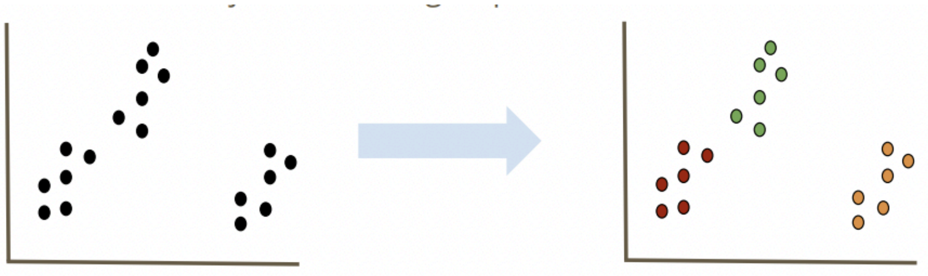
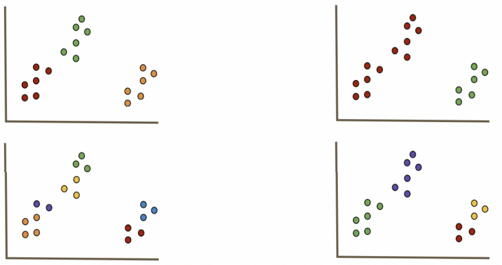
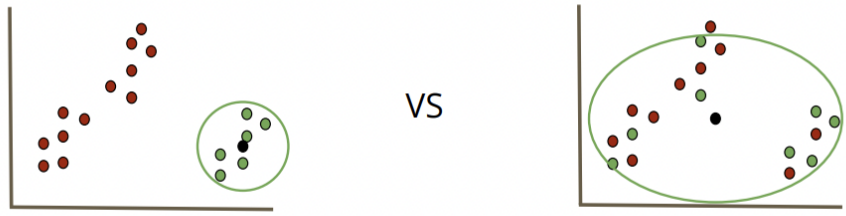
CAS CS 506
Lec 05

Clustering - Kmeans

1. What is Clustering
    a. A clustering is a grouping / assignment of objects (data points) such that objects in the same group / cluster are
        i. similar to one another
        ii. dissimilar to objects in other groups
        iii. 
2. Applications
    a. Outlier detection / anomaly detection
        i. Data Cleaning / Processing
        ii. Credit card fraud, spam filter, etc.
    b. Feature Extraction
    c. Filling gaps in the data
        i. Using the same marketing strategy for similar people
        ii. Infer probable values for gaps in the data (similar users could have similar hobbies, likes / dislikes, etc.)
3. Clusters can be Ambiguous
    a. 
4. Types of Clusterings
    a. Partitional
        i. Each object belongs to exactly one cluster
    b. Hierarchical
        i. A set of nested clusters organized in a tree
    c. Density-Based
        i. Defined based on the local density of points
    d. Soft Clustering
        i. Each point is assigned to every cluster with a certain probability

5. Partitional Clustering
    a. Goal: partition dataset into k partitions
    b. 

6. Example
    a. Given a distance function d, we can find points (not necessarily part of our dataset) for each cluster called centroids that are at the center of each cluster.
    b. 

    c. Q: When d is Euclidean, what is the centroid (also called center of mass) of m points $\{x_1, \ldots, x_m\}$?
    d. Looking at the sum of the distances of points in a cluster to its centroid also captures the "spread" (variance) of a cluster

ance) of a cluster

Mean of cluster i

$$\sum_{i}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

Cluster i

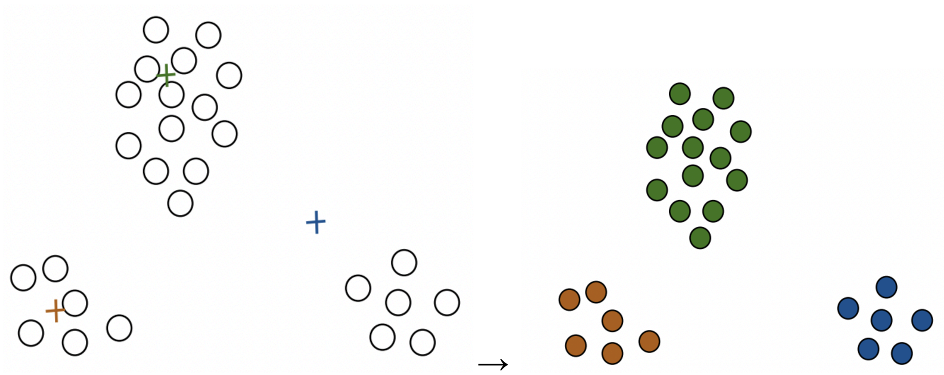
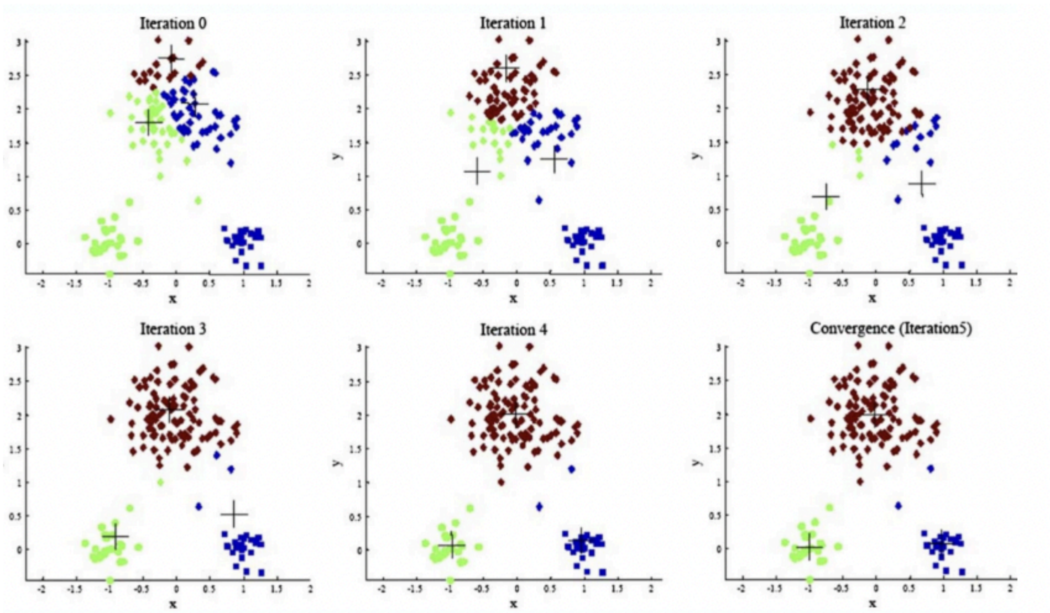
7. Cost Function
    a. Way to evaluate and compare solutions
    b. Hope: can find some algorithm that find solutions that make the cost small
    c. Q: Can you suggest a cost function to use for partitional clustering?

$$\sum_{i}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

8. K-means
    a. Given $X = \{x_1, \ldots, x_n\}$ our dataset and k
    b. Find k points $\{u_1, \ldots, u_k\}$ that minimize the cost function:

$$\sum_{i}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

    c. When k = 1 and k = n this is easy. Why?
        i. If k = n, every data point is a cluster
        ii. If k = 1, the whole data point is a cluster
    d. When $x_i$ lives in more than 2 dimensions, this is a very difficult (NP-hard) problem

9. K-means - Lloyd's Algorithm
    a. Randomly pick k centers $\{u_1, \ldots, u_k\}$
    b. Assign each point in the dataset to its closest center
    c. Compute the new centers as the means of each cluster
    d. Repeat 2 & 3 until convergence
    e. 
    f. 


10.