MACHINE INTELLIGENCE

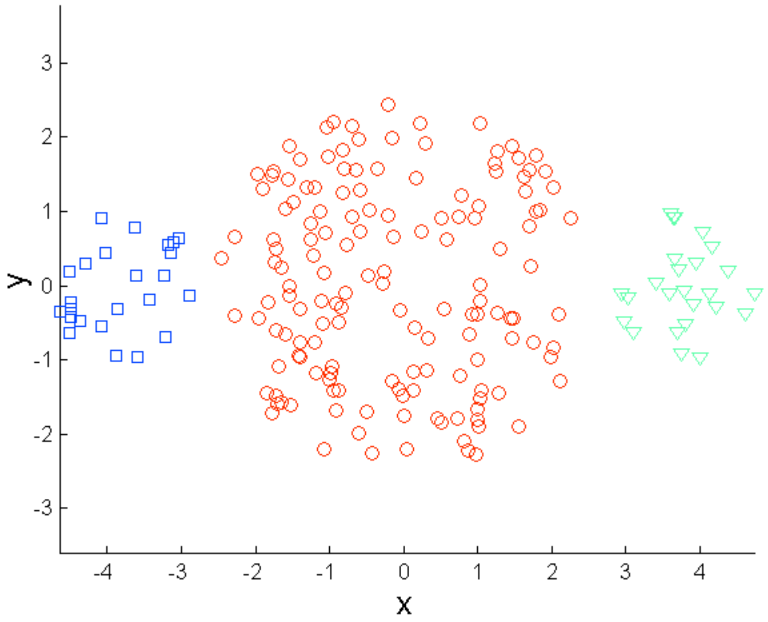
Dr. N MEHALA
Department of Computer Science
and Engineering

MACHINE INTELLIGENCE

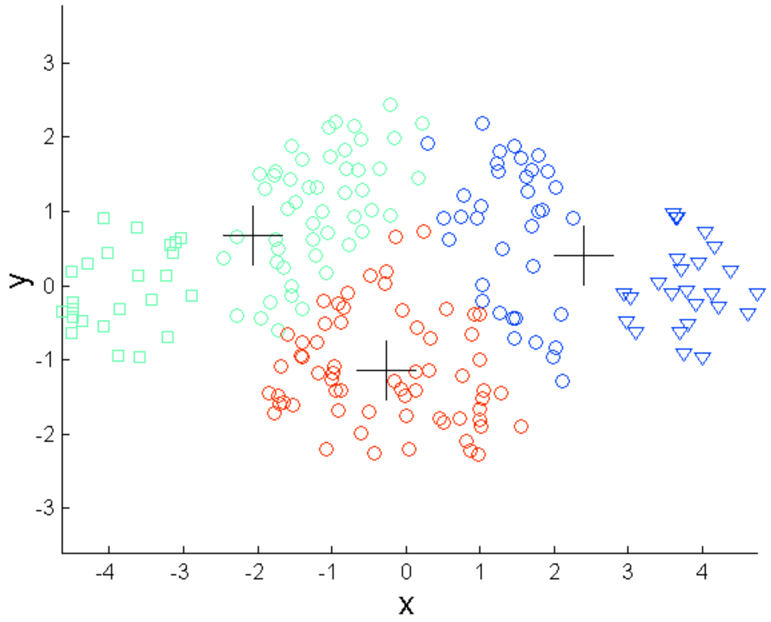
Module 4 [Unsupervised Learning]

Dr. N MEHALA
Department of Computer Science and Engineering

Limitations of K-means: Differing Sizes

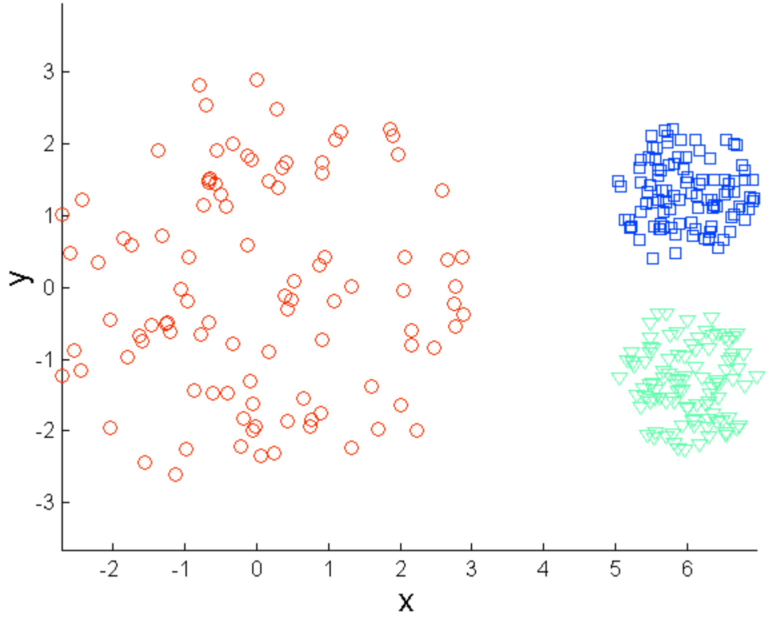


Original Points

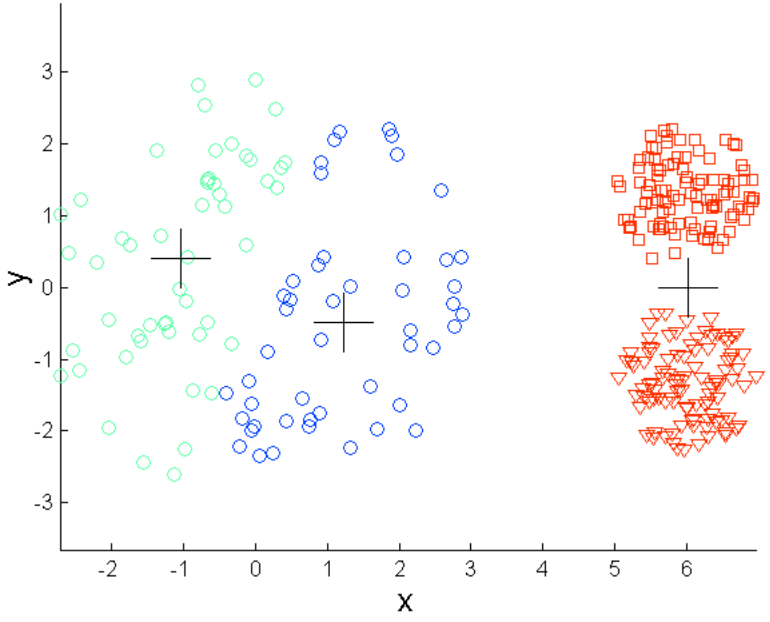


K-means (3 Clusters)

Limitations of K-means: Differing Density

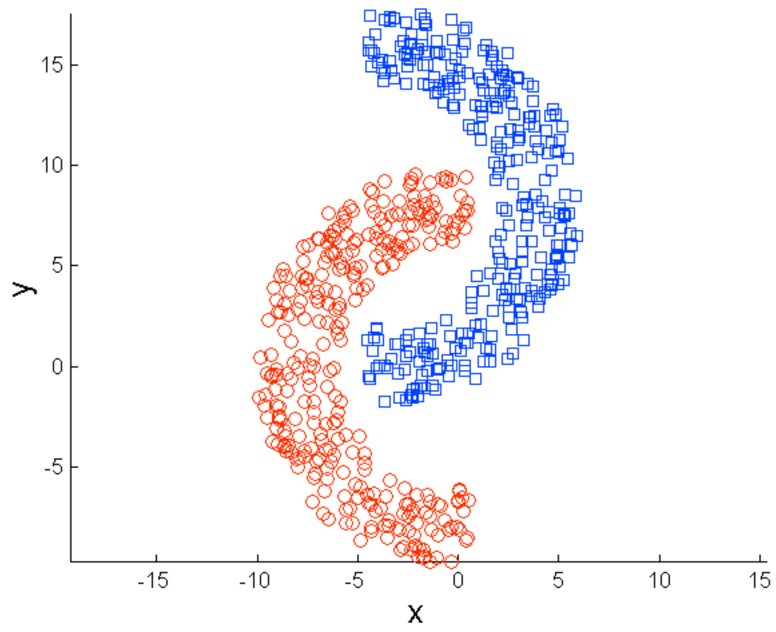


Original Points

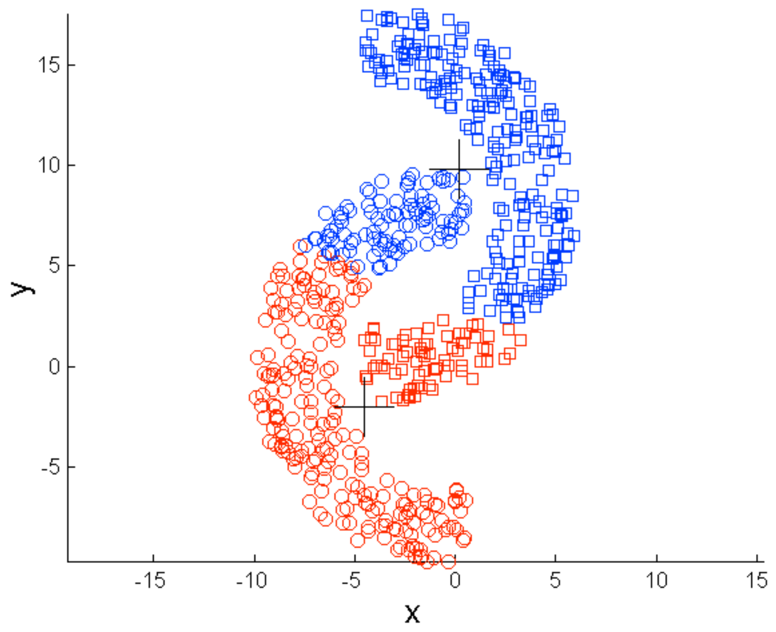


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

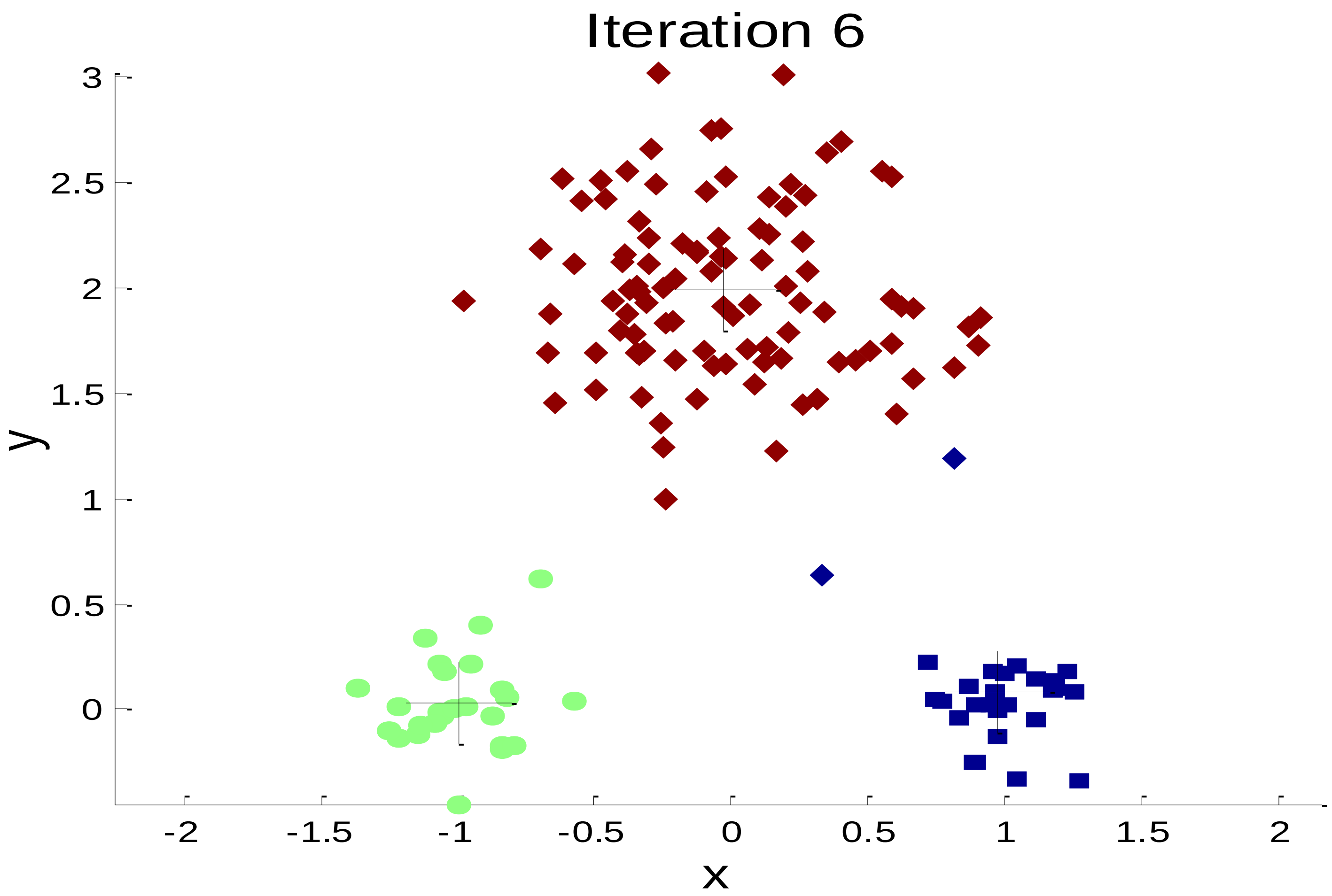


Original Points

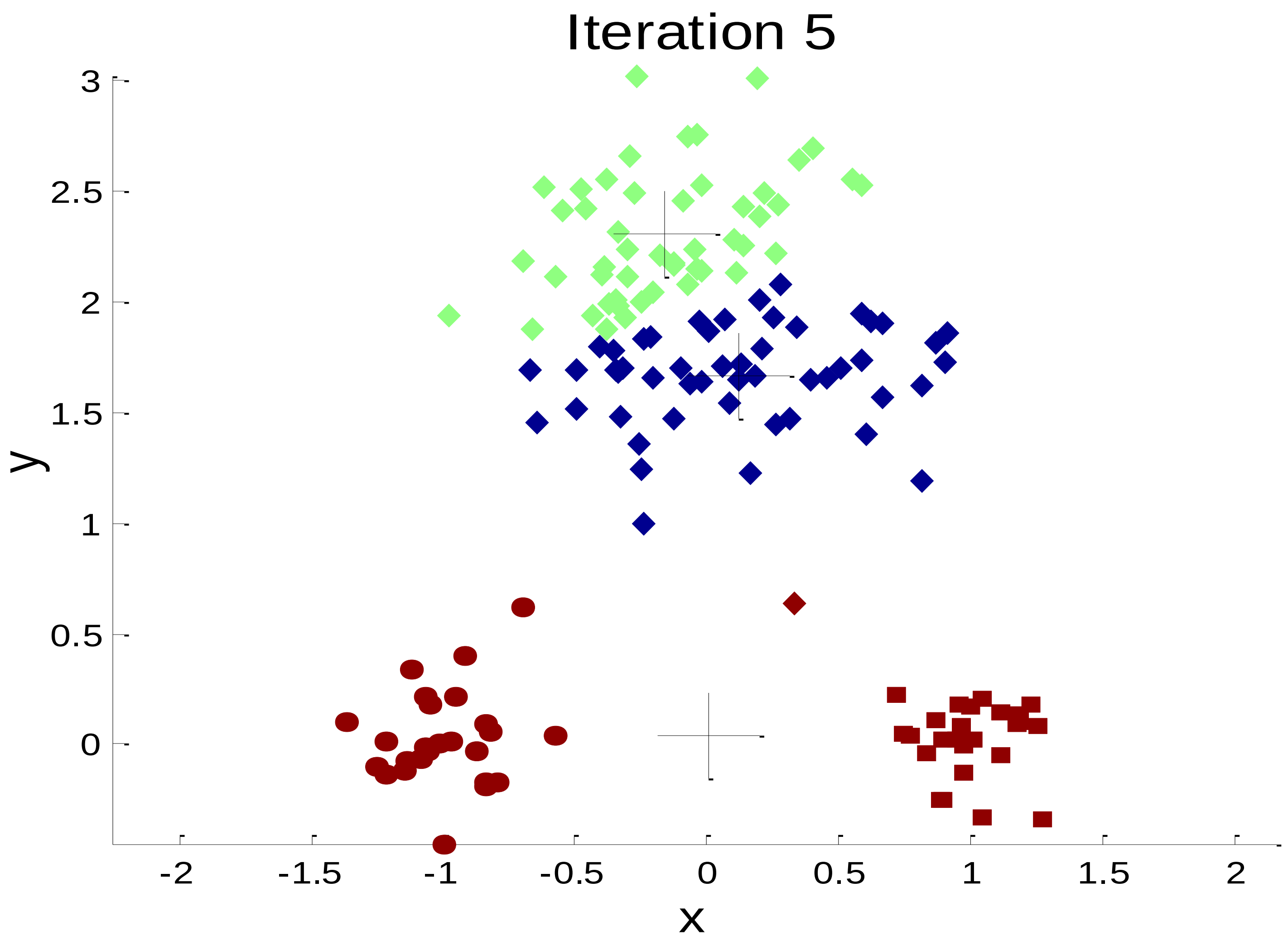


K-means (2 Clusters)

Importance of Choosing Initial Centroids



Importance of Choosing Initial Centroids



Problems with Selecting Initial Points

- If there are K ‘real’ clusters then the chance of selecting one centroid from each cluster is small.
  - Chance is relatively small when K is large
  - If clusters are the same size, n, then

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

  - For example, if K = 10, then probability = $10!/10^{10}$ = 0.00036
  - Sometimes the initial centroids will readjust themselves in ‘right’ way, and sometimes they don’t

Solutions to Initial Centroids Problem

- Multiple runs
  - Helps, but probability is not on our side
- Sample and use hierarchical clustering to determine initial centroids
- Select more than k initial centroids and then select among these initial centroids
  - Select most widely separated
- Postprocessing
- Generate a larger number of clusters and then perform a hierarchical clustering
- Bisecting K-means
  - Not as susceptible to initialization issues

Bisecting K-Means Clustering Algorithm

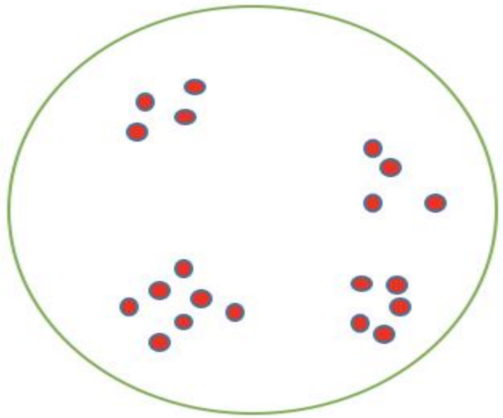
- To overcome the problem of poor clusters because of k-means getting caught in a local minimum
- Variant of K-means that can produce a partitional or a hierarchical clustering
- Instead of partitioning the data set into K clusters in each iteration, bisecting k-means algorithm splits one cluster into two sub clusters at each bisecting step (by using k-means) until k clusters are obtained.
- Hybrid approach between Divisive Hierarchical Clustering (top down clustering) and K-means Clustering.

Note: A local minimum means that the result is good but not necessarily the best possible. A global minimum is the best possible

How it Works?

Step1: Set K to define the number of cluster

Step2: Set all data as a single cluster

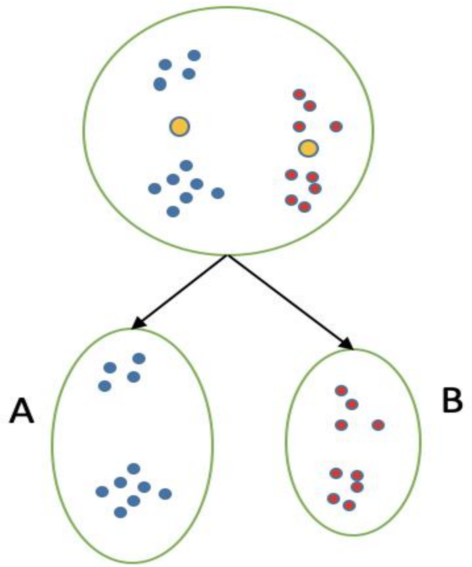


How it Works?

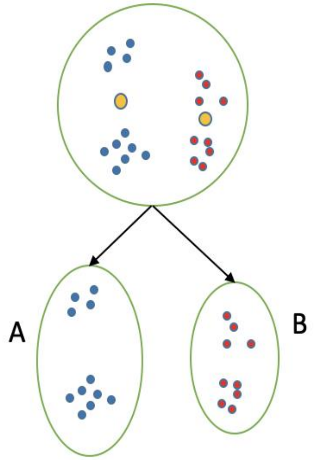
Step1: Set K to define the number of cluster

Step2: Set all data as a single cluster

Step3: Use K-means with K=2 to split the cluster



How it Works?



Step1: Set K to define the number of cluster

Step2: Set all data as a single cluster

Step3: Use K-means with K=2 to split the cluster

Step4: Measure the distance for each intra cluster

- Sum of square Distance

$$\sum_{i=0}^{n} (X_i - \overline{X})^2$$

Step5: Select the cluster that have the largest distance and split to 2 cluster using K-means

How it Works?

Step1: Set K to define the number of cluster

Step2: Set all data as a single cluster

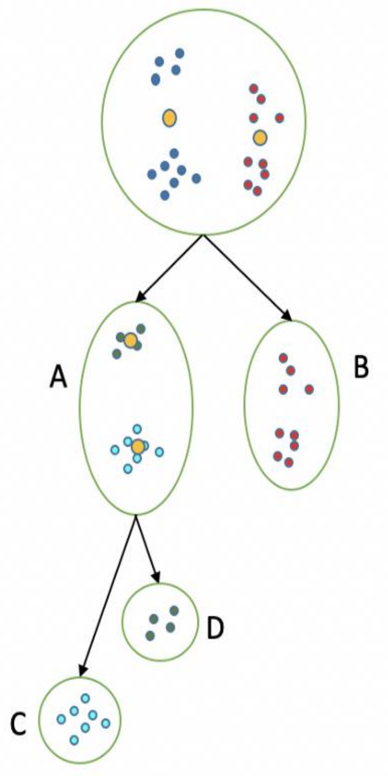
Step3: Use K-means with K=2 to split the cluster

Step4: Measure the distance for each intra cluster

- Sum of square Distance

$$\sum_{i=0}^{n} (X_i - \overline{X})^2$$

Step5: Select the cluster that have the largest distance and split to 2 cluster using K-means



How it Works?

Step1: Set K to define the number of cluster

Step2: Set all data as a single cluster

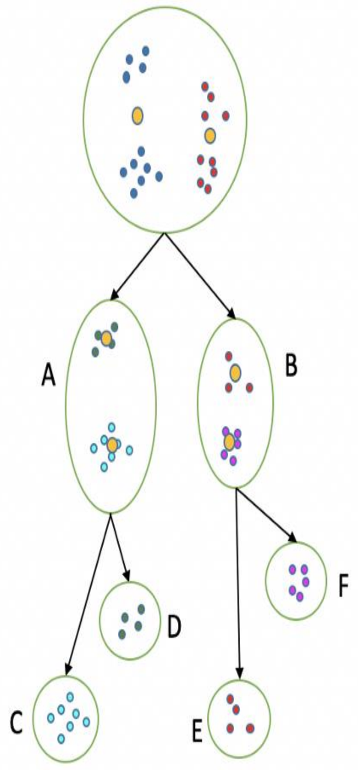
Step3: Use K-means with K=2 to split the cluster

Step4: Measure the distance for each intra cluster

- Sum of square Distance

Step5: Select the cluster that have the largest distance and split to 2 cluster using K-means

Step6: Repeat step 3–5 until the number of leaf cluster = K.



Advantage of Bisecting K-Means over K-Means

- Bisecting k-means is more efficient when **K** is large.
- Bisecting k-means produce clusters of similar sizes, while k-means is known to produce clusters of widely different sizes.

Summary

- K-Means Issues
- Bisecting K-Means Clustering

Resources

- http://www2.ift.ulaval.ca/~chaib/IFT-4102-7025/public_html/Fichiers/Machine_Learning_in_Action.pdf
- http://wwwusers.cs.umn.edu/~kumar/dmbook/.
- ftp://ftp.aw.com/cseng/authors/tan
- http://web.ccsu.edu/datamining/resources.html

THANK YOU

Dr. N MEHALA
Department of Computer Science and Engineering

mehala@pes.edu