CLUSTERING

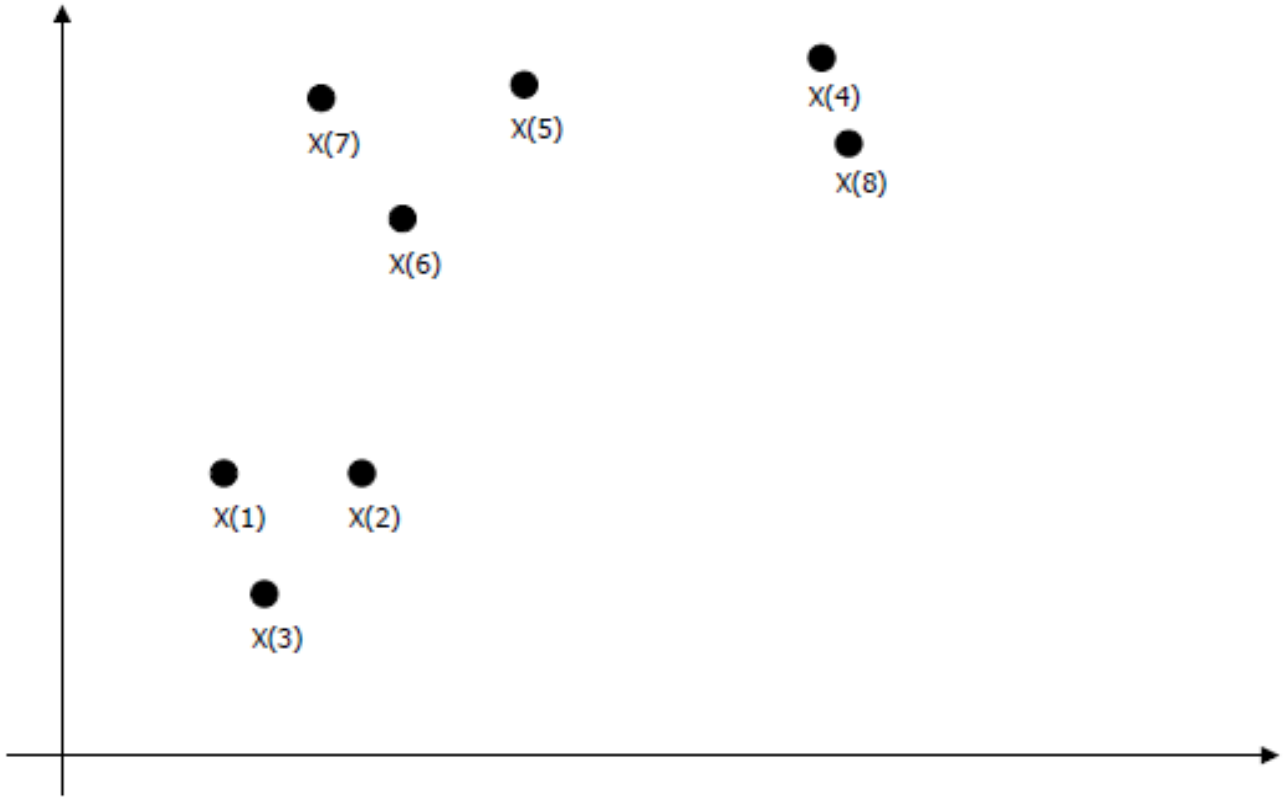
What is clustering?

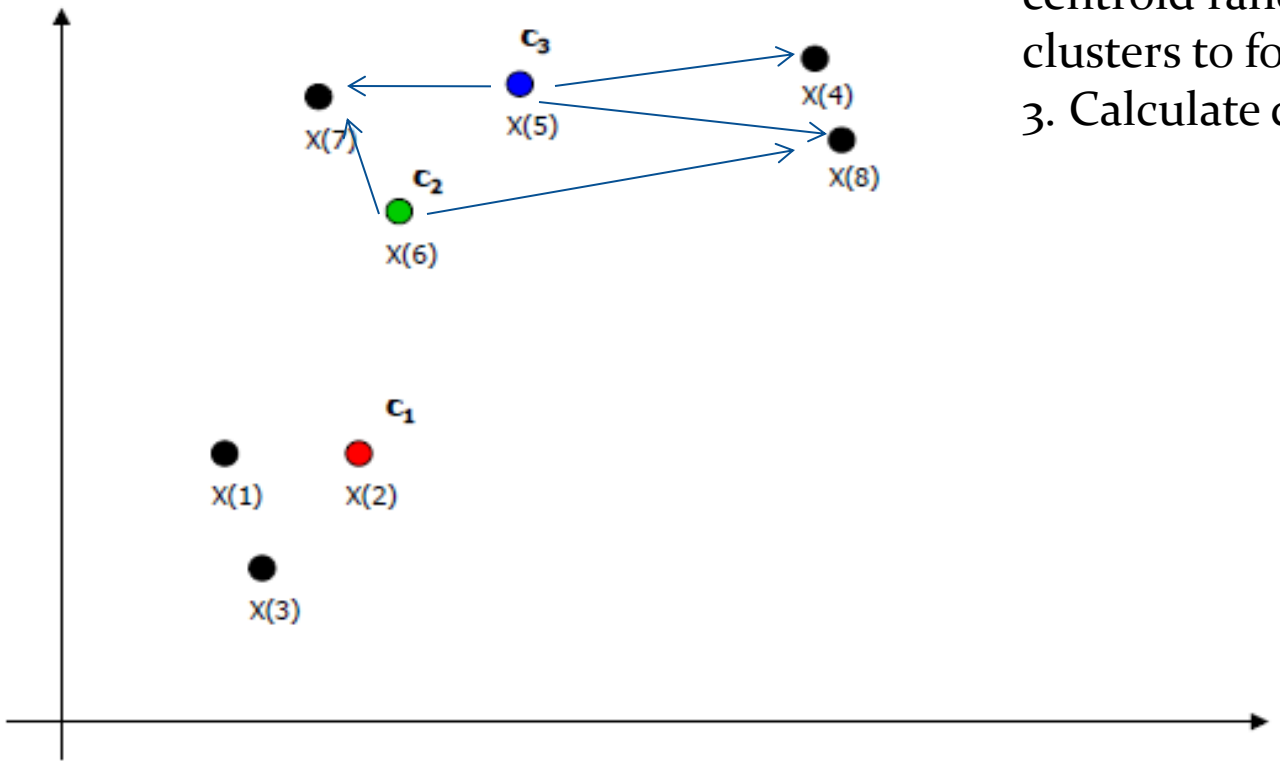
- Unsupervised learning is also called as clustering.
- What data?
  - Unlabeled data
- Training?
  - No
- Output
  - Groups/clusters of data objects such that clusters in same group have high similarity.

Types

- Partitioning
  - Constructs k partitions; each partition represents a cluster (k<n)
- Hierarchical
  - Agglomerative (bottom up) or divisive (top down)
  - Merge to form the number of clusters or split
- Density-based
  - Continue building a cluster as long as density (no. of objects) in neighborhood exceeds some threshold.
- Grid-based
  - A grid structure is maintained where the object space is quantised.

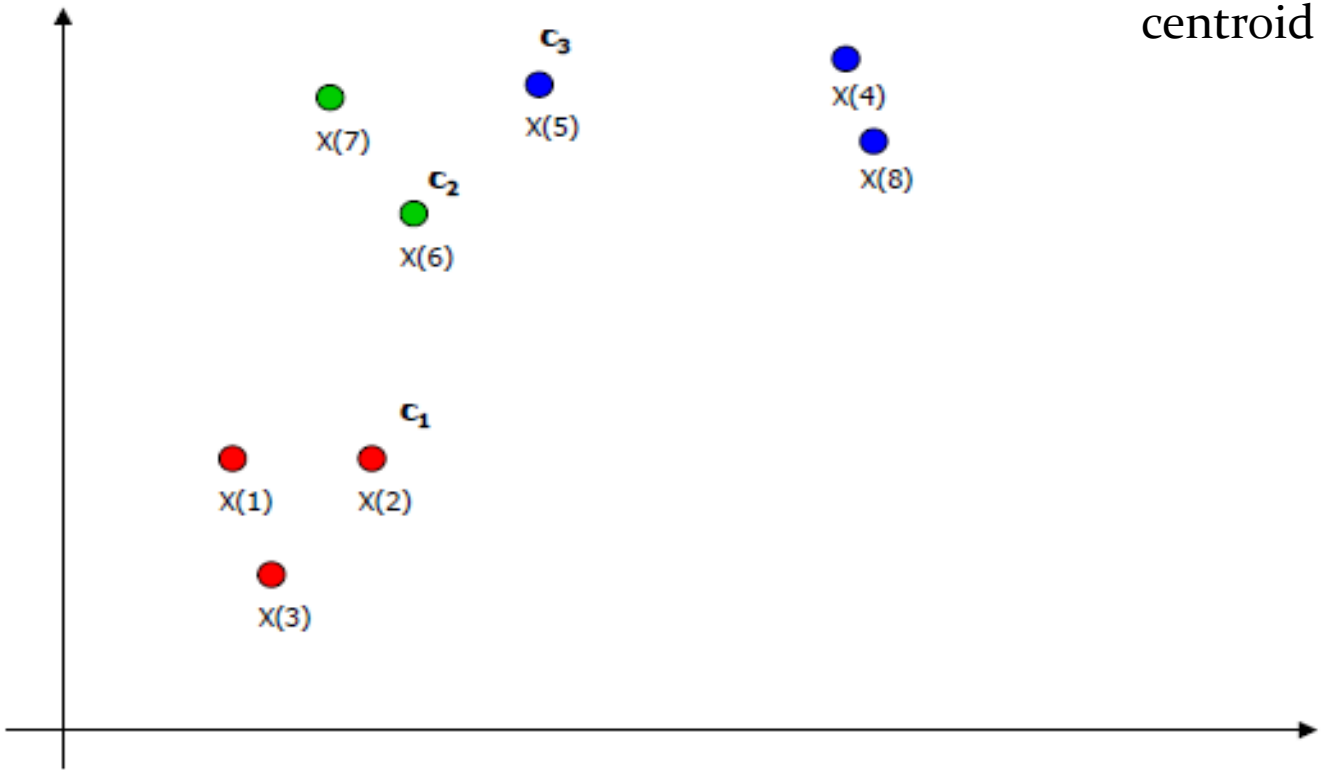
K-means demo

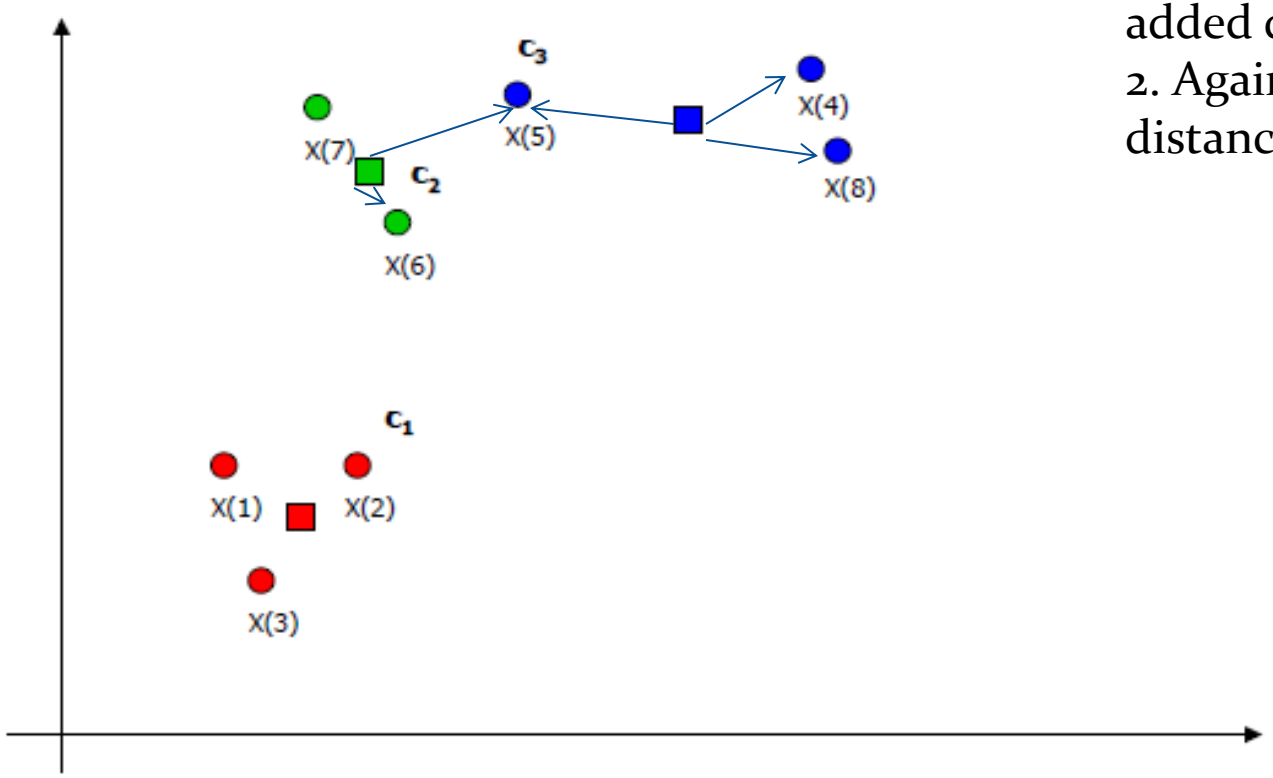




1. Take input of no. of clusters to form
2. Select data points as cluster centroid randomly = no. of clusters to form
3. Calculate distance

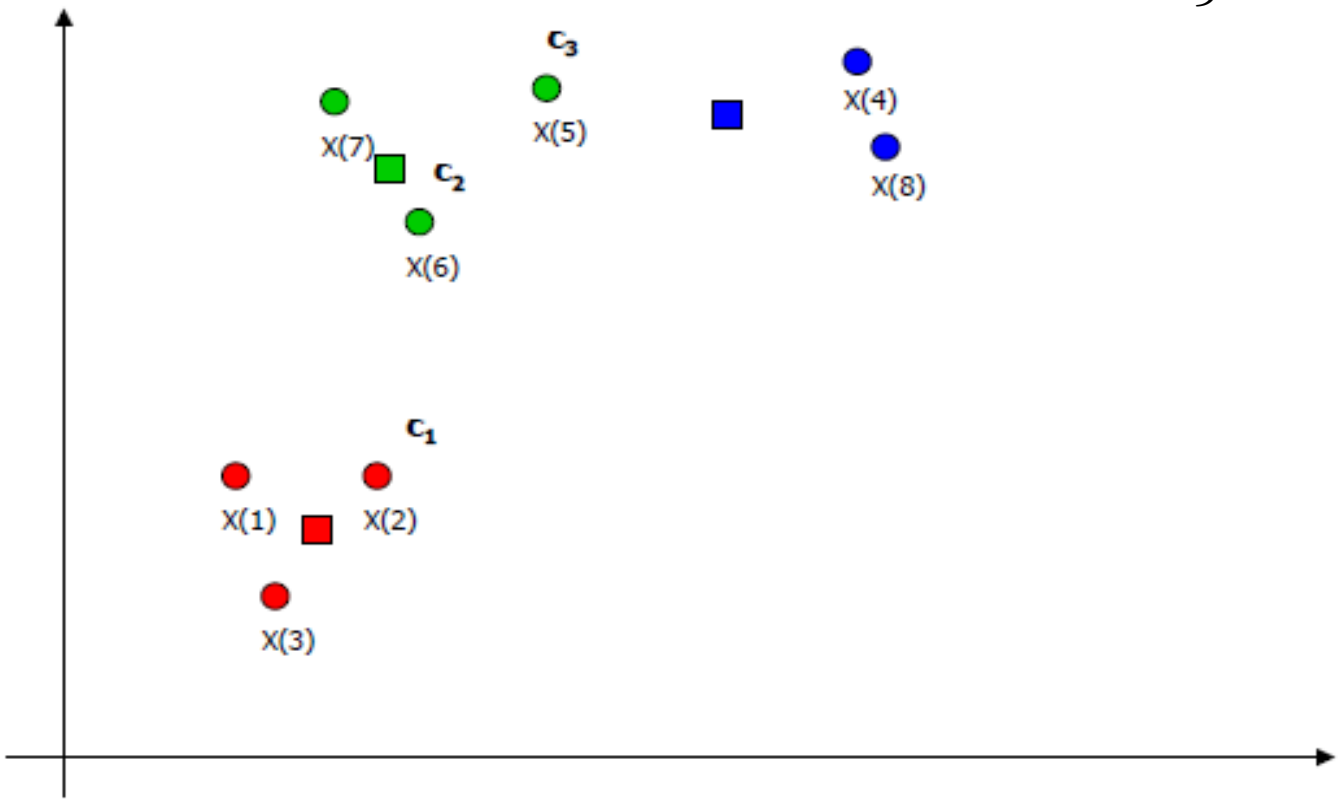
Assign to cluster with shortest distance from centroid

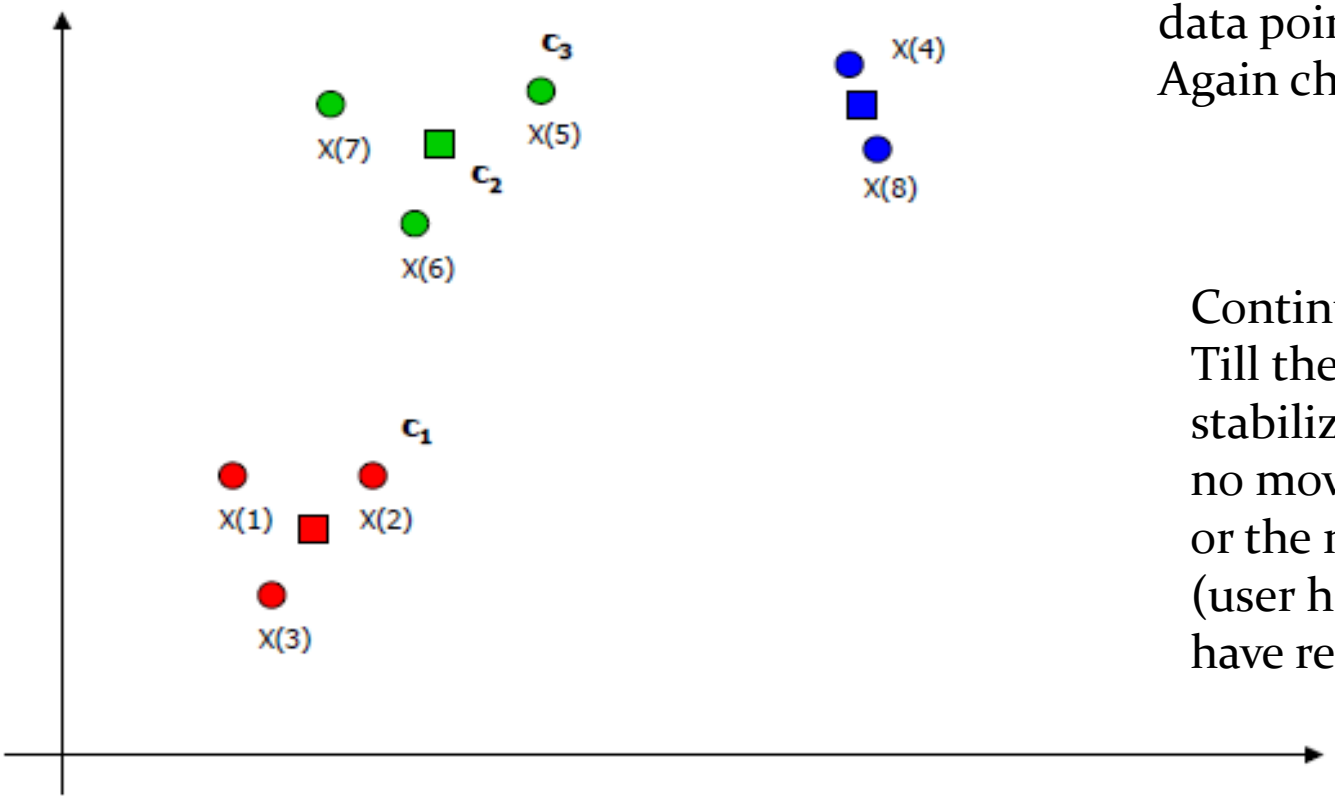




1. Recompute centroids (mean of the added data points)
2. Again check distance

Notice the change in X5… Cluster changed

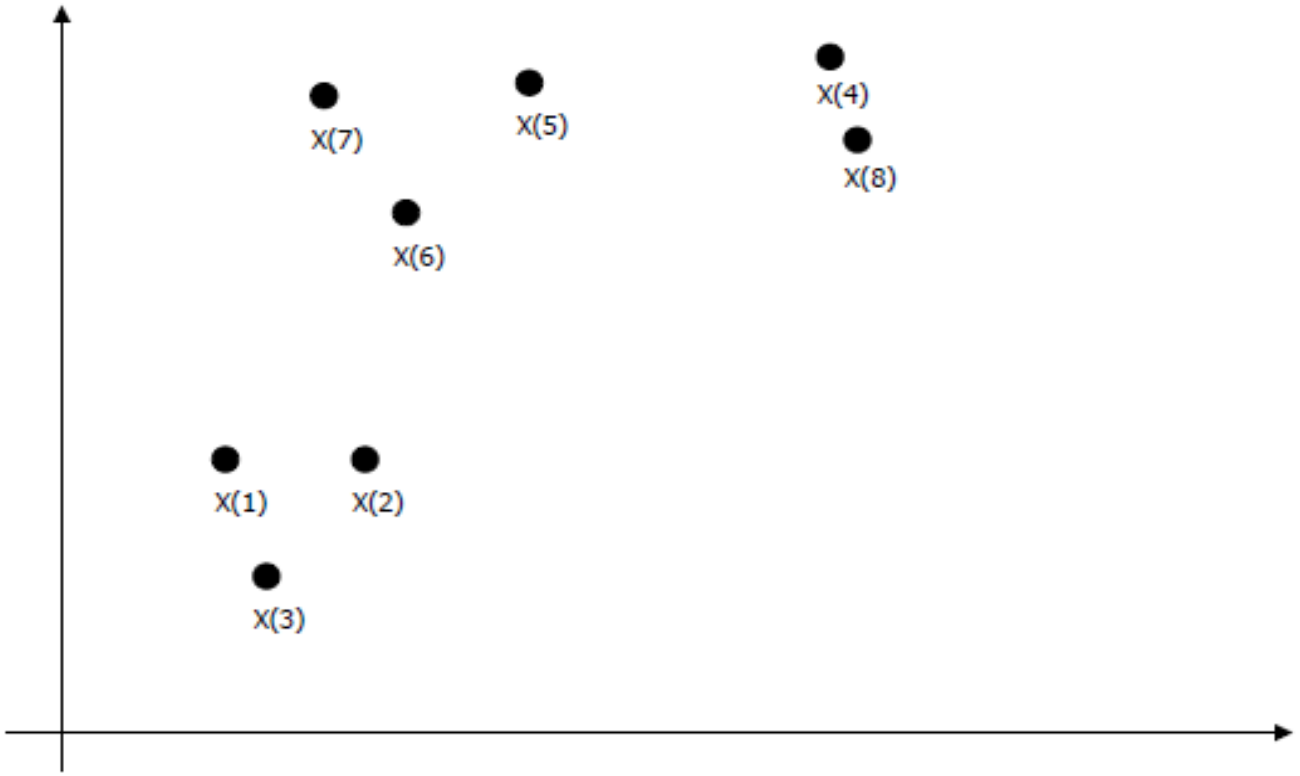


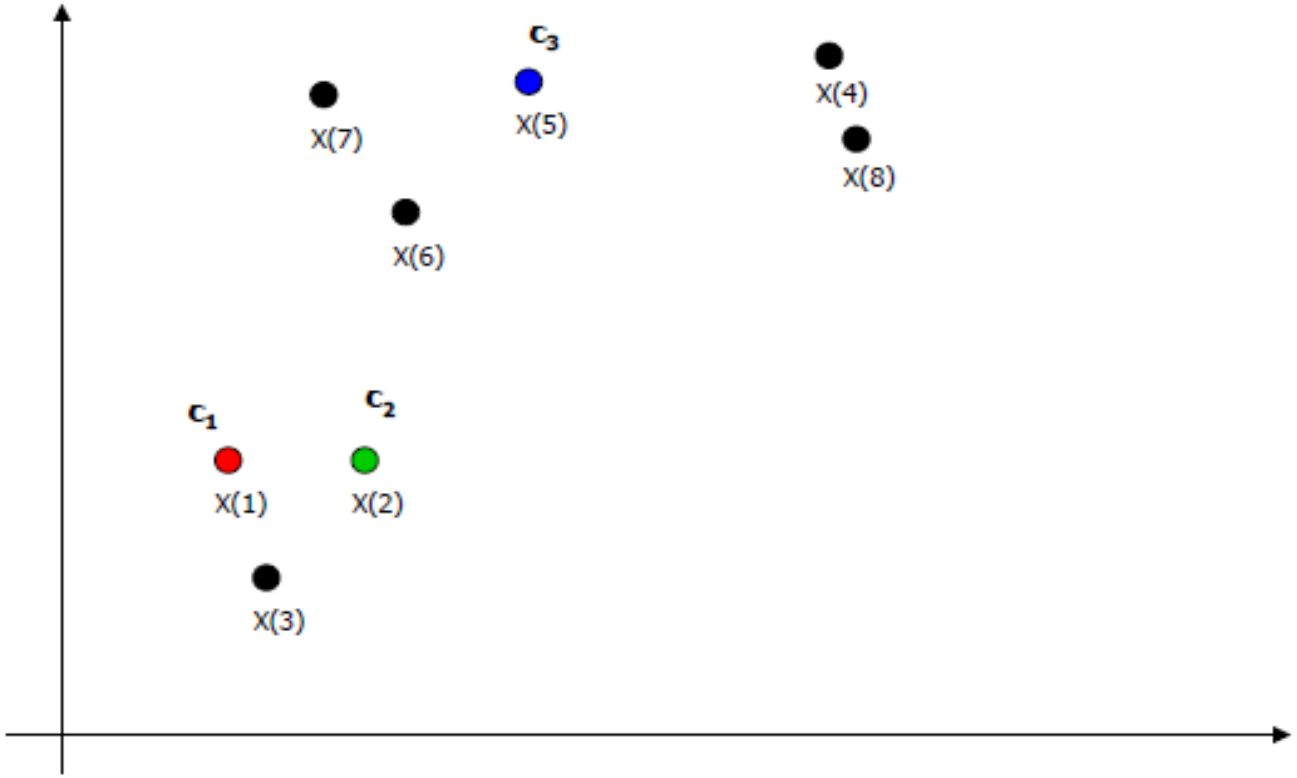


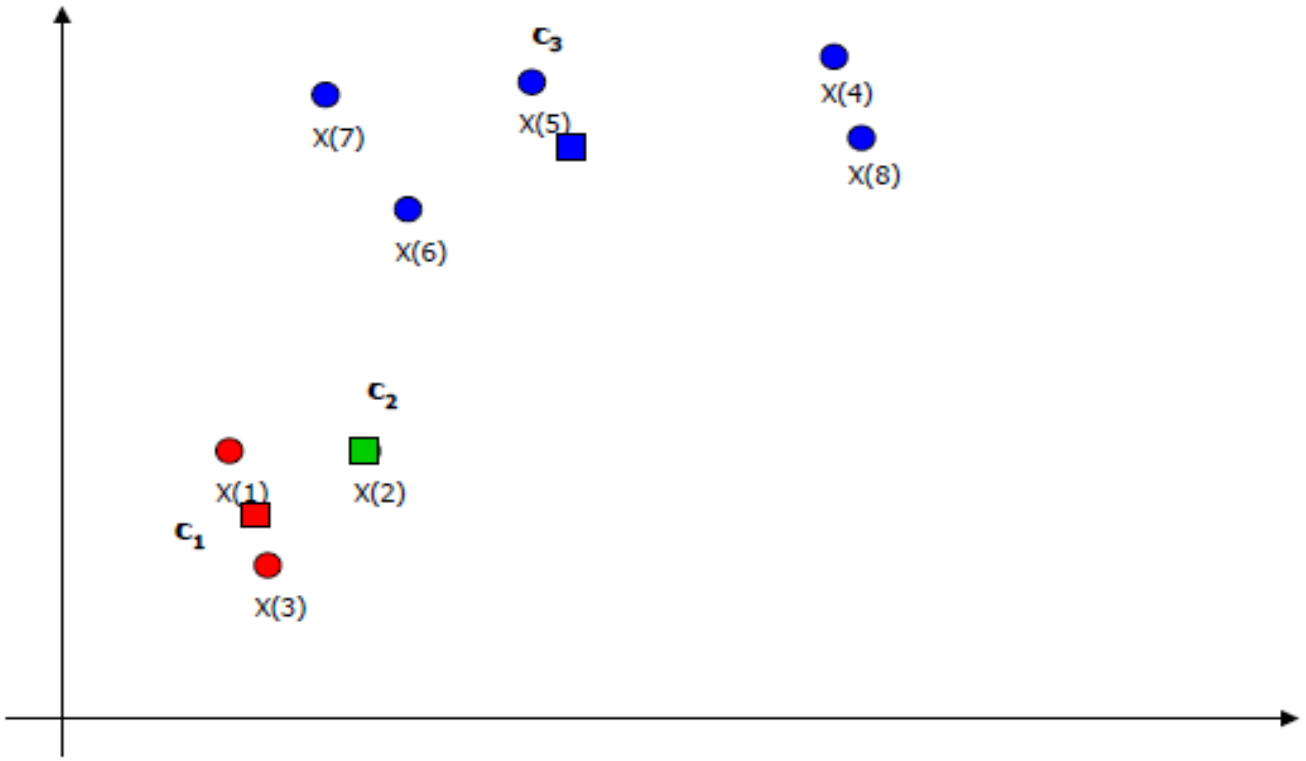
Recompute centroids (mean of the added data points)
Again check distance

Continue till when????
Till the cluster stabilizes : i.e. there is no movement of data or the no. of iterations (user has specified) have reached.

Another example k-means







The algorithm summarized

- Input: The data set and the number of clusters to form.
- Steps:

1. Select centroids / seed points = the number of clusters to form (k)
2. For every data item:
   - Calculate the distance between the centroids and the data item
   - Assign the data item to the cluster to which it is nearer.
3. Recompute the centroids = mean of the data points of that cluster
4. Continue with step 2 till the termination criteria is reached /satisfied

Advantages/ disadvantages

- Time complexity: O(ntk)
- Dependency on the initial centroids selected.
- Number of clusters to be formed must be known prior
- Euclidean distance is used....

$$dis\tan ce = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

References

- Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.
- Dunham, Margaret H. *Data mining: Introductory and advanced topics*. Pearson Education India, 2006.