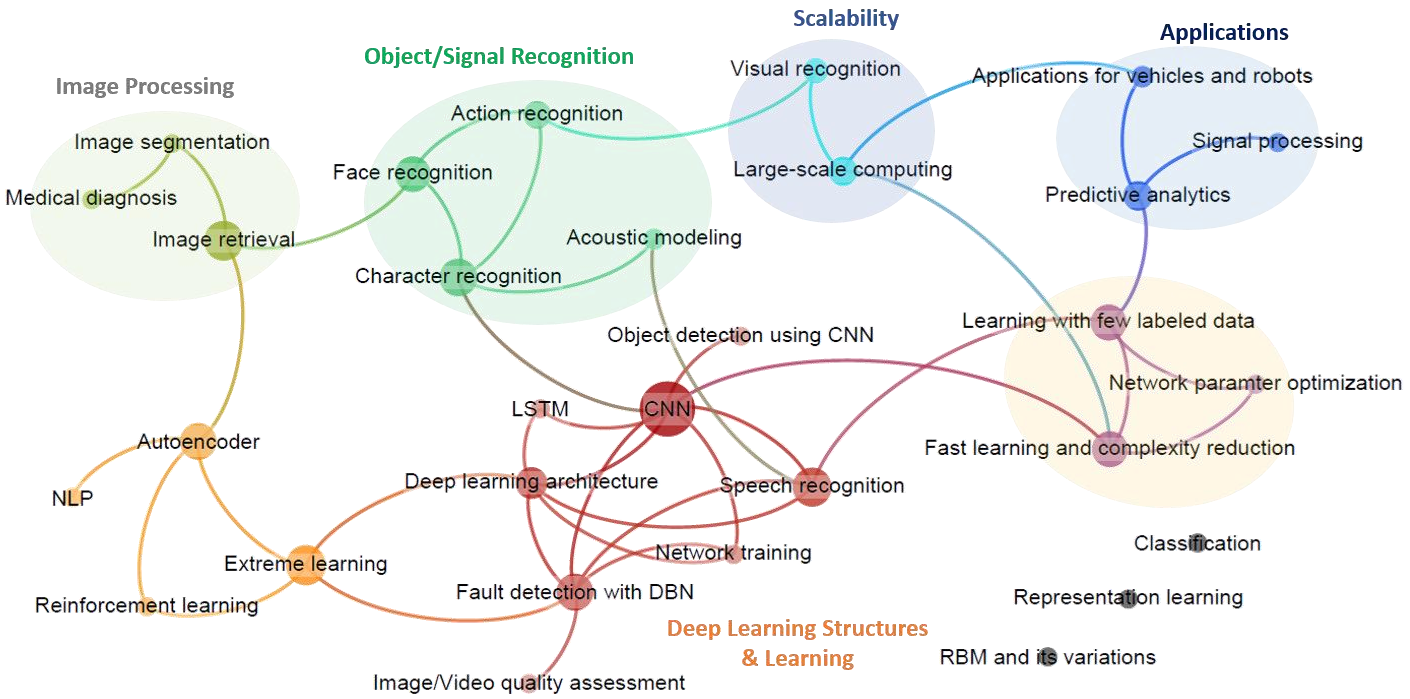


Clustering: K-Means Clustering

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

01 Clustering: Overview

02 K-Means Clustering

03 Hierarchical Clustering

04 Density-based Clustering: DBSCAN

04 R Exercise

K-Means Clustering

- K-Means Clustering (KMC)
  - ✓ Partitional clustering approach
    - Each cluster is associated with a centroid
    - Each point is assigned to the cluster with the closest centroid
    - Number of cluster, K, must be specified

$$\mathbf{X} = C_1 \cup C_2 \ldots \cup C_K, \quad C_i \cap C_j = \phi, \quad i \neq j$$

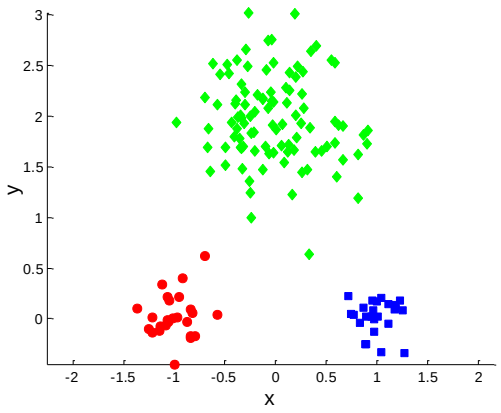
$$\arg\min_{\mathbf{C}} \sum_{i=1}^{K} \sum_{\mathbf{x}_j \in C_i} ||\mathbf{x}_j - \mathbf{c}_i||^2$$

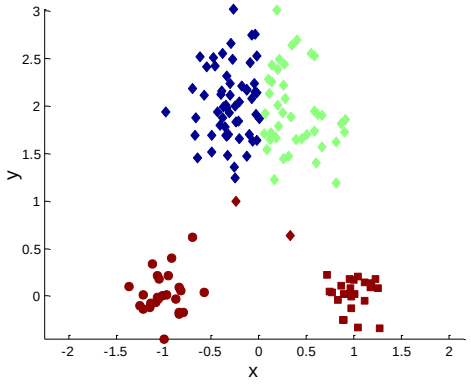
K-Means Clustering

- K-Means Clustering Procedure

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

✓ Initial centroids are often chosen randomly: clustering results vary according to the initial centroid selection





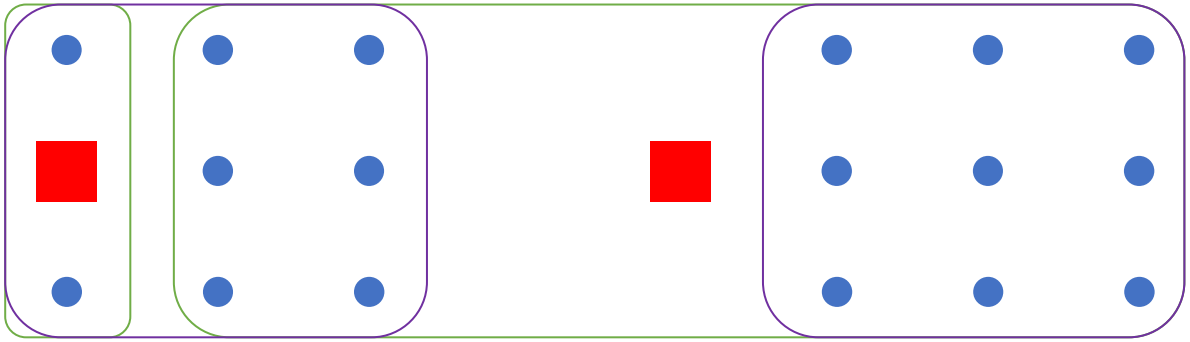
K-Means Clustering

- Example
  - ✓ Step 1: Initializing K centroids

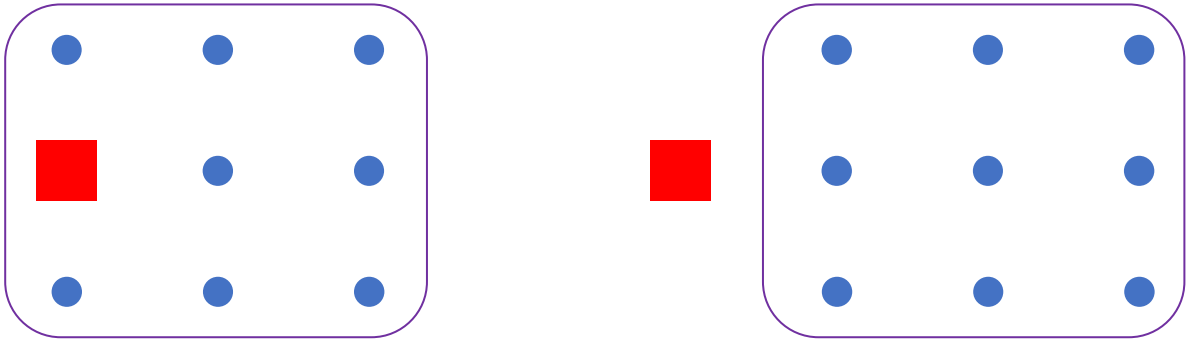
  - ✓ Step 2-1 (1st): Assign each instance to the closest center
  - ✓ Step 2-2 (1st): Re-compute the centroids based on the assigned instances

K-Means Clustering

- Example
  - ✓ Step 2-1 (2^{nd}): Assign each instance to the closest center



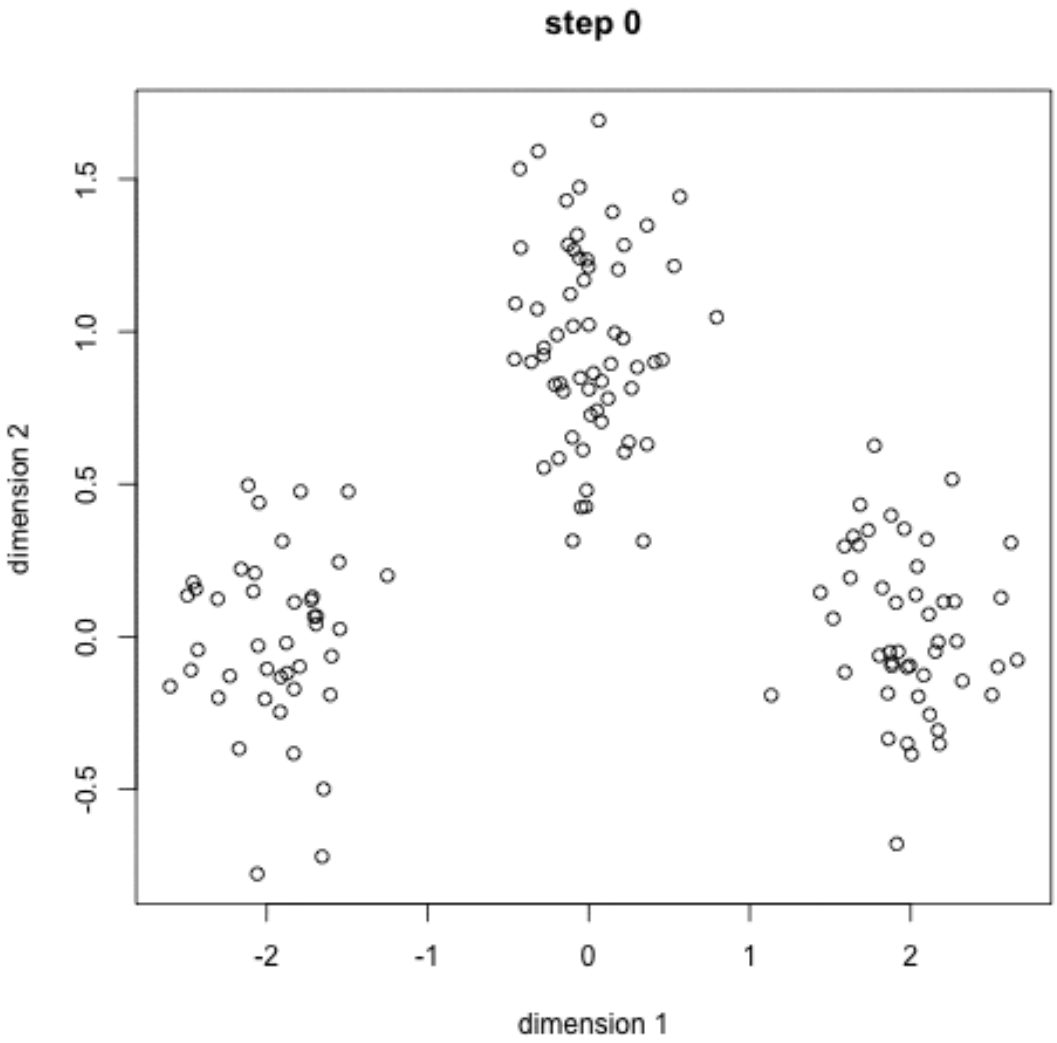
  - ✓ Step 2-2 (2^{nd}): Re-compute the centroids based on the assigned instances



  - ✓ Stop the algorithm because there is no change for centroids and membership assignment

K-Means Clustering

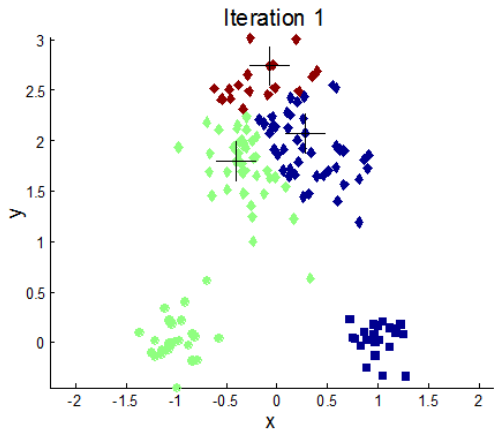
- KMC example

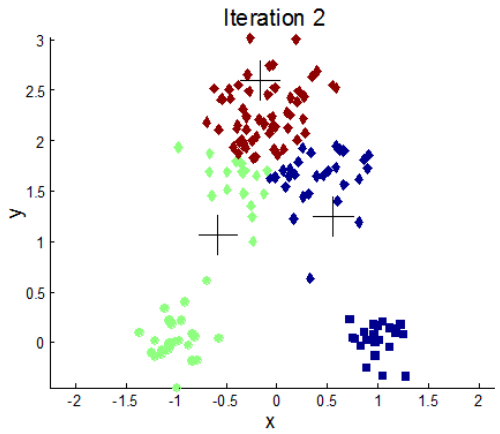


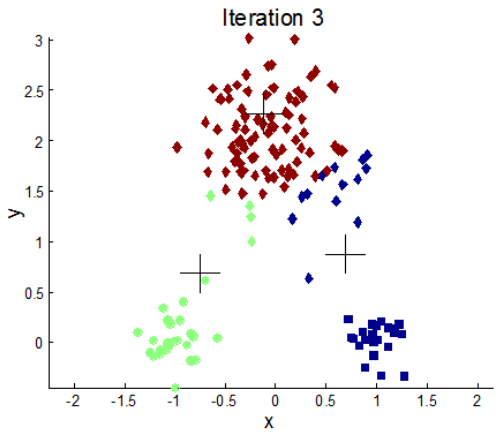
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

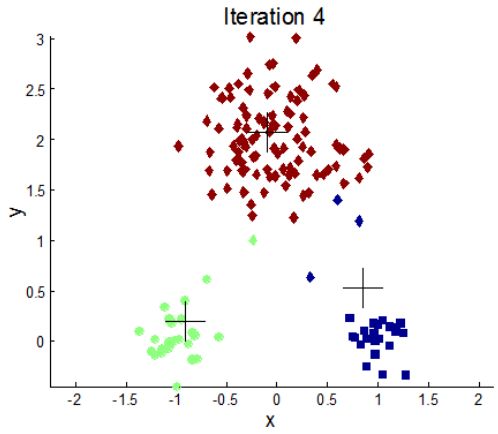
K-Means Clustering

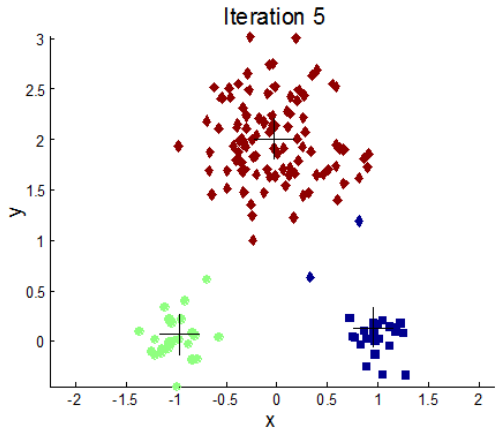
- Effect of initial centroids
  - ✓ Desirable centroid selection

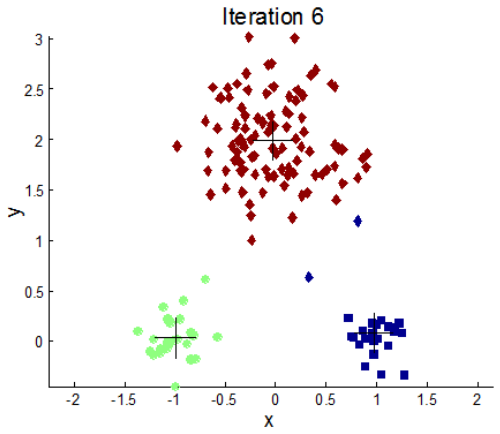






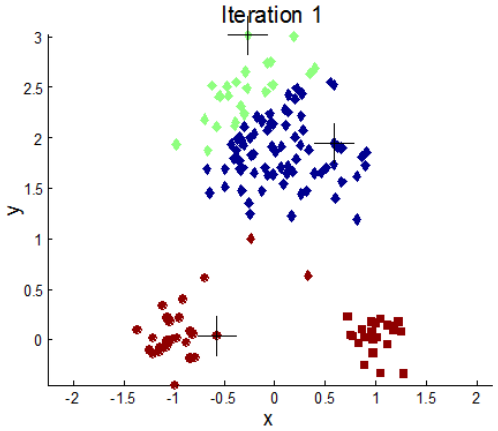


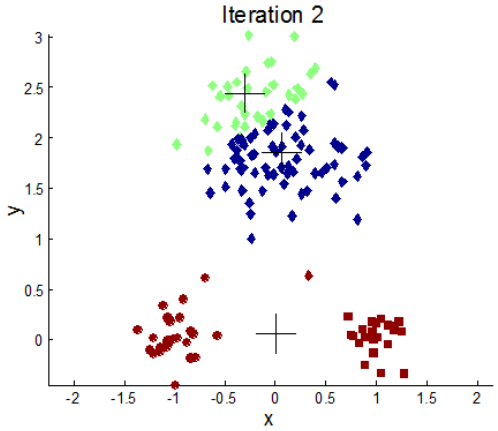


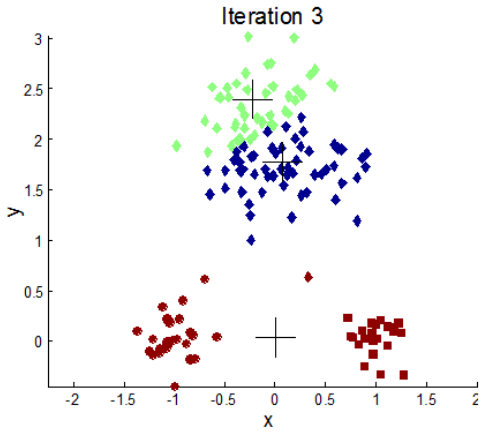


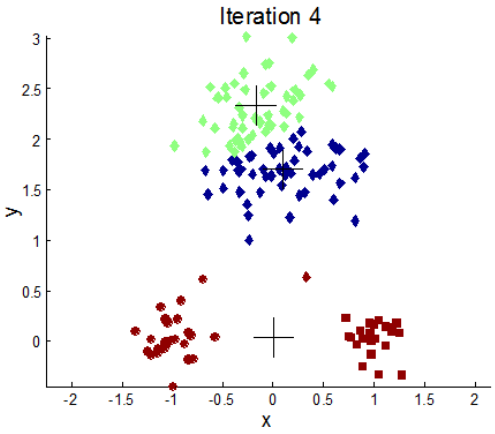
K-Means Clustering

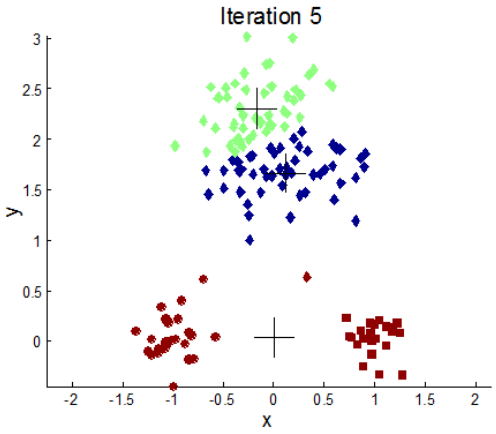
- Effects of initial centroids
  - ✓ Undesirable centroid selection







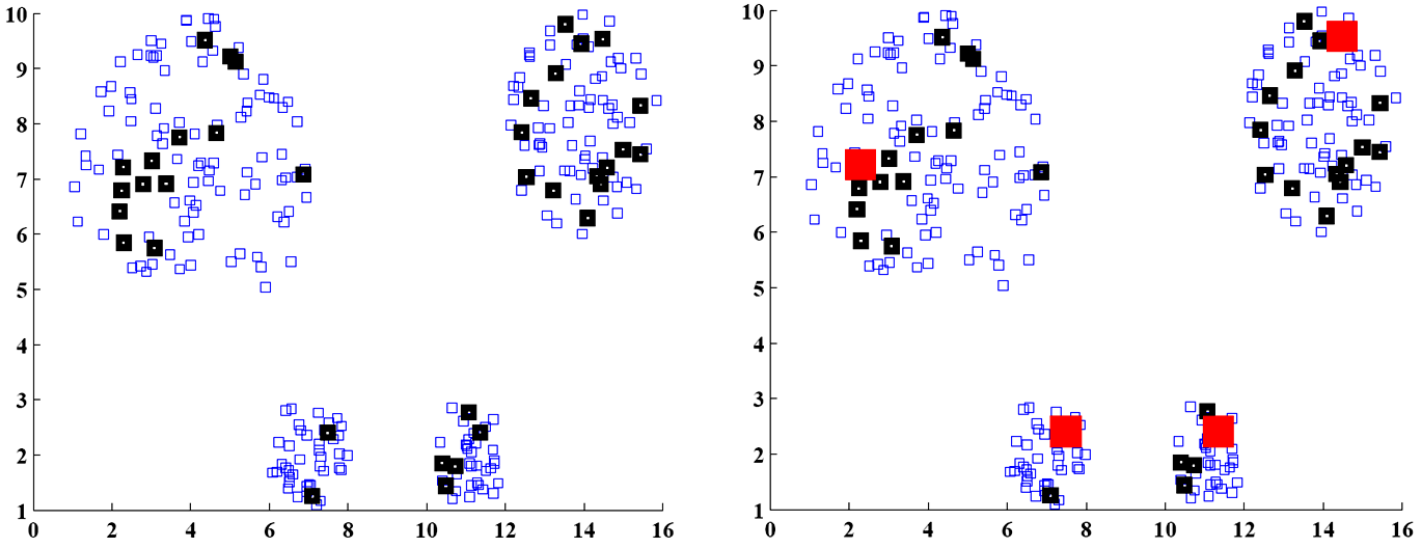




K-Means Clustering

- Some remedies for initial centroid selection
  - ✓ Multiple runs
  - ✓ Sample and use hierarchical clustering to determine initial centroids
  - ✓ Preprocessing & Postprocessing

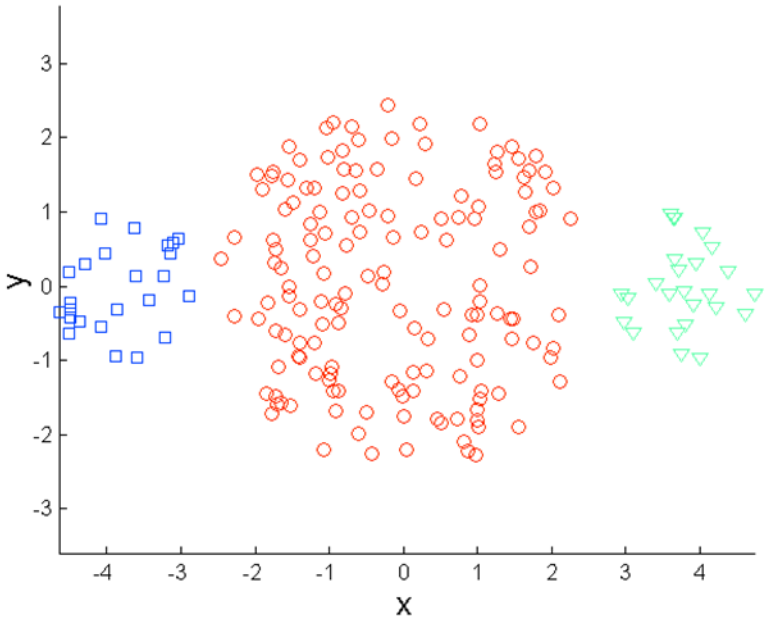
$$\mathcal{L}(\mathbf{x}_s|\mathbf{S},\mathbf{C}) = d_G(\mathbf{x}_s,\mathbf{S}) \times \frac{1}{1+\exp\big(-d_R(\mathbf{x}_s,\mathbf{S})\big)}$$

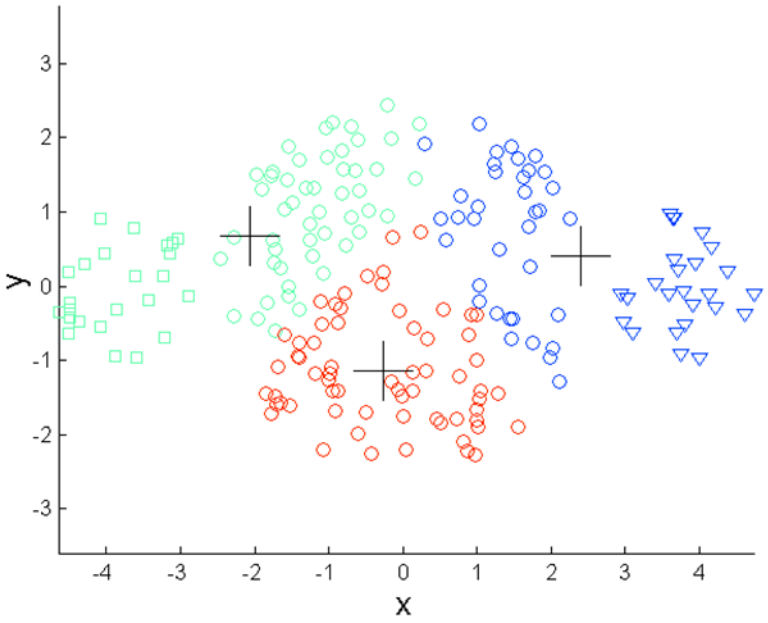


Pilsung Kang and Sungzoon Cho. (2009). K-Means clustering seeds initialization based on centrality, sparsity, and isotropy. *The 13th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2009)*, Burgos, Spain. E. Corchado and H. Yin (Eds.), ***Lecture Notes in Computer Science LNCS 5788***, 109-117.

K-Means Clustering

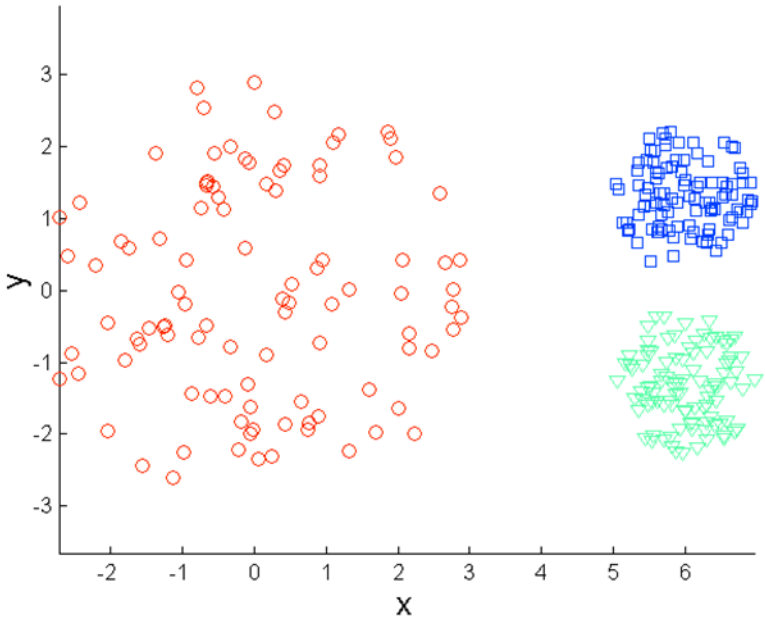
- Limitations of K-Means Clustering
  - ✓ Cannot cope with different sizes

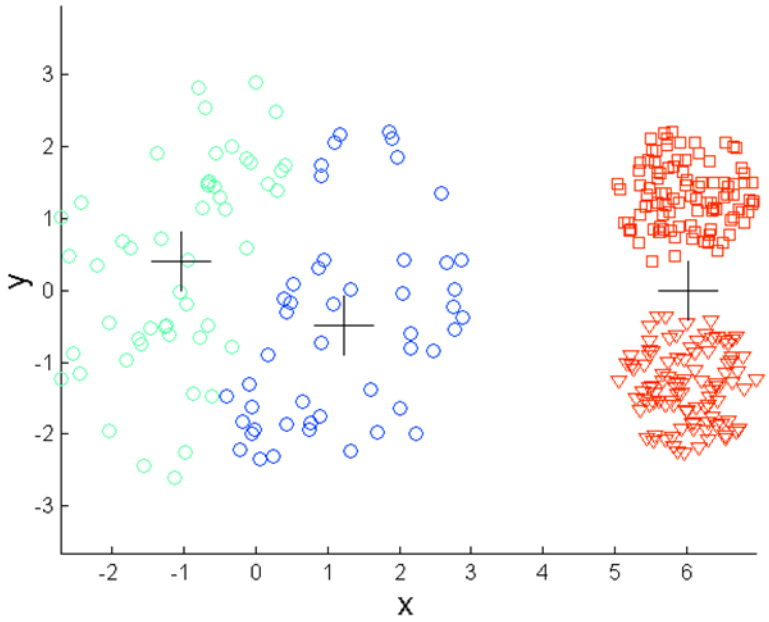




K-Means Clustering

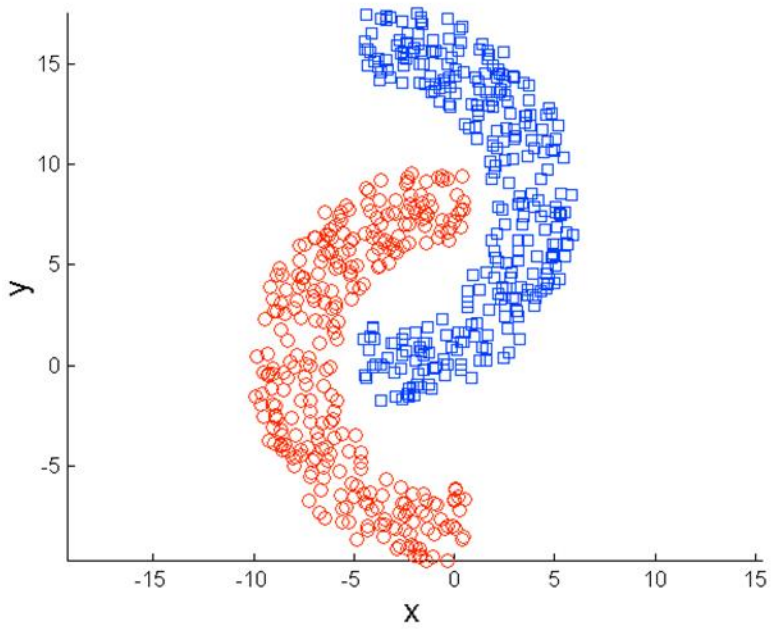
- Limitations of K-Means Clustering
  - ✓ Cannot cope with different densities

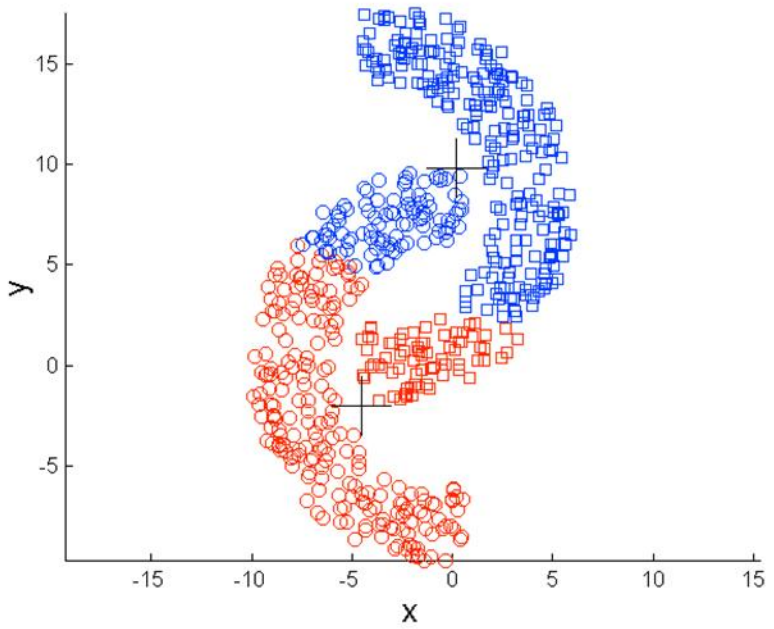




K-Means Clustering

- Limitations of K-Means Clustering
  - ✓ Cannot cope with non-globular shapes






ANY
questions ?