Machine Learning

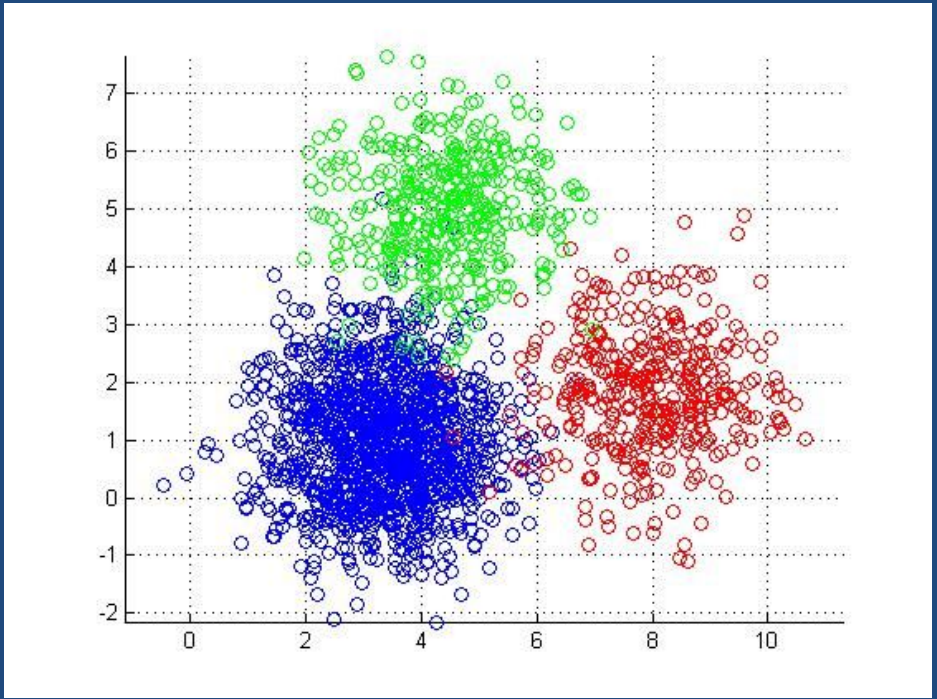
Session 7

31st March 2018

Rama Krishna Bhupathi
ramakris@gmail.com

Agenda

K Means Clustering



K Means Clustering

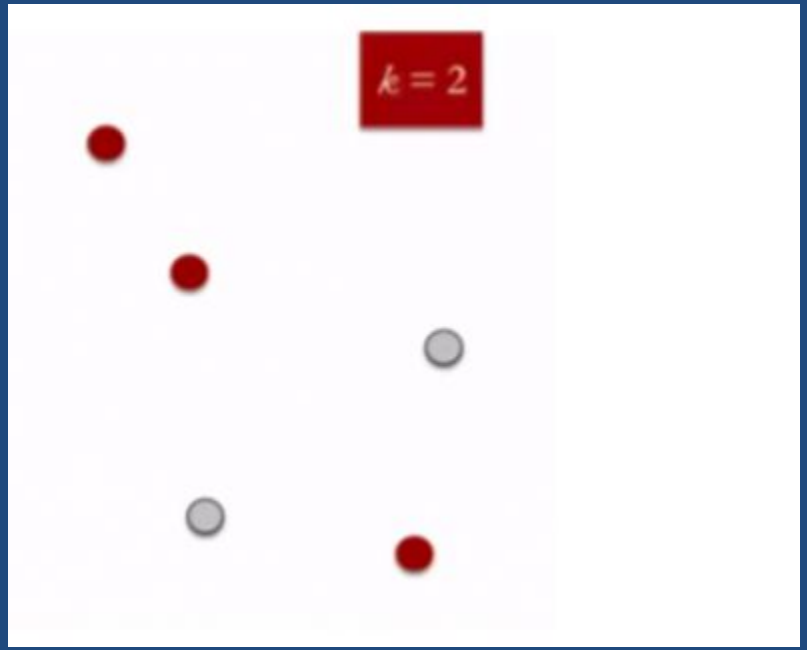
- Clustering is used to find groups of similar observations within a set of unlabeled data.
- It is unsupervised algorithm
- Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

K Means Clustering...examples

- In marketing, clustering is used to find segments of similar consumers
- Social networks can be clustered to identify communities and to suggest missing connections between people.
- In biology, clustering is used to find groups of genes with similar expression patterns.
- Recommendation systems sometimes employ clustering to identify products or media that might appeal to a user.

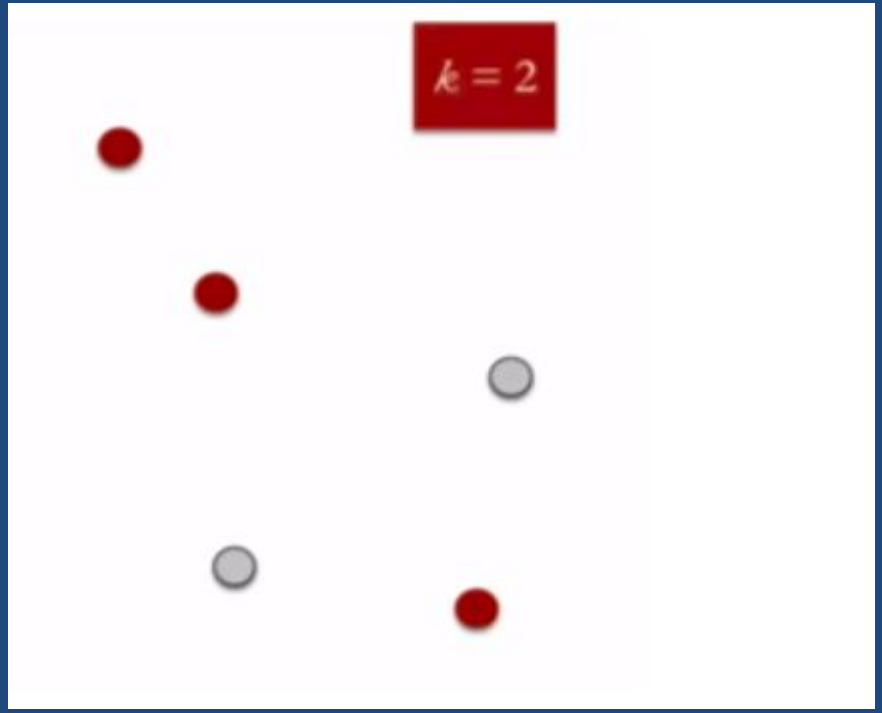
K-Means Clus. Algorithm

Step 1 :Specify the desired number of clusters K : Let us choose k=2 for these 5 data points in 2-D space.



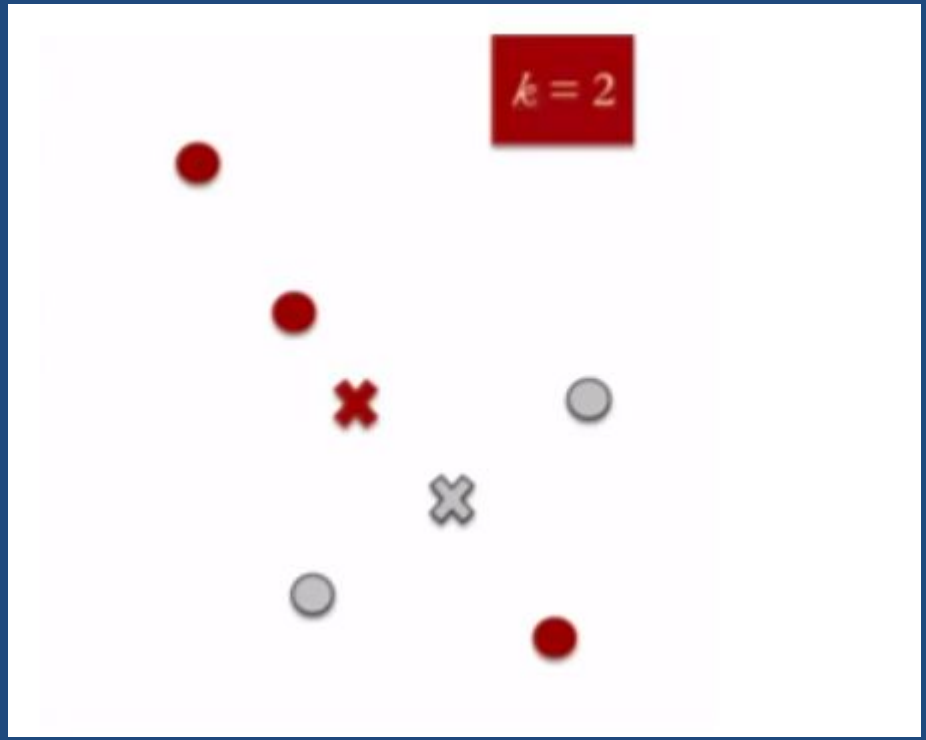
K-Means Clus. Algorithm

Step 2 :Randomly assign each data point to a cluster : Let's assign three points in cluster 1 shown using red color and two points in cluster 2 shown using grey color.



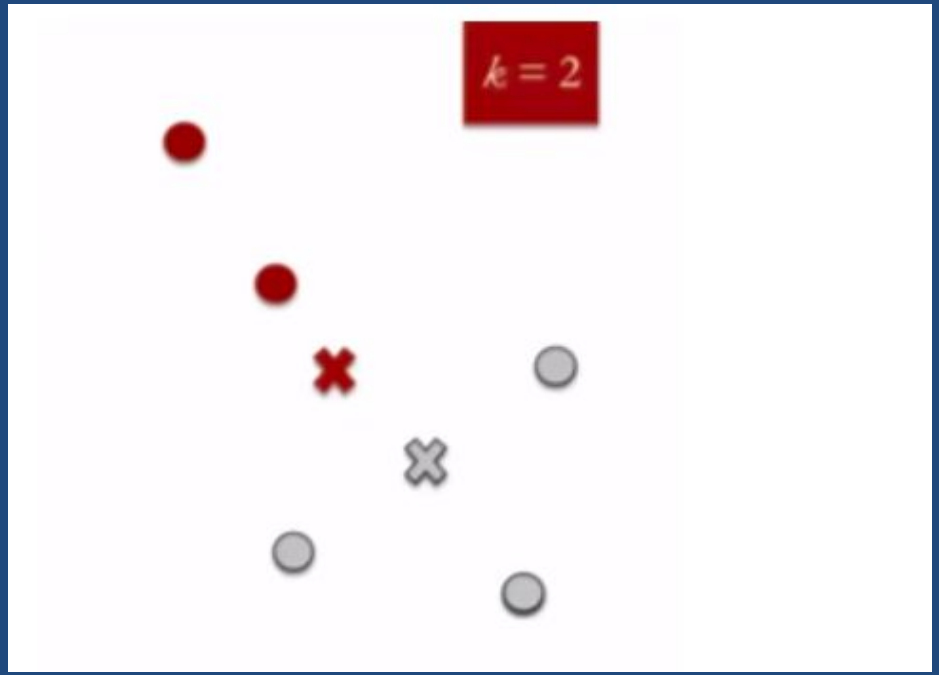
K-Means Clus. Algorithm

Step 3 : Compute cluster centroids : The centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.



K-Means Clus. Algorithm

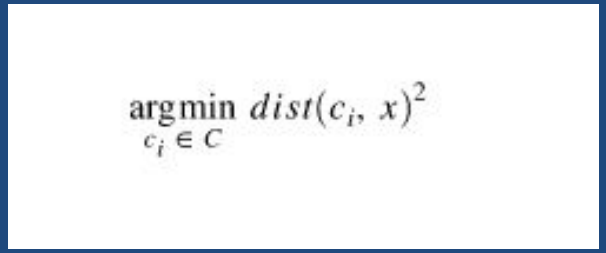
Step 4 : Re-assign each point to the closest cluster centroid : Note that only the data point at the bottom is assigned to the red cluster even though it's closer to the centroid of grey cluster. Thus, we assign that data point into grey cluster



K-Means Clus. Algorithm

But the how do you find the closest centroid?

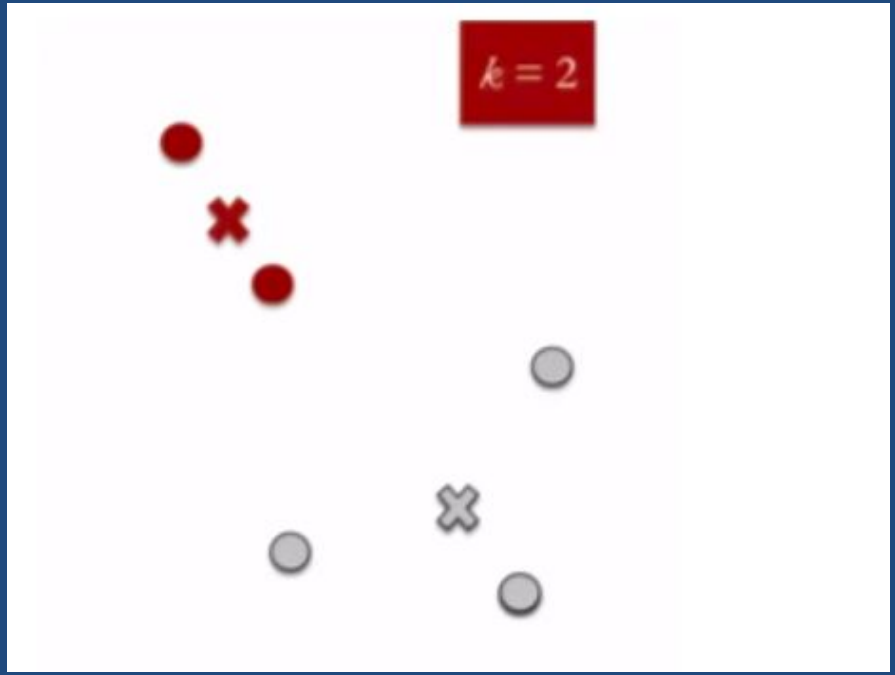
Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if ci is the collection of centroids in set C, then each data point x is assigned to a cluster based on



where dist(·) is the standard (L2) Euclidean distance. Let the set of data point assignments for each ith cluster centroid be Si.

K-Means Clus. Algorithm

Step 5 : Re-compute cluster centroids : Now, re-computing the centroids for both the clusters.



Cost/Loss Function

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} \left\| x_i - \mu_k \right\|^2$$