Data Mining

CLUSTERING: K-MEANS ISSUES AND K-MEDOID

Prof. Dr. Hikmat Ullah Khan
Department of Information Technology

UNIVERSITY OF SARGODHA, SARGODHA

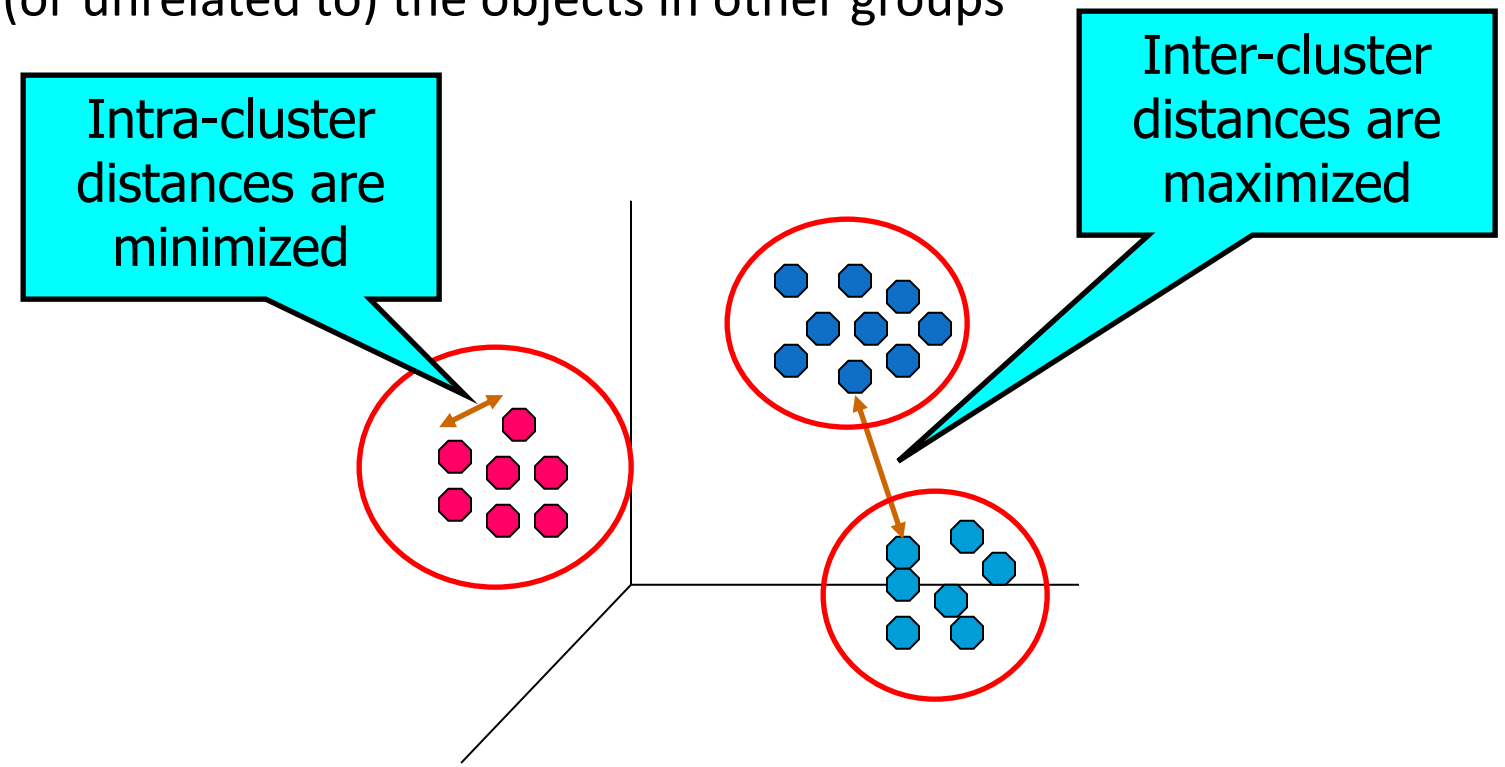
Lesson from Holy Quran

يَوْمَئِذٍ يَصْدُرُ النَّاسُ أَشْتَاتًا لِّيُرَوْا أَعْمَالَهُمْ - فَمَن يَعْمَلْ مِثْقَالَ ذَرَّةٍ خَيْرًا يَرَهُ - وَمَن يَعْمَلْ مِثْقَالَ ذَرَّةٍ شَرًّا يَرَهُ

"That Day mankind will proceed in scattered groups that they may be shown their deeds. So whosoever does good equal to the weight of an atom (or a small ant), shall see it. And whosoever does evil equal to the weight of an atom (or a small ant), shall see it,"
[Az-Zalzalah 99: 6- 8].

What is Cluster Analysis?

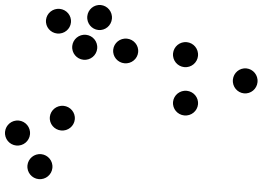
- Finding groups of objects
  - objects in a group will be similar (or related) to one another
  - different from (or unrelated to) the objects in other groups

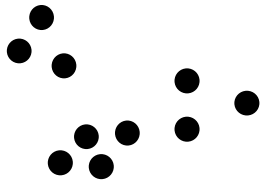


What is not Cluster Analysis?

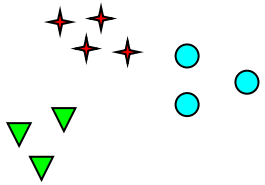
- Supervised classification
  - Have class label information
- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query
  - Groupings are a result of an external specification

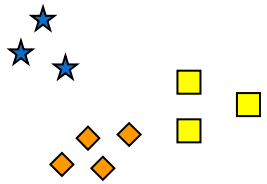
Notion of a Cluster can be Ambiguous



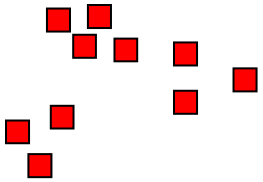


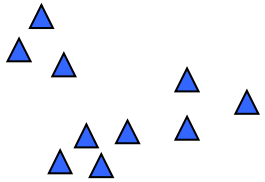
How many clusters?



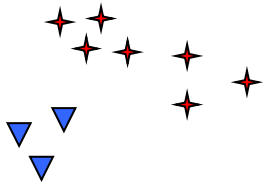


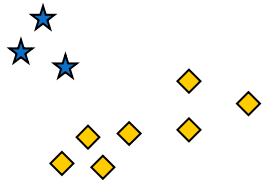
Six Clusters





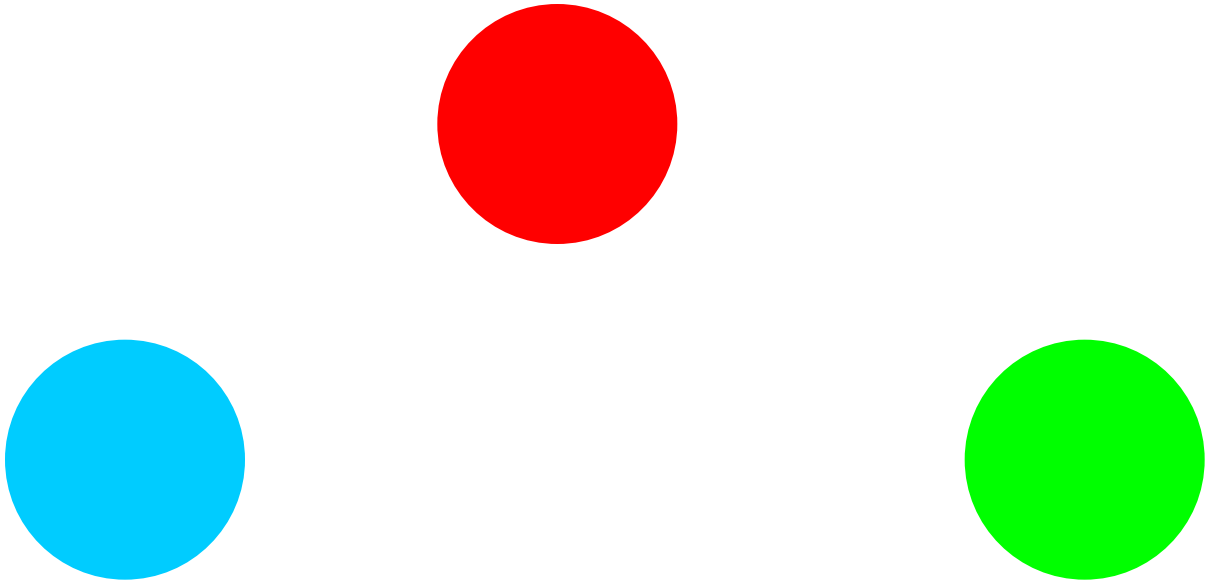
Two Clusters





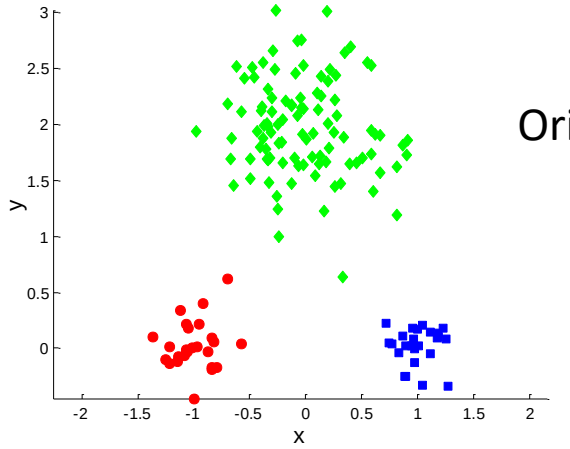
Four Clusters

- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.

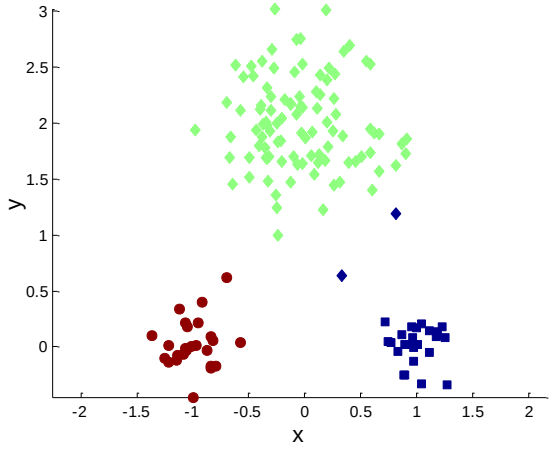


3 well-separated clusters

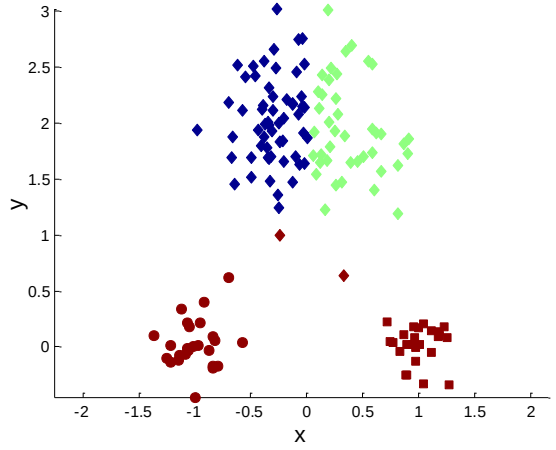
Two different K-means Clusterings



Original Points

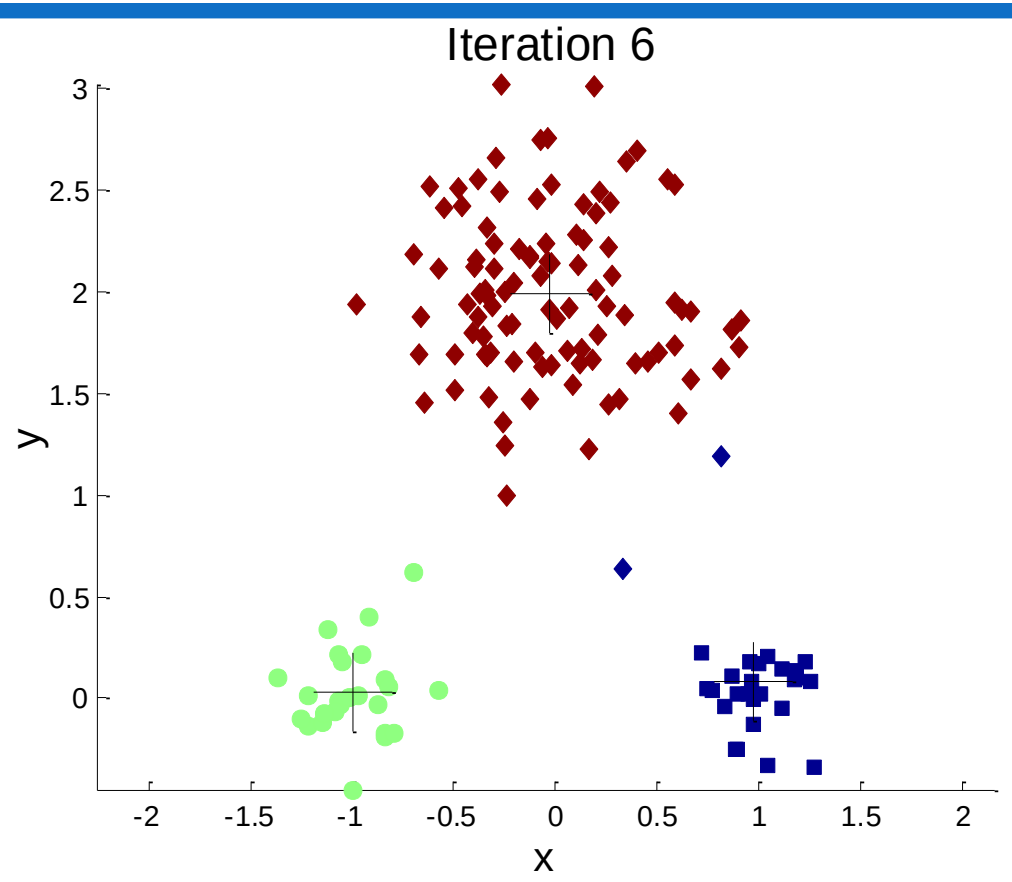


Optimal Clustering

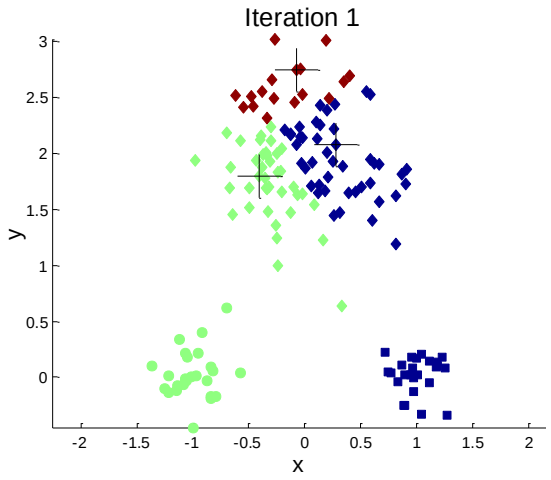


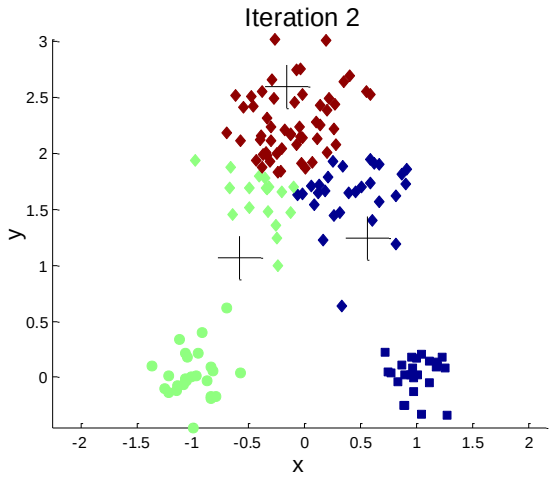
Sub-optimal Clustering

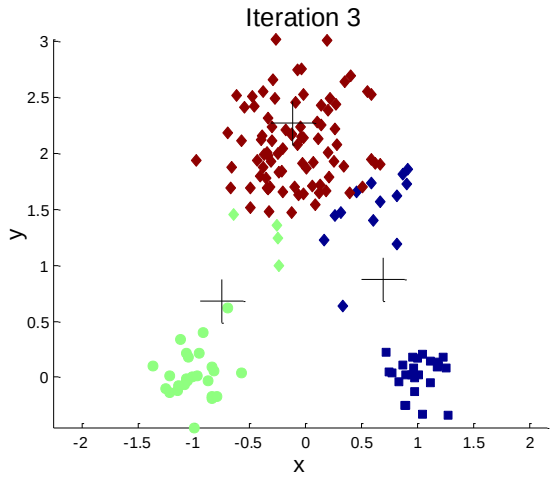
Importance of Choosing Initial Centroids

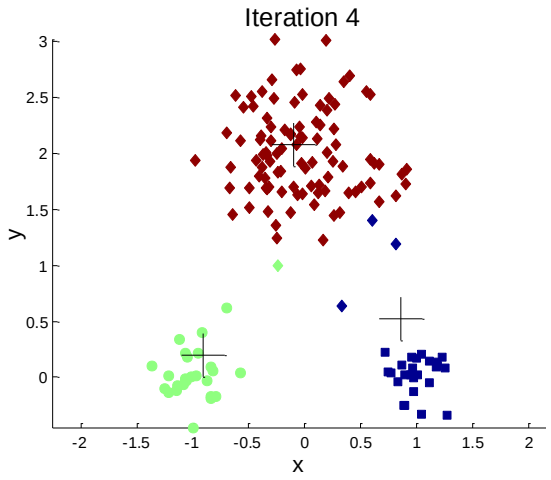


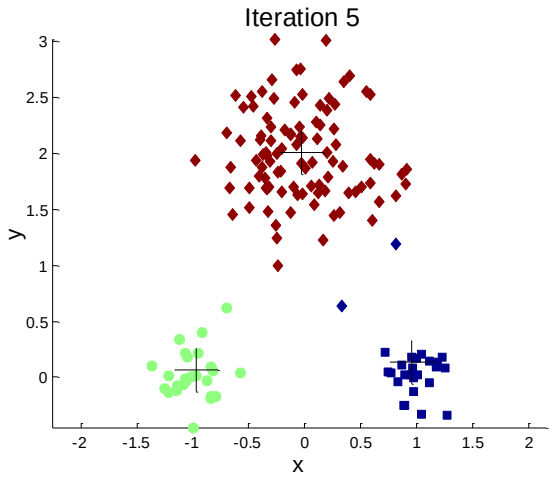
Importance of Choosing Initial Centroids

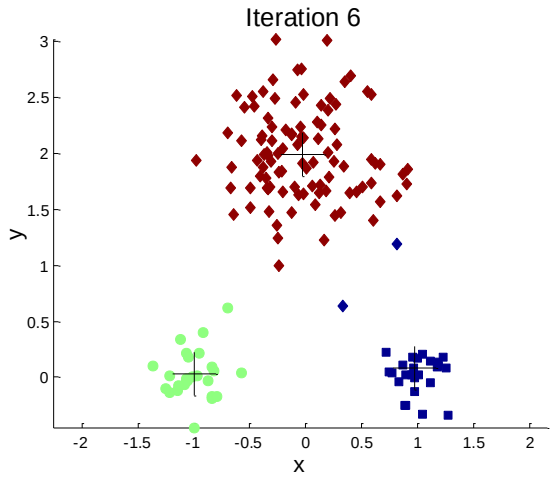




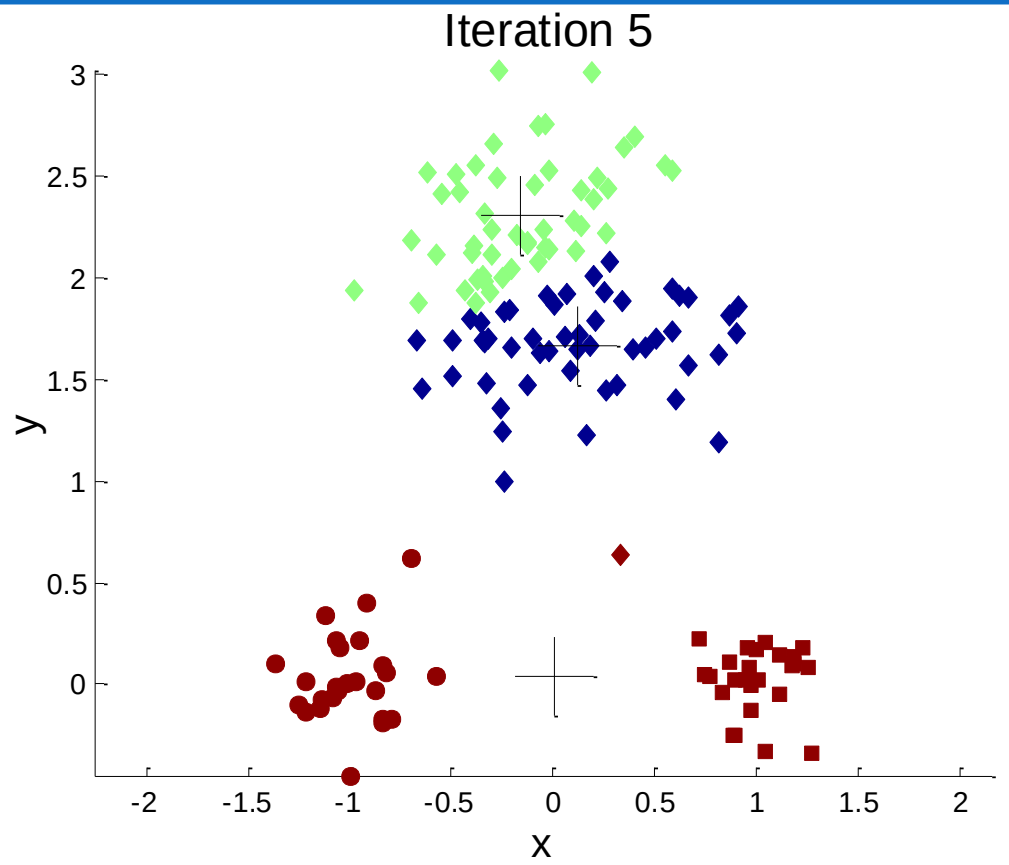




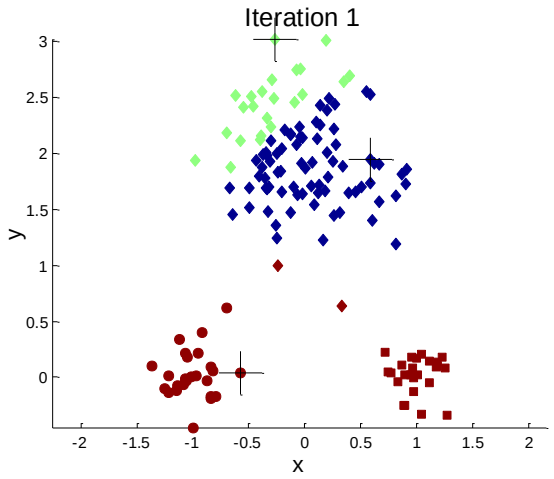


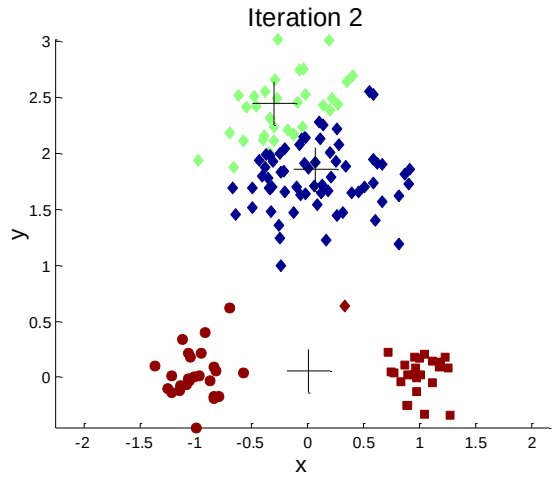


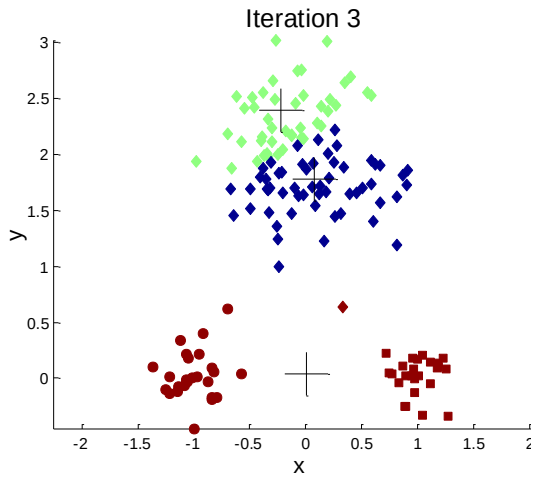
Importance of Choosing Initial Centroids …

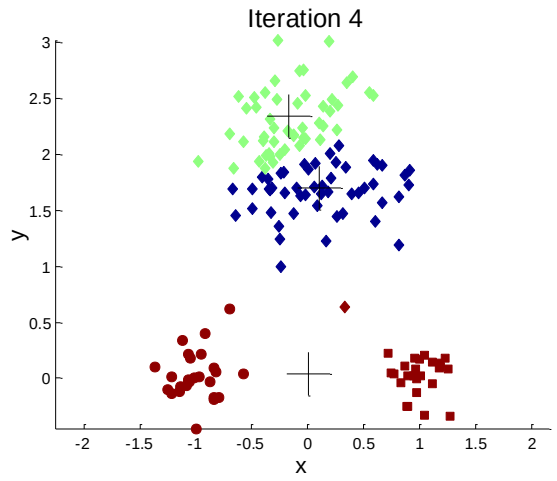


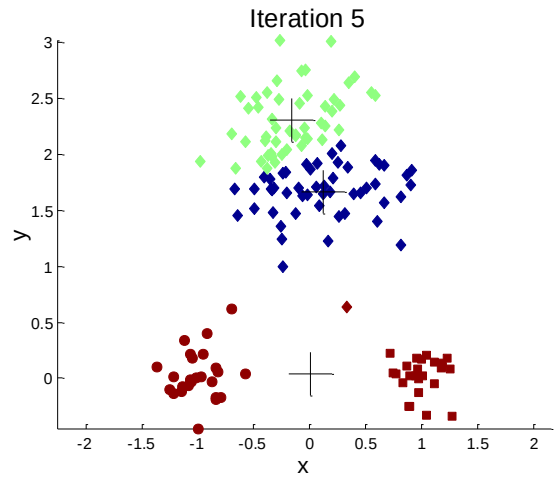
Importance of Choosing Initial Centroids …



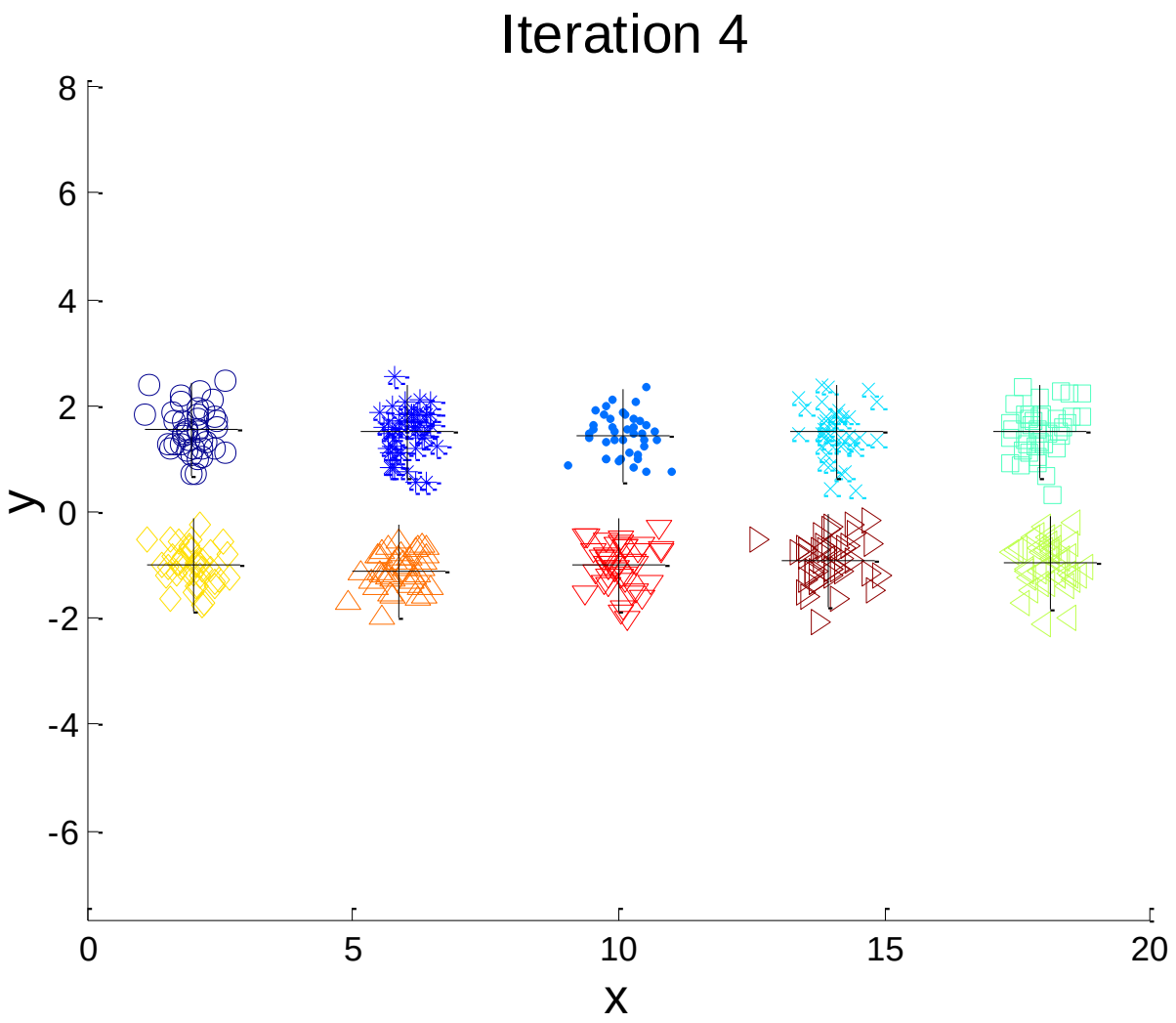






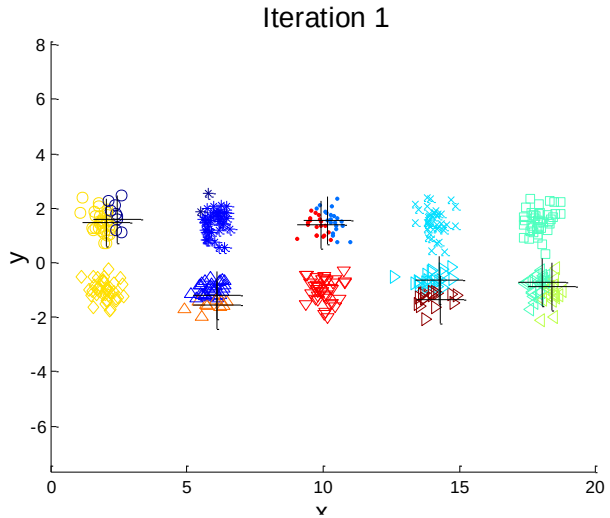


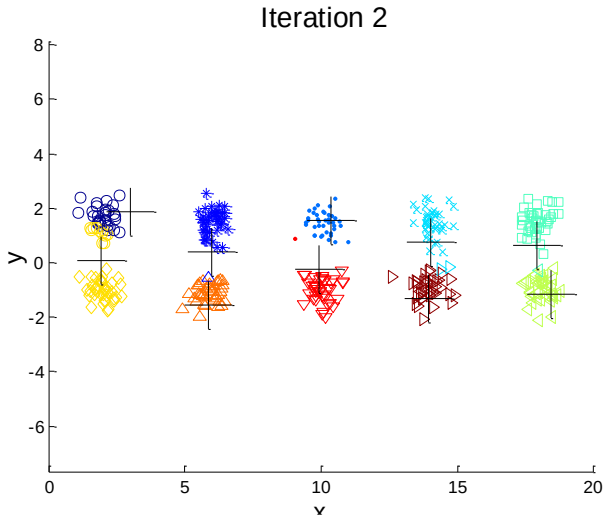
10 Clusters Example

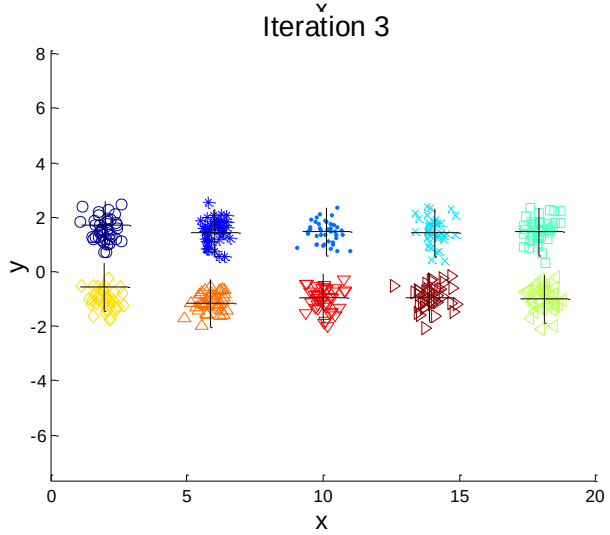


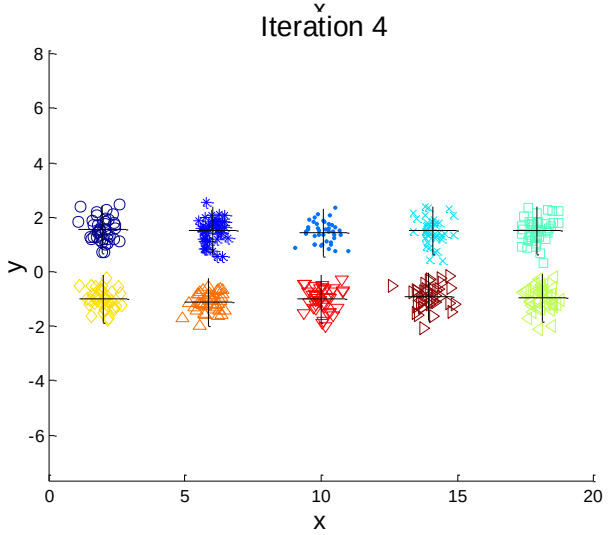
Starting with two initial centroids in one cluster of each pair of clusters

10 Clusters Example



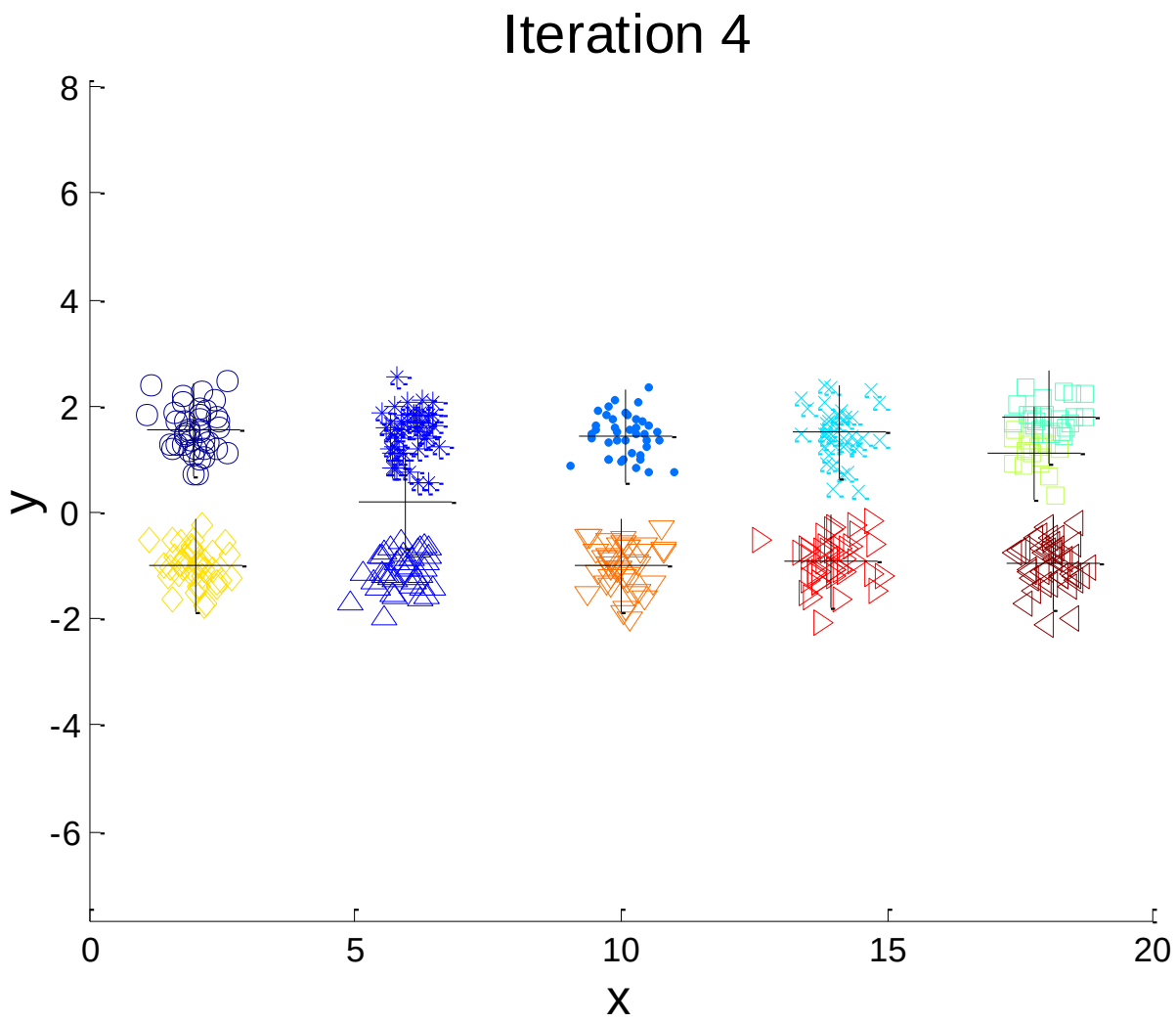






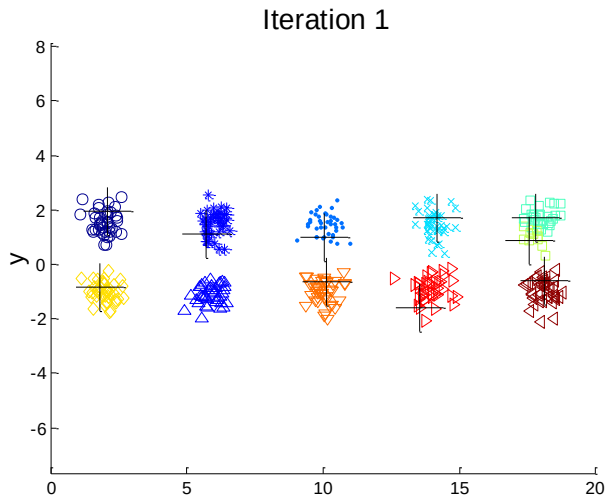
Starting with two initial centroids in one cluster of each pair of clusters

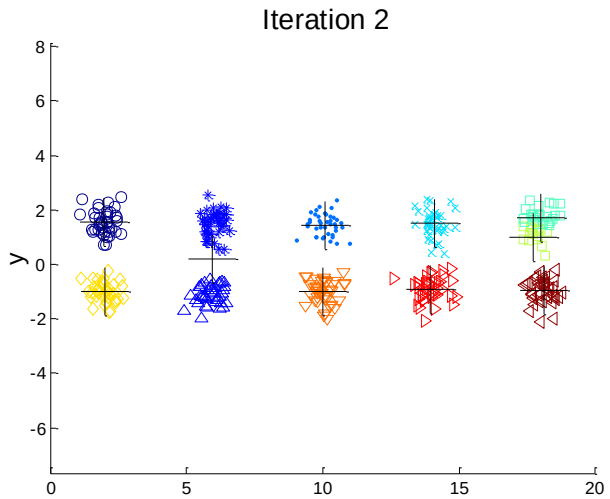
10 Clusters Example

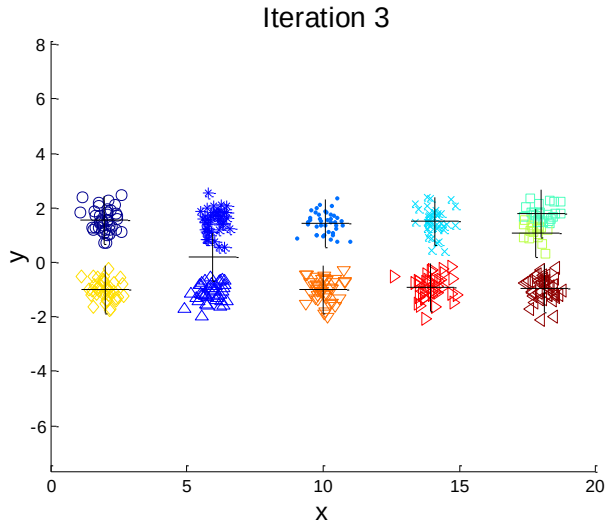


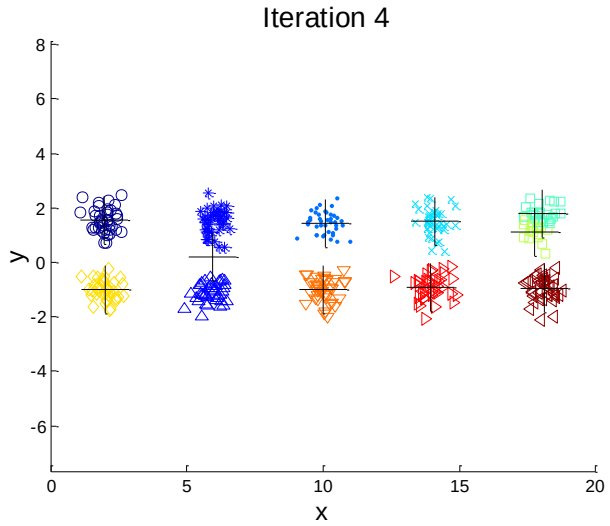
Starting with some pairs of clusters having three initial centroids, while other have only one.

10 Clusters Example



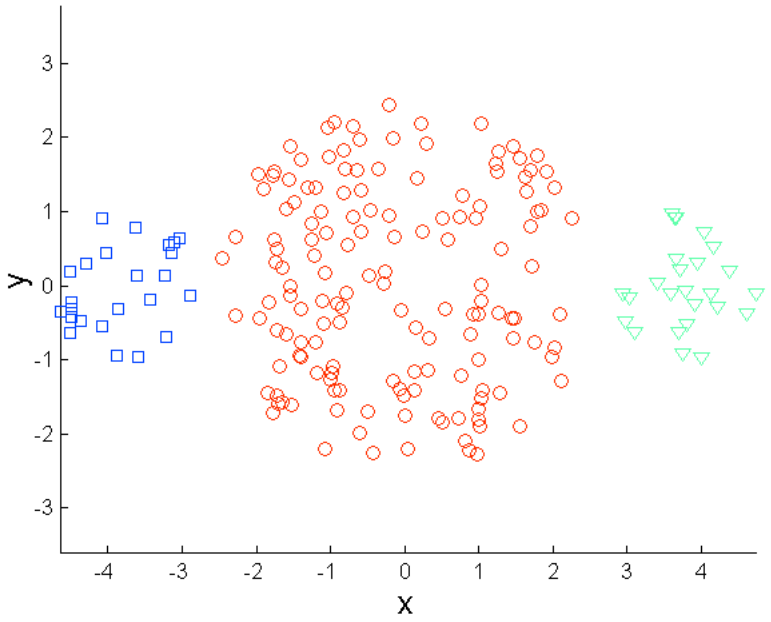




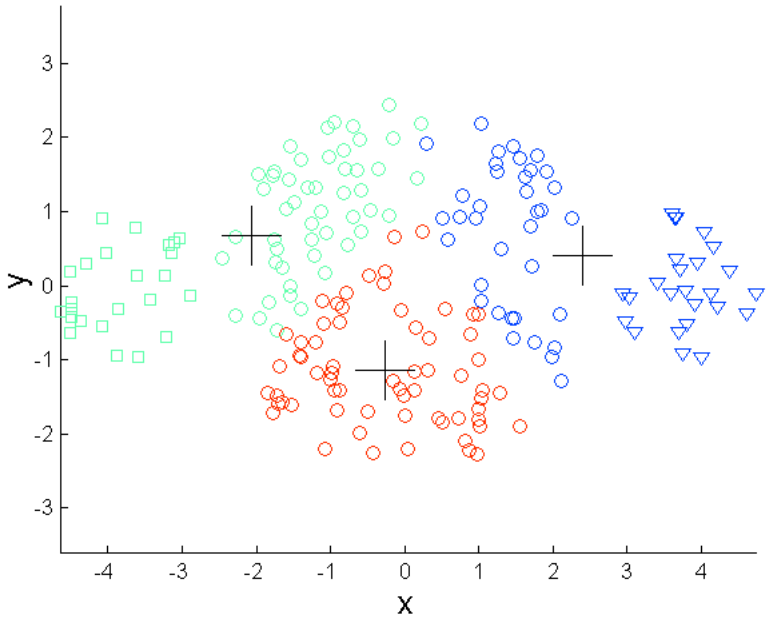


Starting with some pairs of clusters having three initial centroids, while other have only one.

Limitations of K-means: Differing Sizes

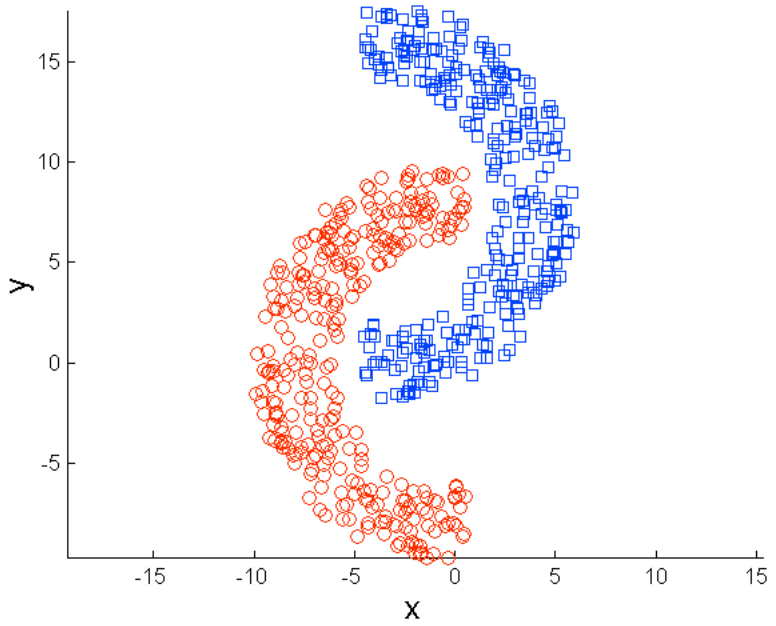


Original Points

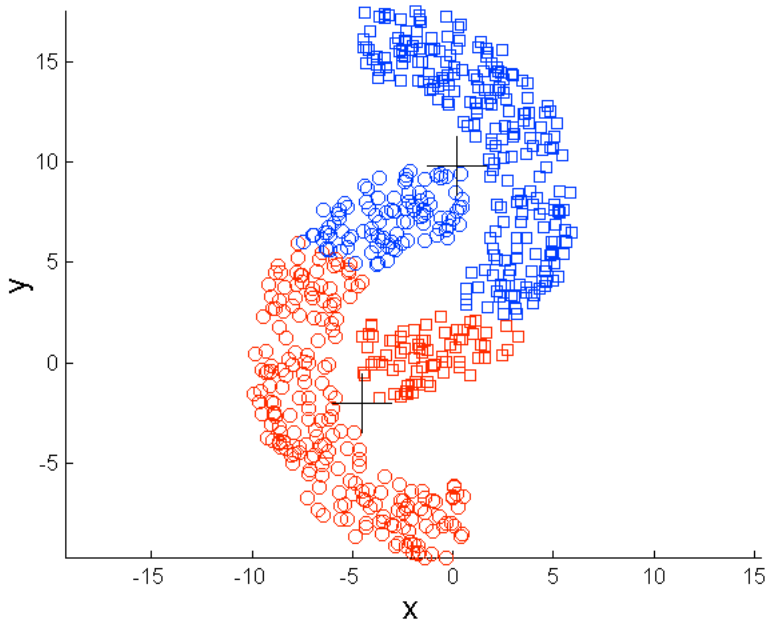


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

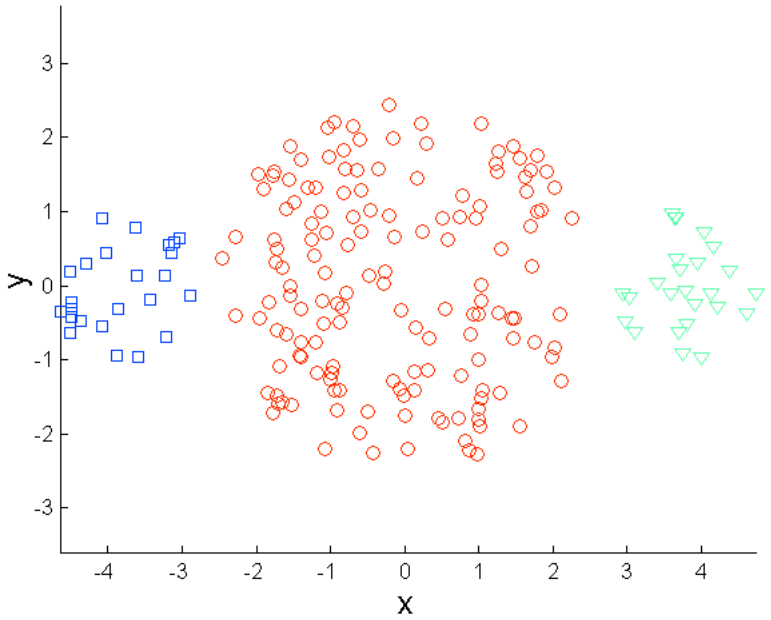


Original Points

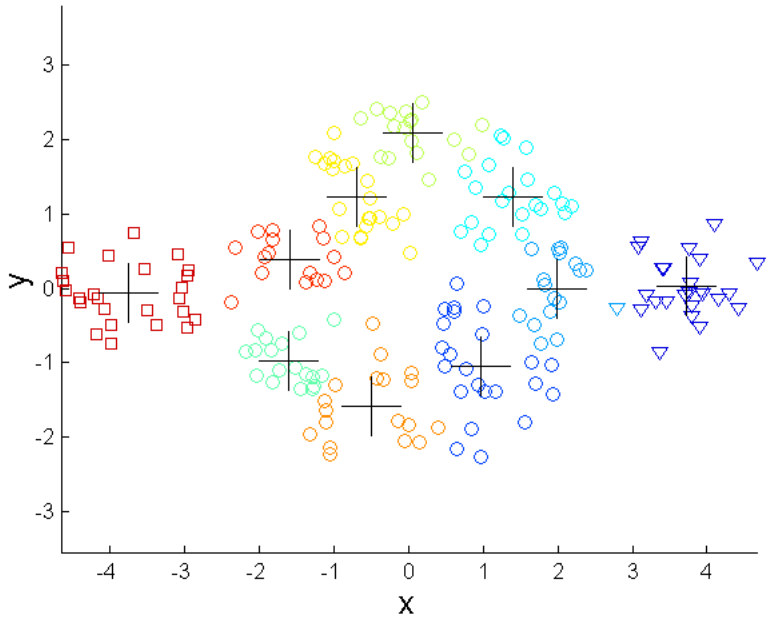


K-means (2 Clusters)

Overcoming K-means Limitations

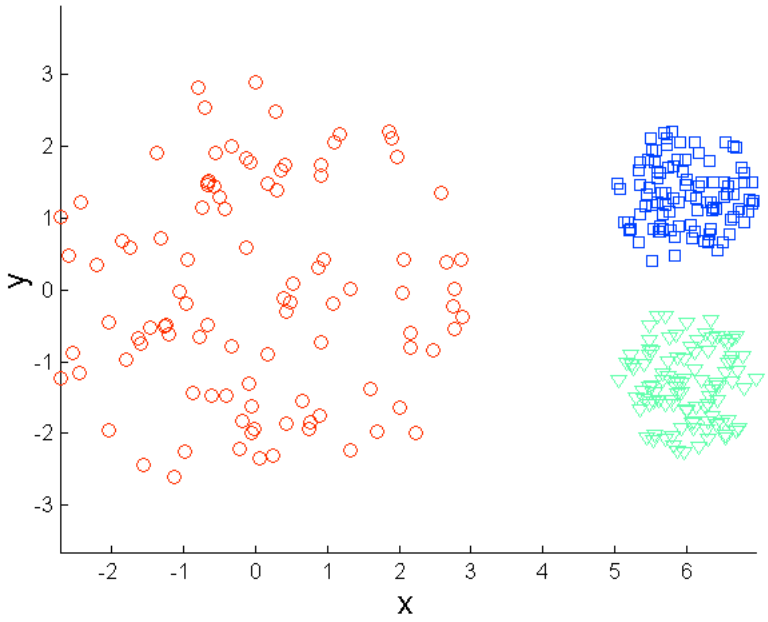


Original Points

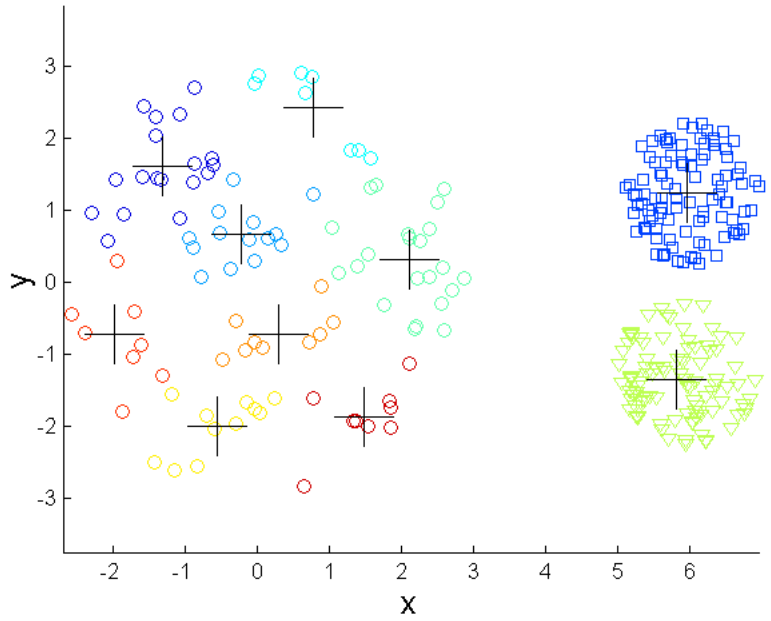


K-means Clusters

Overcoming K-means Limitations

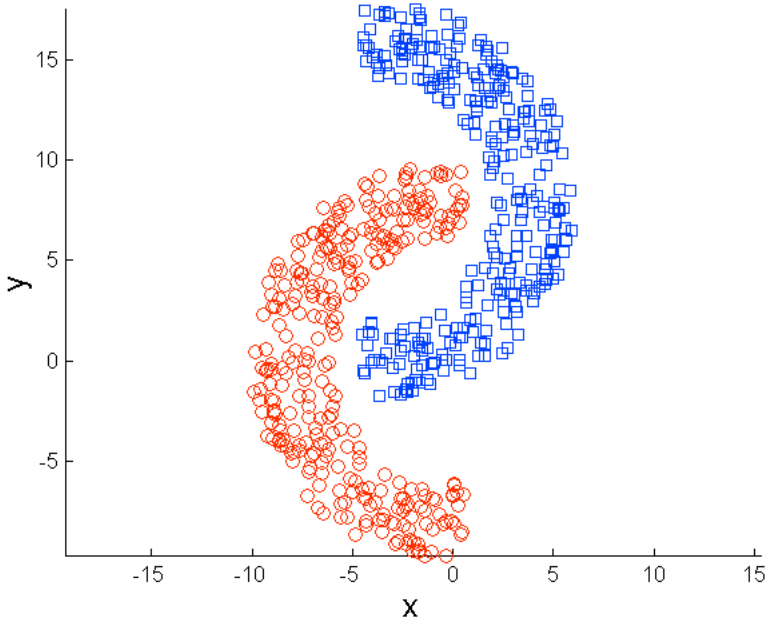


Original Points

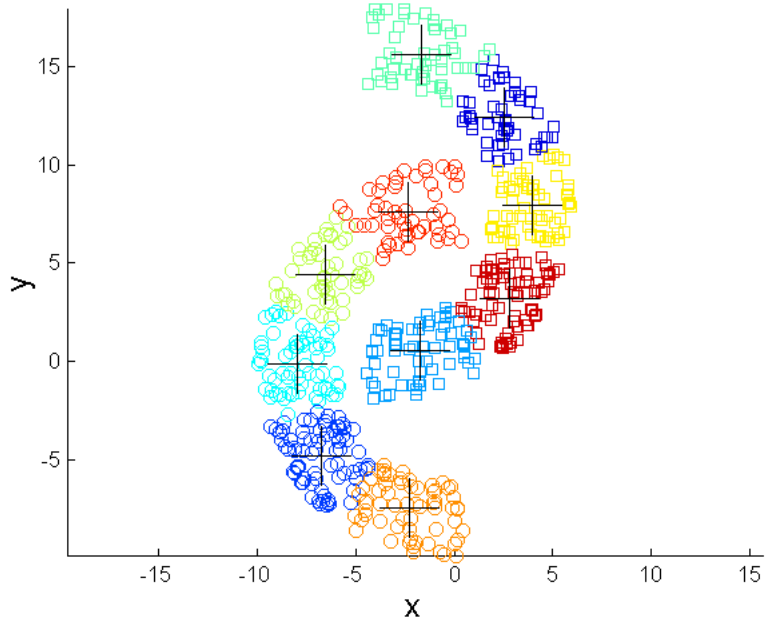


K-means Clusters

Overcoming K-means Limitations



Original Points



K-means Clusters

Solutions to Initial Centroids Problem

- Multiple runs
  - Helps, (but probability is not on your side)
- Use hierarchical clustering to determine initial centroids
- Pre-processing
  - Normalize the data
  - Eliminate outliers
- Post-processing
  - Eliminate small clusters that may represent outliers
  - Split ‘loose’ clusters,
  - Merge clusters that are ‘close’

K-Medoid

An Extension of K-Means Algorithm

K-MEDOID CLUSTERING

- The **k-medoid algorithm, an extension of k-means algorithm,**
- An algorithm to cluster *n* objects based on attributes into *k* partitions, where $k < n$.

K-Medoid

- A Partitioning based algorithm
- In k-means means are chosen as Centroid
- In k-medoid, data points are chosen to be the medoid
- A medoid is an object of a cluster whose average dissimilarity to all the objects in the cluster is minimal.

K-Medoid

- Phases:
- BUILD-Phase:
  - Sequentially selects *k* "centrally located" objects, to be used as initial medoids
- SWAP-Phase:
  - If the objective function can be reduced by interchanging (swapping) a selected object with an unselected object, then the swap is carried out. This is continued till the objective function can no longer be decreased.

Steps

1. Initially select k random points as the medoids from the given n data points of the data set.
2. Associate each data point to the closest medoid by using any of distance metrics.
3. Compute the Sum of distance of each mediod with each object in the dataset
4. Select the one with least Cost.

Example?

Point	x-axis	y-axis
1	7	6
2	2	6
3	3	8
4	8	5
5	7	4
6	4	7
7	6	2
8	7	3
9	6	4
10	3	4

Distance

- The distance formula to be used in this example is Mannhatton distance

- It is simply the
  - SUM OF THE ABSOLUTE DIFFERENCE OF POINTS AND THEIR SUM

  - Randomly selected initial central points are 5^{th} and 10^{th} objects
  - (3,4) and (7,4)

Calculation

Total Cost = cost((3, 4), (2, 6)) + cost((3, 4), (3, 8)) + cost((3, 4), (4, 7)) + cost((7, 4), (6, 2))+ cost((7, 4), (6, 4))+ cost((7, 4), (7, 3))+ cost((7, 4), (8, 5))+ cost((7, 4), (7, 6))

= 3 + 4 + 4 + 3 + 1 + 1 + 2 + 2

= 20.

- Replace (7,6) as centroid with (7,4)
- Note that (3,4) is still the other centroid.

Calculating the total cost = cost((3, 4), (2, 6)) + cost((3, 4), (3, 8)) + cost((3, 4), (4, 7)) + cost((7, 3), (7, 6)) + cost((7, 3), (8, 5)) + cost((7, 3), (6, 2)) + cost((7, 3), (7, 4)) + cost((7, 3), (6, 4))

= 3 + 4 + 4 + 3 + 3 + 2 + 1 + 2

= 22.

NOTE the Points

1. There is no convergence point
2. There can be as many computations as many combination of centroid for any value of k <n.
3. K-medoid is more robust to noise


www.words4ever.com
Life is something that everyone should try at least once
Henry J. Tillman