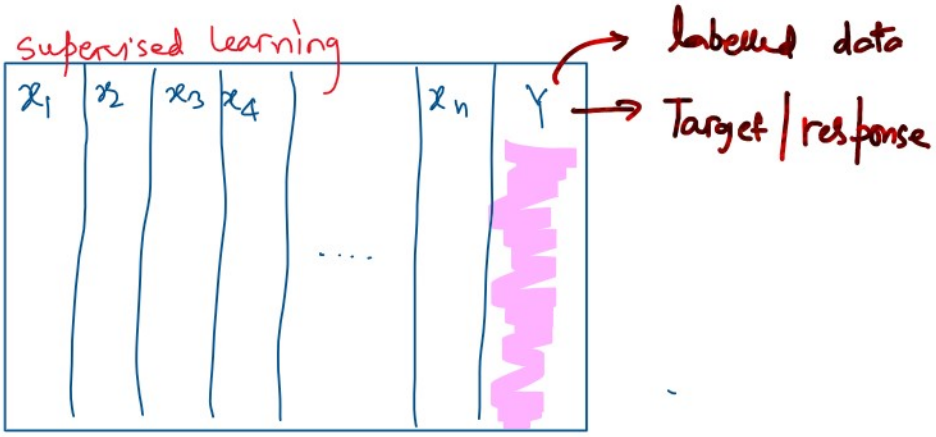
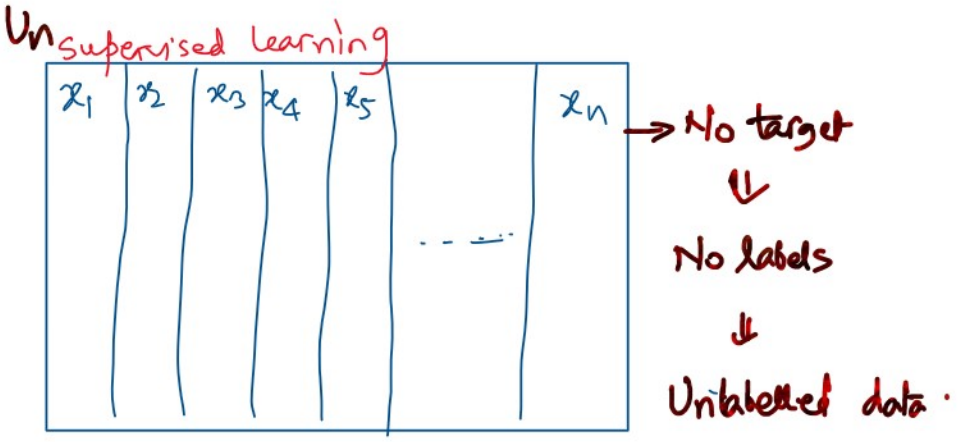
K-Means Clustering

19 January 2024 07:04

Unsupervised Learning

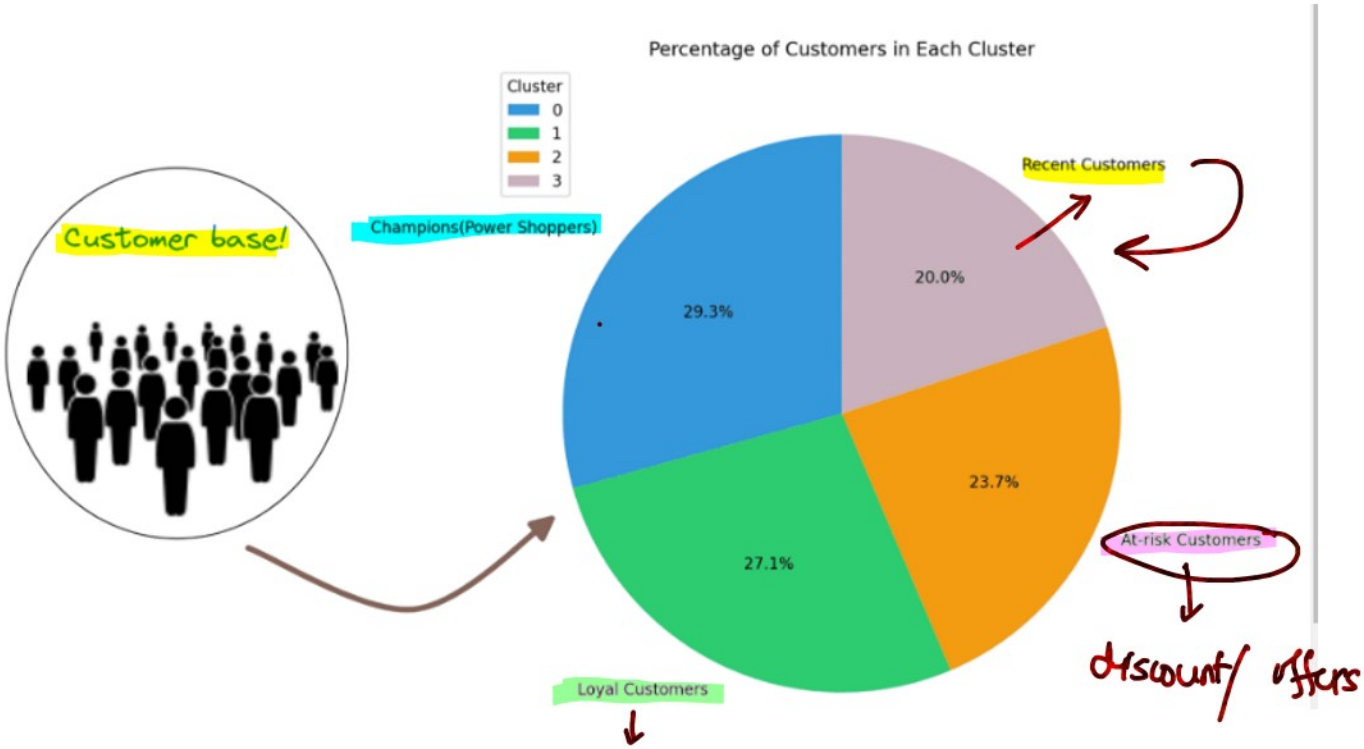




In unsupervised learning, there is no output variable to guide the learning process, and data is explored by algorithms to find patterns. Given that data has no labels, the algorithm identifies similarities on the data points and groups them into clusters

K-Means clustering

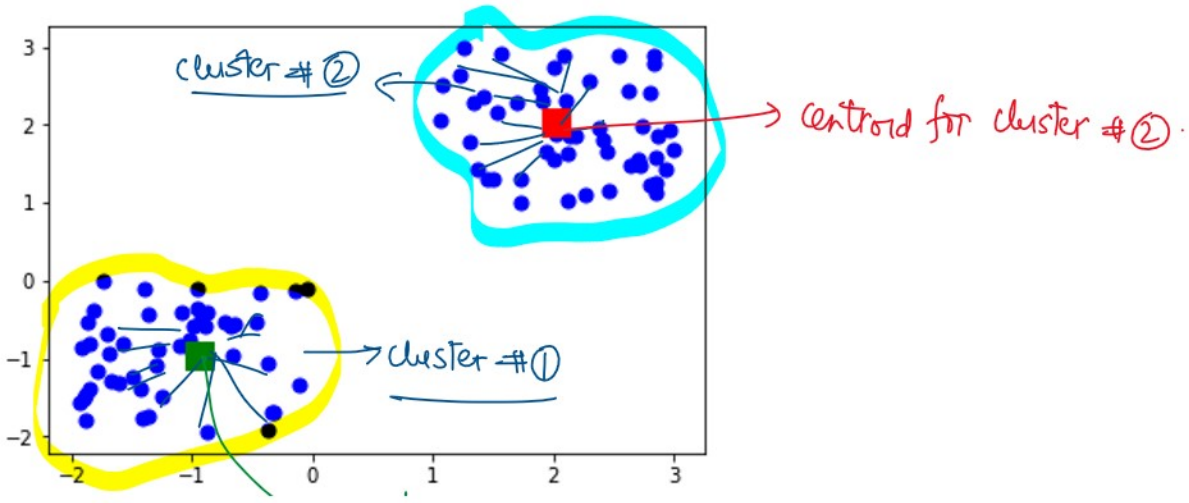
Demographic Data :

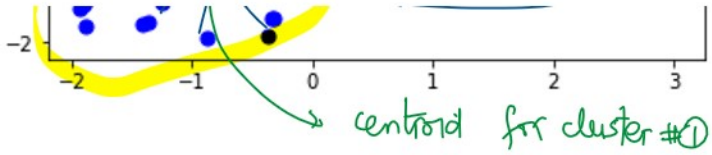


K-Means clustering

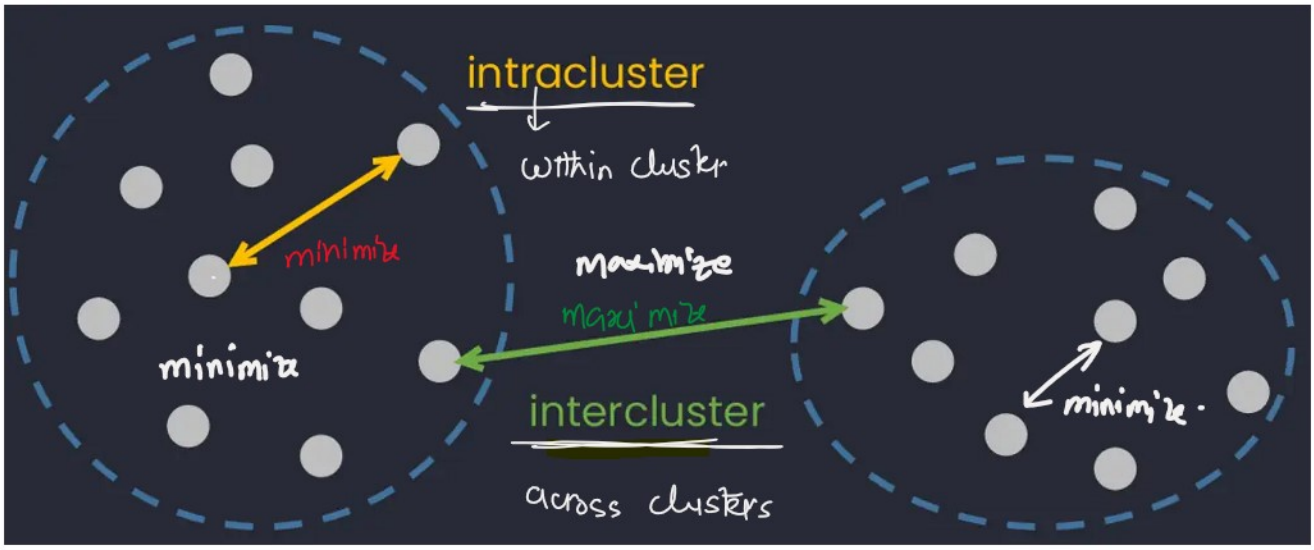
In K-Means clustering technique, each cluster is represented by its center (called a centroid), which corresponds to the arithmetic mean of data points assigned to the cluster.

- A centroid is a single data point within the cluster which represents the center of the cluster (the mean) and it might not necessarily be a member of the data set.



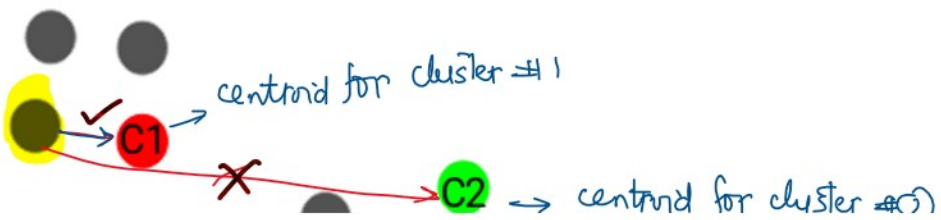


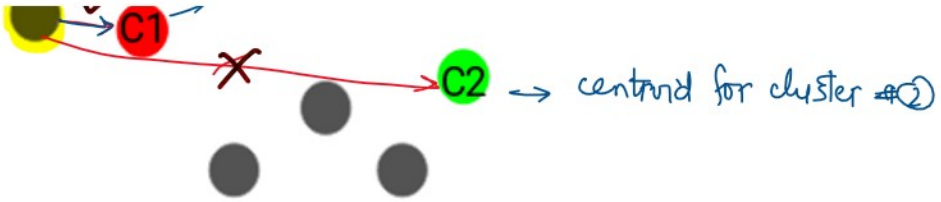
K-means clustering algorithm tries to minimize the distance of the points in a cluster with their centroid



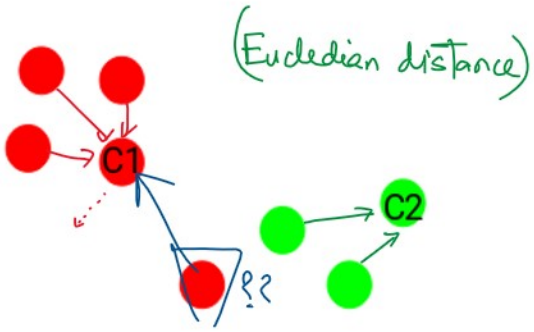
K-Means algorithm steps

1. Specify number of clusters 'K' — user input
   - generally K=3 or K=4.
2. Initialize centroids by first shuffling the dataset and then randomly selecting 'K' data points for the centroids without replacement.

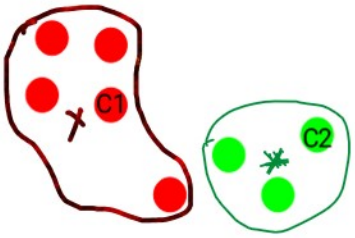




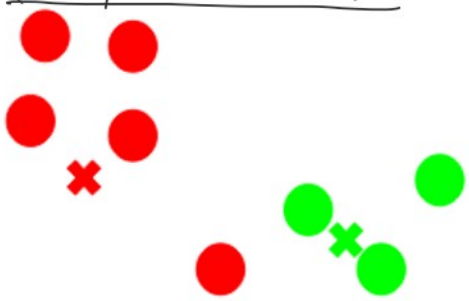
3. Assign all the points to closest cluster centroid

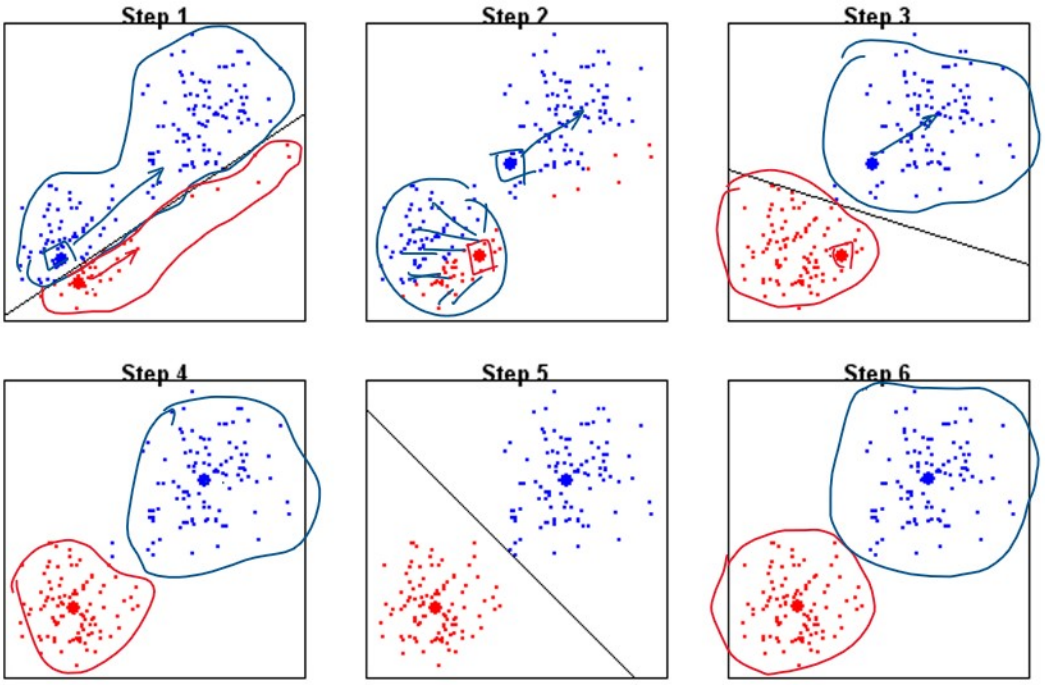


④ Recompute the centroids of newly formed clusters



(Recomputed centroids)

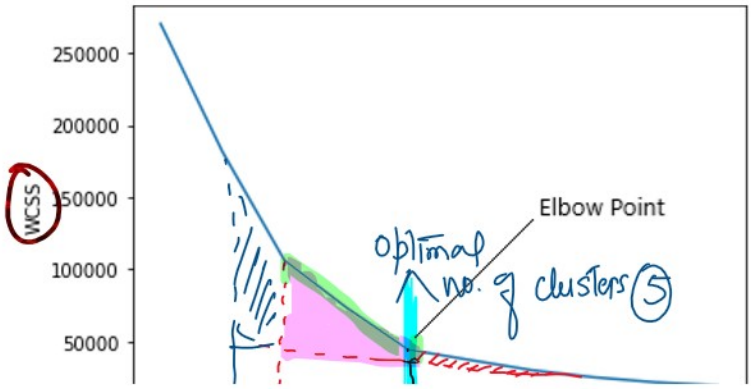


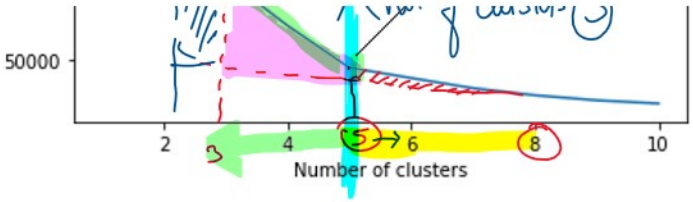


Real life applications of K-Means Clustering

1. Customer Profiling / Segmentation
2. Market segmentation
3. Documents classification
4. Geostatistics
5. Astronomy

Finding the value of 'k'





Testing different number of clusters and measuring the resulting sum of squared errors (SSE), and choosing the 'K' value at which an increase('K') will cause a very small decrease in error sum while a decrease('K') will sharply increase the error sum.

Stopping criteria for K-Means clustering

1. centroids of newly formed clusters do not change..
2. Points remain in the same cluster
3. Maximum no of iterations is reached..