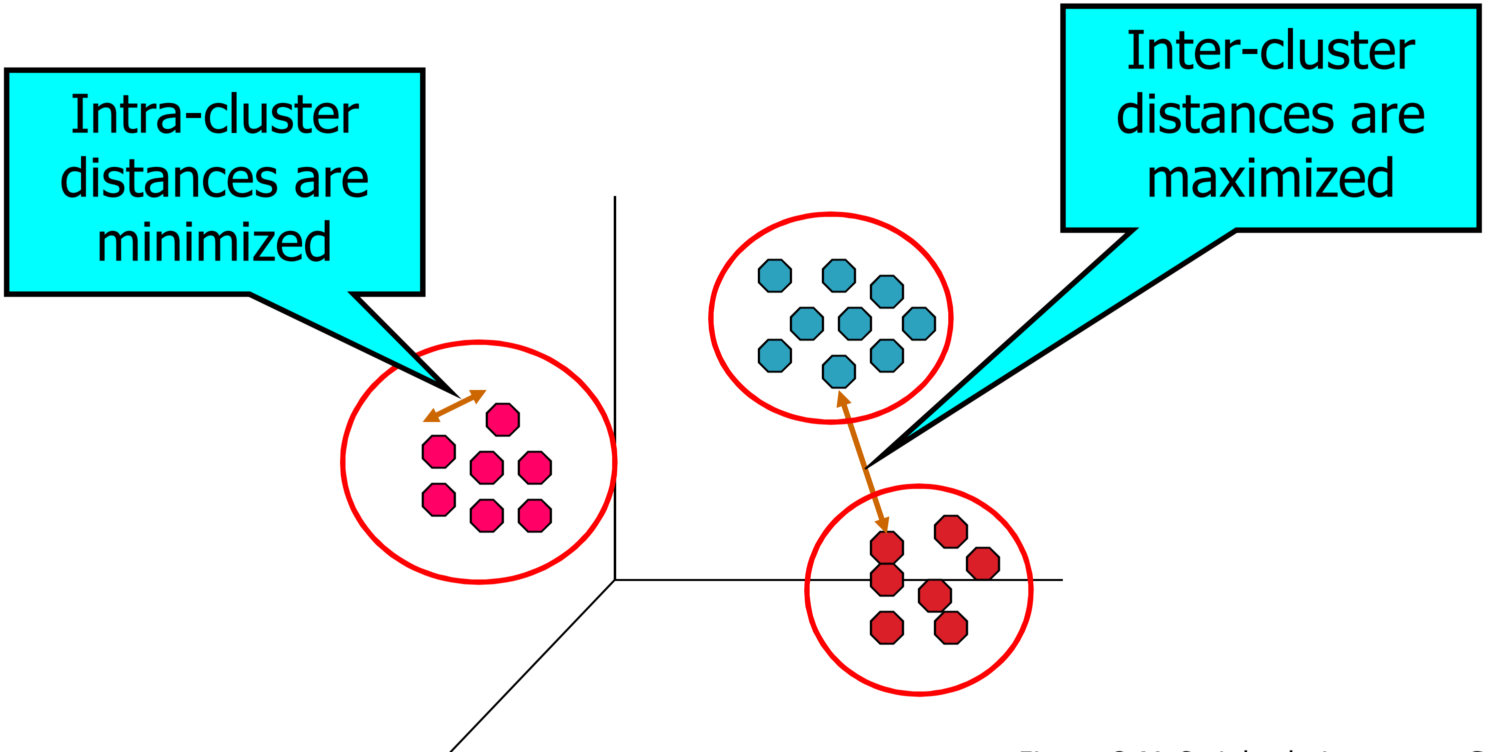


Figure © M. Steinbach, Lecture on Clustering

Applications

- Document clustering (news, …)
- Community Detection in Social Networks
- Sentiment analysis (customer reviews, …)
- Gene expression clustering
- Clustering of patients based on phenotypic and genotypic factors for efficient disease diagnosis
- Market Segmentation
- Anomaly detection
- Fraud detection
- Finding groups of driver behaviors based upon patterns of automobile motions (normal, drunken, sleepy, rush hour driving, etc.)
- …

Major Clustering Approaches

- Partitioning-based 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, CLARA, CLARANS
- Density-based approach
  - Based on connectivity and density functions
  - Typical methods: DBSCAN, OPTICS, DenClue
- Hierarchical approach
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Agnes, Diana, BIRCH, CURE, CHAMELEON
- Model-based approach
  - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
  - Typical methods: EM, SOM

Major Clustering Approaches (cont.)

- Grid-based approach
  - Based on a multiple-level granularity structure
  - Typical methods: STING, CLIQUE, WaveCluster
- Frequent Pattern-based approach
  - Based on the analysis of frequent patterns
  - Typical methods: p-Cluster
- Support Vector approach
  - Based on the idea of mapping data points into higher dimensional feature space via a kernel function.
  - Typical methods: SVC, Kernel K-means
- Graph Theoretic approach
  - Typical methods: Spectral Clustering
- …

Partitioning-based Approach

- Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors.

- Example: K-means

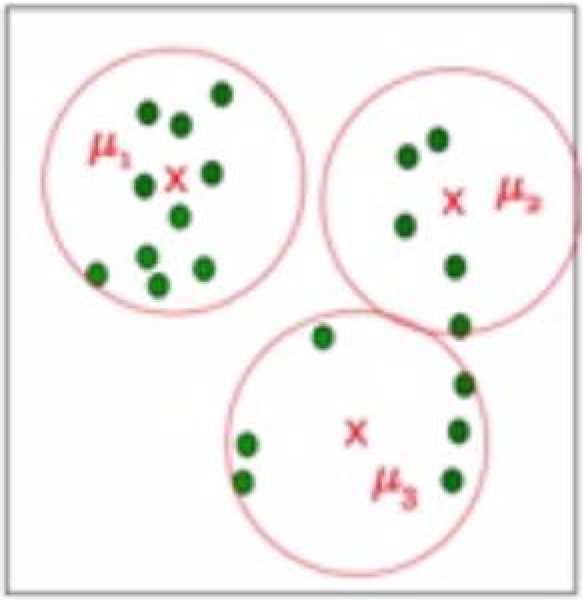
K-means Clustering

- Assume K clusters
- Iterate between two following steps:
  - Updating the assignment of data to clusters
  - Updating the cluster's summarization

  - sklearn.cluster.KMeans

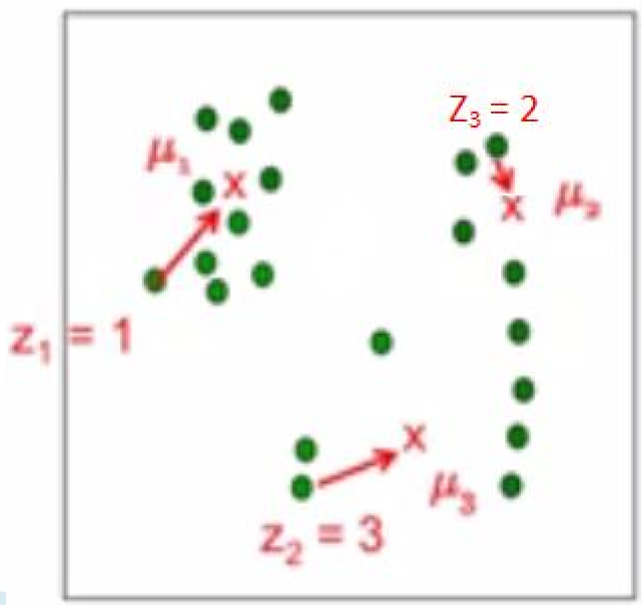
K-means Clustering (cont.)

- Assume K clusters
- Iterate between two following steps
  A. Updating the assignment of data to clusters
  B. Updating the cluster's summarization
- Each cluster C is described by a centroid $\boldsymbol{\mu_c}$



K-means Clustering (cont.)

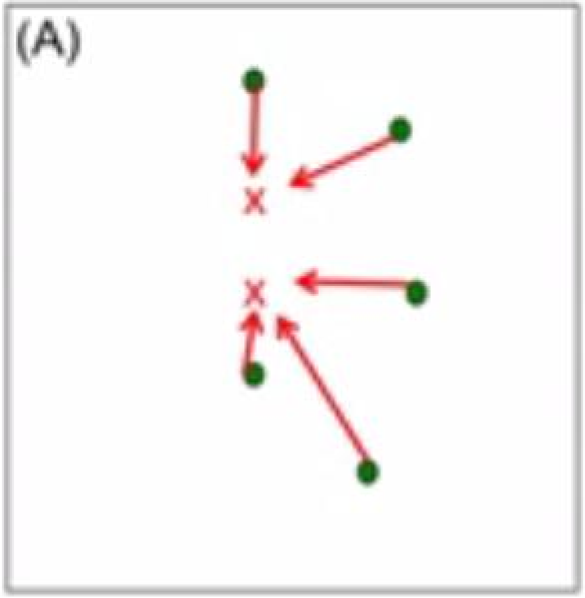
- Assume K clusters
- Iterate between two following steps:
  A. Updating the assignment of data to clusters
  B. Updating the cluster's summarization
- Assignment of i_{th} example: $z_i \in 1..K$



K-means Clustering (cont.)

- Iterate until convergence
  A. For each data, find the closest centroid:

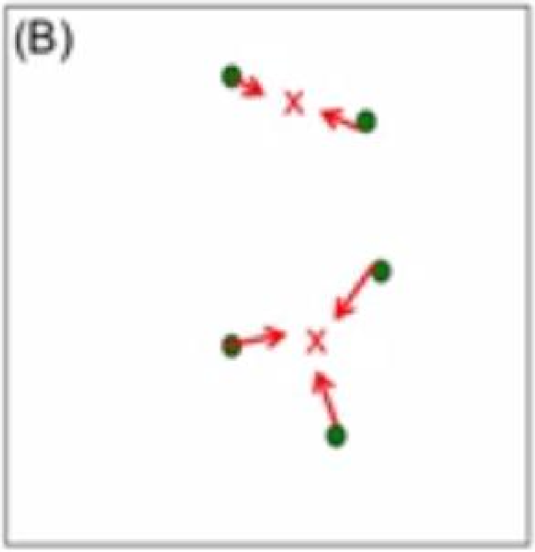
$$z_i = \underset{c}{\operatorname{argmin}} \left|\left|\boldsymbol{x_i} - \boldsymbol{\mu_c}\right|\right|^2, \forall i$$



K-means Clustering (cont.)

- Iterate until convergence
  - B. Set each cluster to the mean of all assigned data:

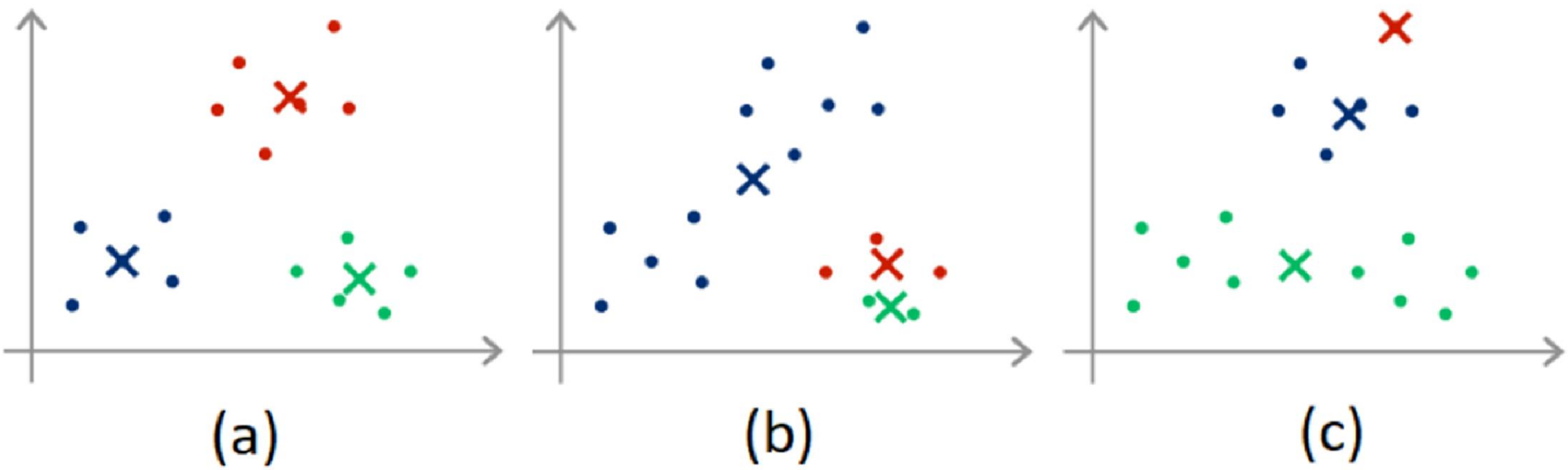
$$\forall c, \qquad \boldsymbol{\mu_c} = 1/\mathrm{m_c} \sum_{i \in S_c} x_i \qquad \mathrm{S_c} = \{\mathrm{i}: \mathrm{z_i} = \mathrm{c}\}, \mathrm{m_c} = |\mathrm{S_c}|$$



Demo

K-means Properties

- Poor initialization may lead to poor clustering



- Solution?
  - Multiple Initializations → randomness
  - K-means++, Intelligent K-means

K-means Properties (cont.)

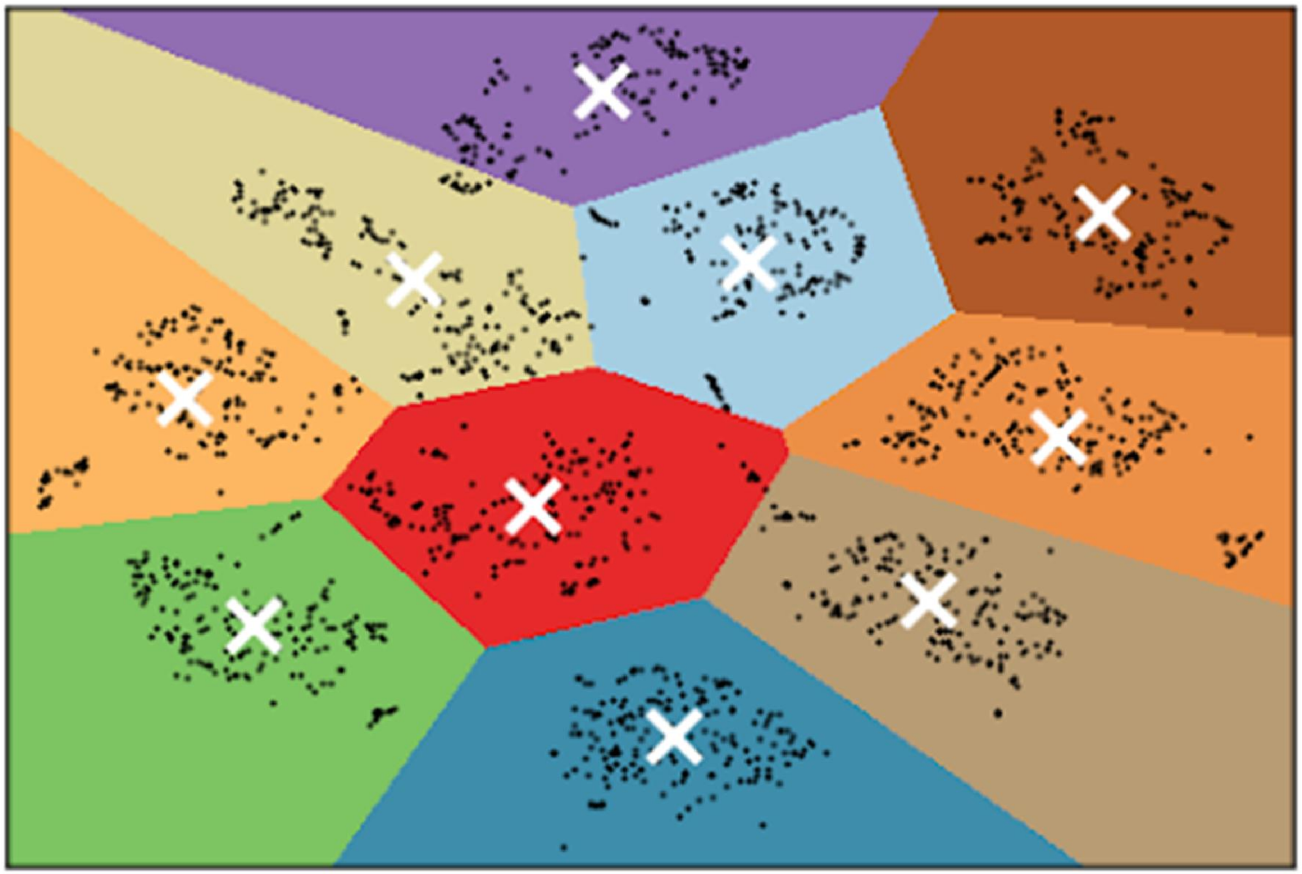
- Distance metrics
  - l_1 norm (Manhattan distance)
  - l_2 norm (Euclidean distance)
  - Cosine distance

- Centroids
  - Mean
  - Median → Sensitivity to outliers?
  - Medoid
    - Most commonly used on data when a mean or centroid cannot be defined, such as graphs.
  - …

K-means Properties (cont.)

- Instance-based
- Time complexity: $O(tkm)$
- Non-parametric
- Linearly separable data

K-means: Linear Separable



Sum of Square Error

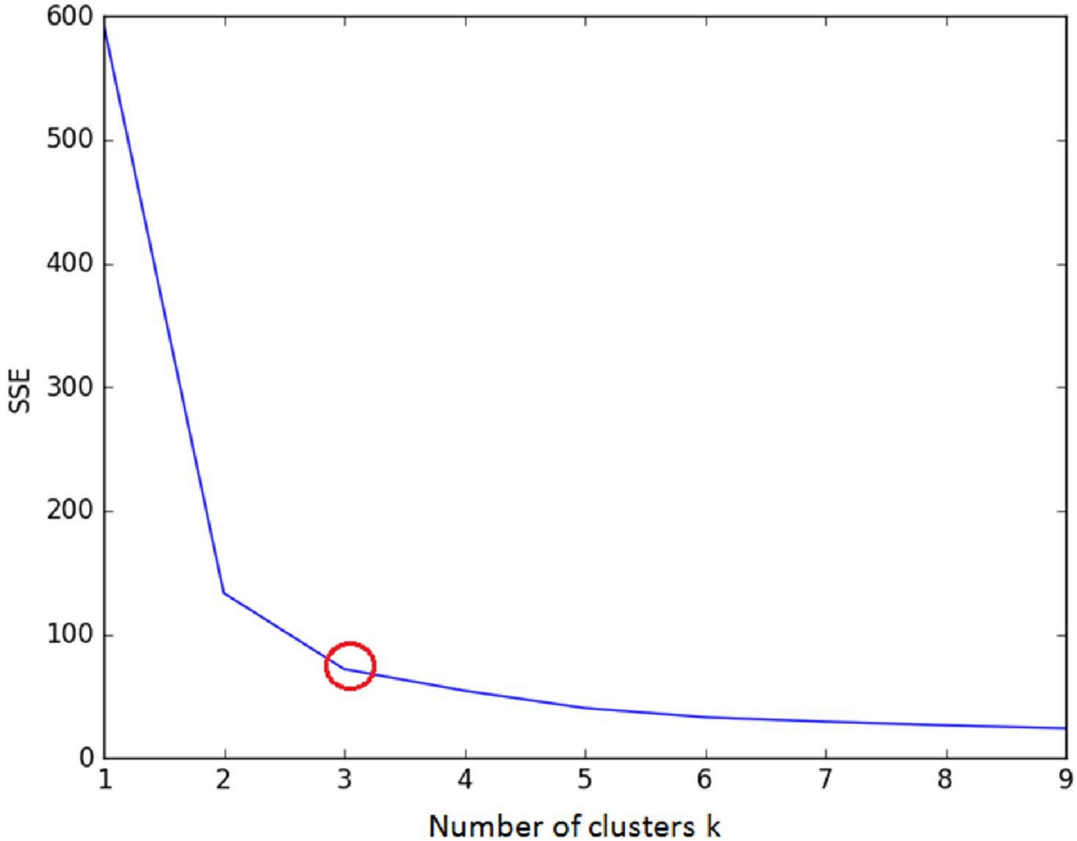
- Sum of Square Error (SSE)

$$SSE = \sum_k \sum_{\boldsymbol{x_i} \in C_k} ||\boldsymbol{x_i} - C_k||^2$$

  - Goal: minimizing within-cluster distance

Optimal number of Clusters

- Elbow method



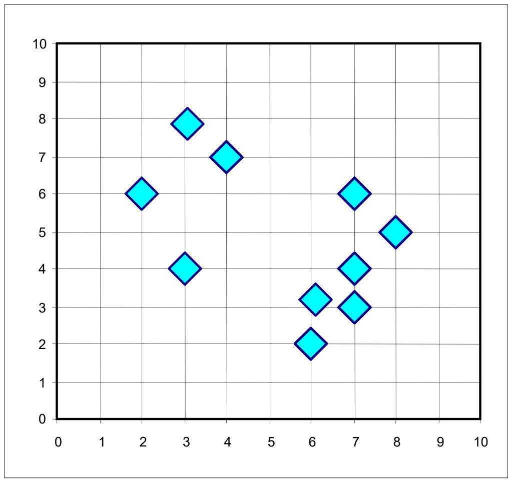
K-means Variations

- K-medoids/PAM (Partitioning Around Medoids)

- CLARA (Clustering Large Applications)

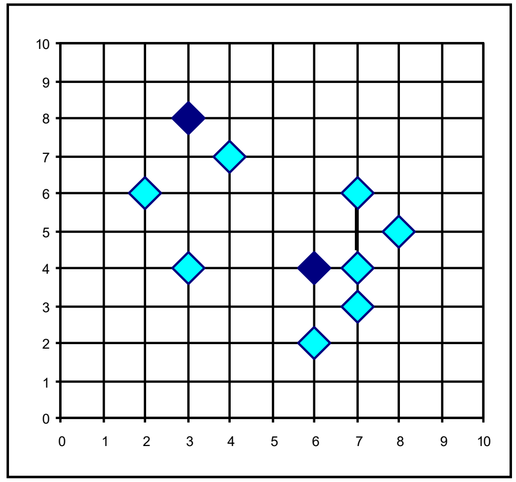
- CLARANS (A Clustering Algorithm based on Randomized Search)

PAM Algorithm



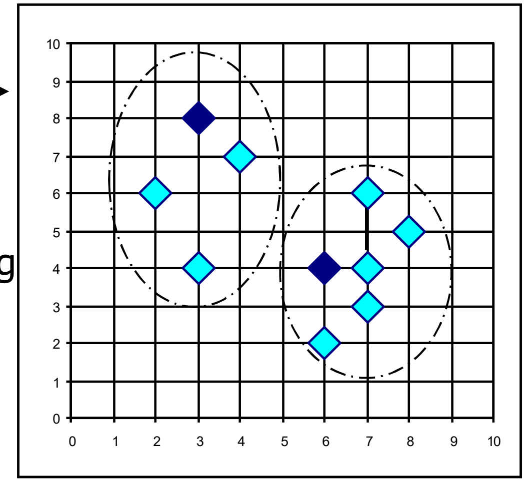
K=2

Arbitrary choose k object as initial medoids

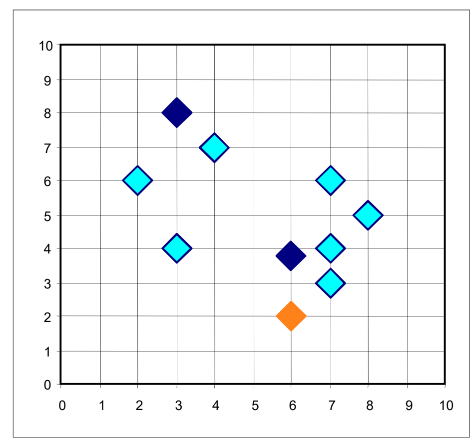


Assign each remaining object to nearest medoids

Total Cost = 20

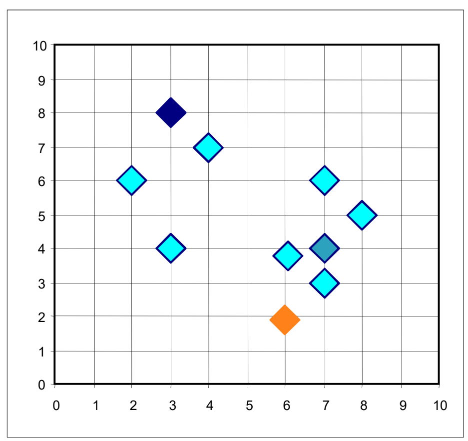


Randomly select a nonmedoid object,O_{ramdom}



Compute total cost of swapping

Total Cost = 26



Swapping O and O_{ramdom}

If quality is improved.

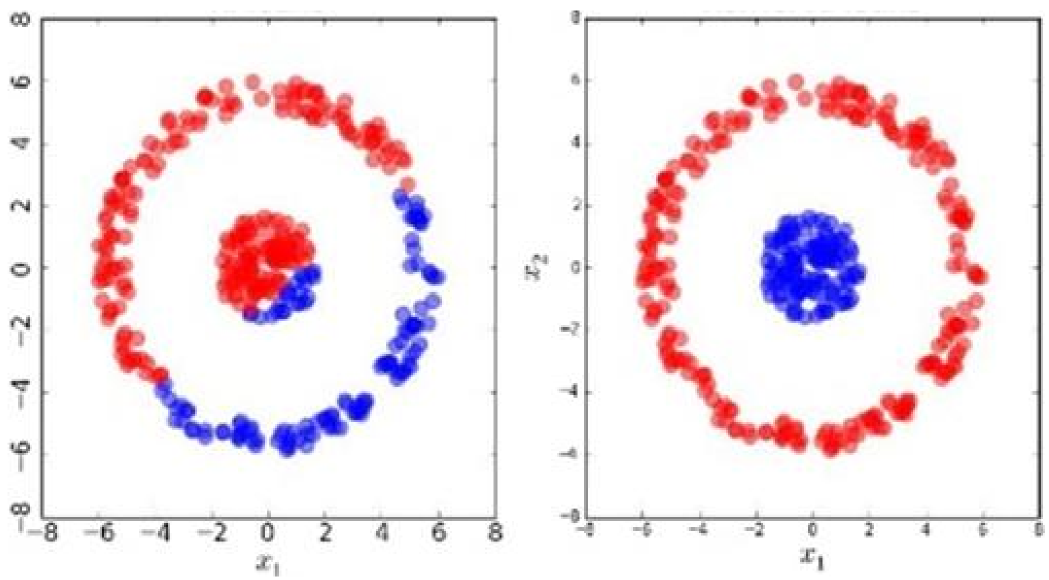
Do loop

Until no change

K-means Variations (cont.)

- Fuzzy C-means

- Kernel K-means



Further Reading

- Clustering Categorical Data
  - ROCK (robust clustering algorithm for categorical attributes)
  - Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, ICDE'99

- Mean Shift Clustering

References

- Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques, 3rd edition, 2006.