UNSUPERVISED MACHINE LEARNING

Mohan M J

INTUITION

- Unsupervised learning is where you only have input data (X) and no corresponding output variables.
- The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.
- These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

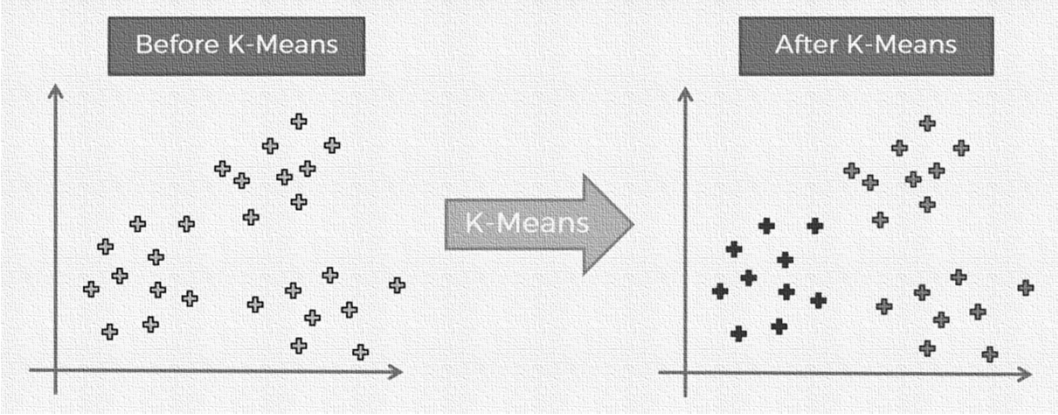
K MEANS CLUSTERING - INTUITION

- Type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups).
- The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*.
- The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided.
- Data points are clustered based on feature similarity.
- The results of the *K*-means clustering algorithm are:
  - The centroids of the *K* clusters, which can be used to label new data
  - Labels for the training data (each data point is assigned to a single cluster)

EXAMPLES

- Netflix want to recommend movie for users
- Wynk / Saavn want to play the songs according to the taste

K MEANS CLUSTERING - INTUITION



STEPS

STEP 1: Choose the number K of clusters

STEP 2: Select at random K points, the centroids (not necessarily from your dataset)

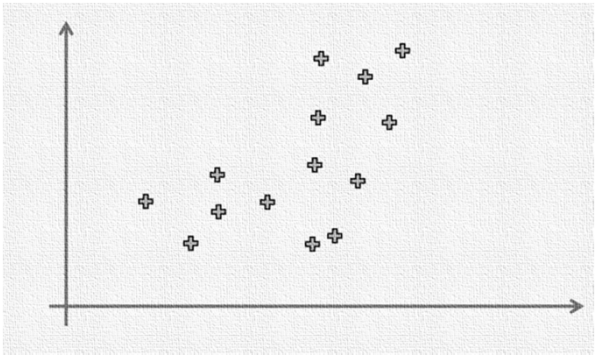
STEP 3: Assign each data point to the closest centroid ➡ That forms K clusters

STEP 4: Compute and place the new centroid of each cluster

STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.

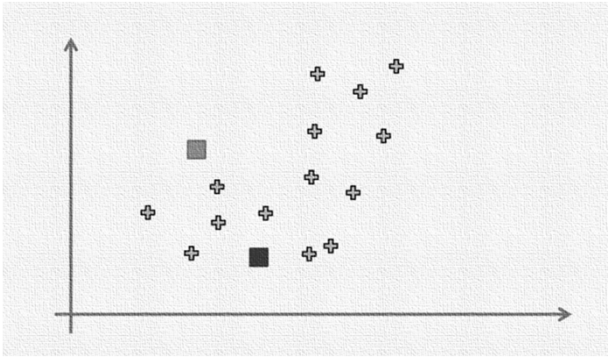
STEP 1: CHOOSE K

- Choose the number of clusters (K)
- Let's assume these are C1,C2,…,Ck and we can say that;
  C=C1,C2,…,Ck
  C is the set of all centroids.
- Here K =2



STEP 2: DATA ASSIGNMENT

- Select K centroids
- Centroids need not be from dataset

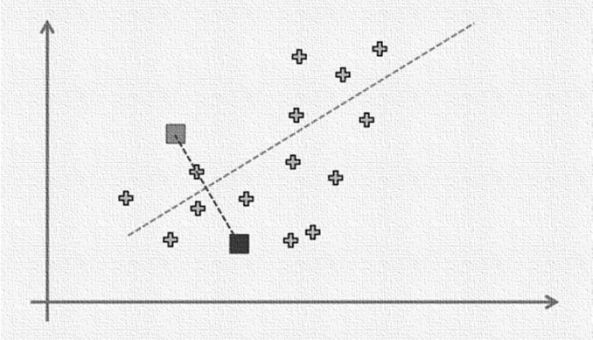


STEP 3: ASSIGN POINTS TO CENTROID

- Each data point is assigned to its nearest centroid, based on the squared Euclidean distance.
- If c_i is the collection of centroids in set C, then each data point x is assigned to a cluster based on

$$\underset{c_i \in C}{\operatorname{argmin}}\ dist(c_i,\ x)^2$$

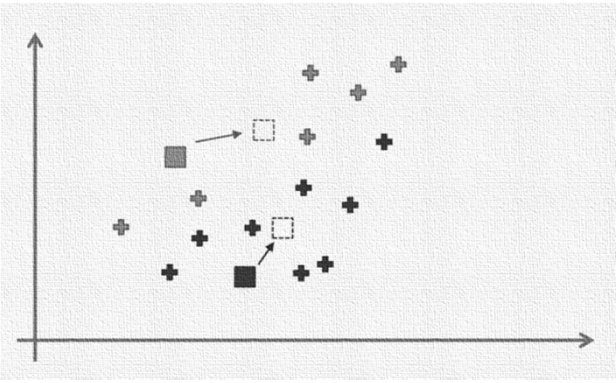
- where $dist(\cdot)$ is the standard (L_2) Euclidean distance. Let the set of data point assignments for each i^{th} cluster centroid be S_i.



STEP 4 : COMPUTE NEW CENTROID

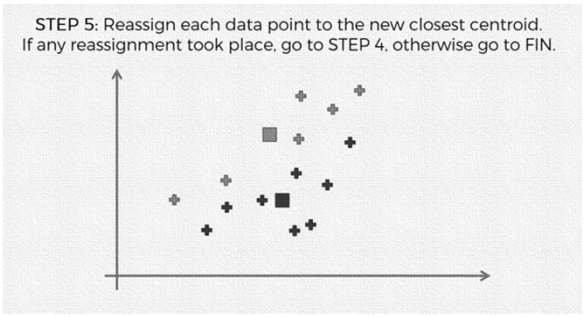
- The centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

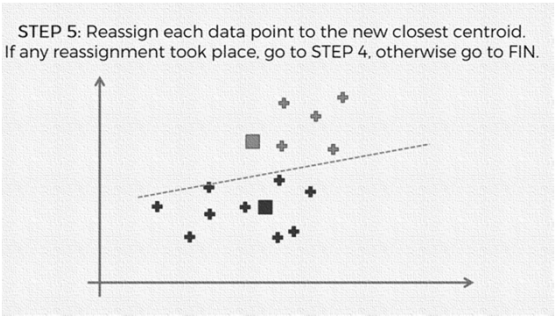
$$c_i = \frac{1}{|S_i|}\sum_{x_i \in S_i} x_i$$

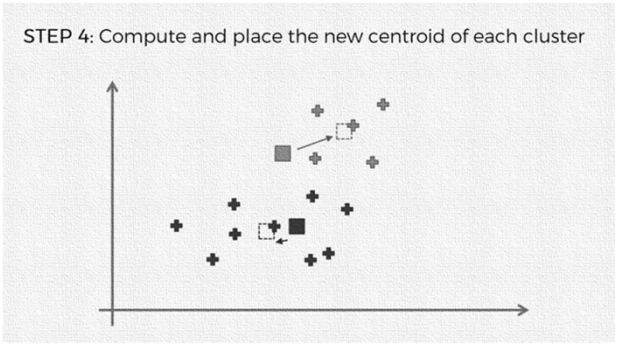


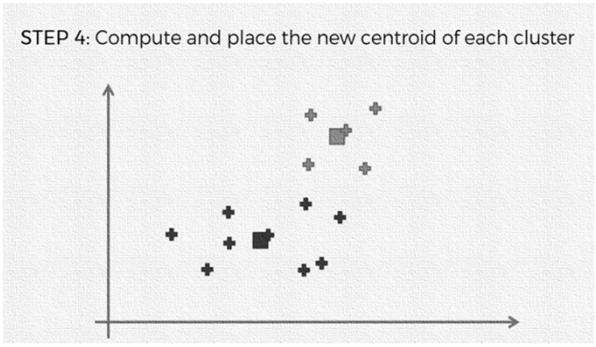
STEP 5: REASSIGN DATA POINTS

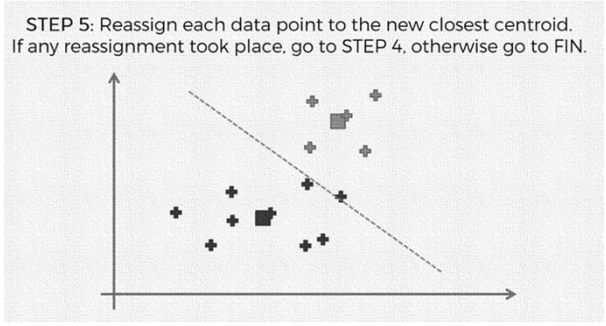
Each data point is assigned to its nearest centroid, based on the Euclidean distance



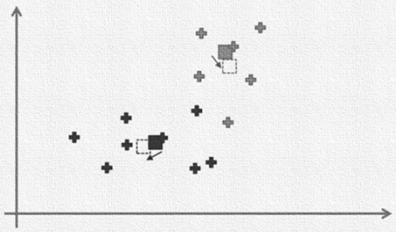




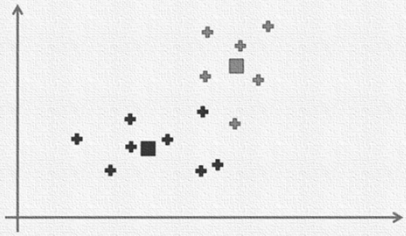




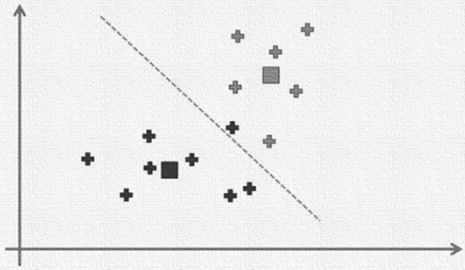
STEP 4: Compute and place the new centroid of each cluster



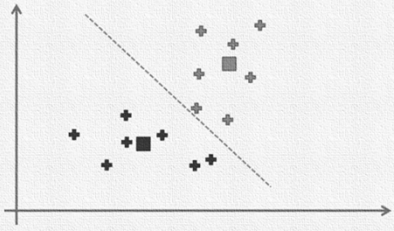
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



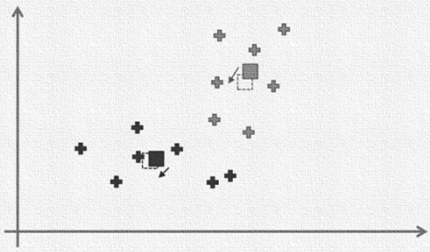
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



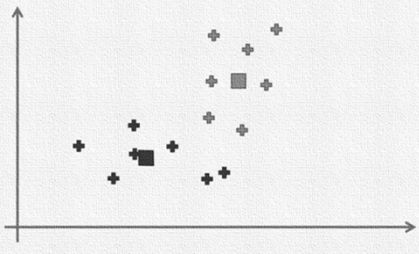
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



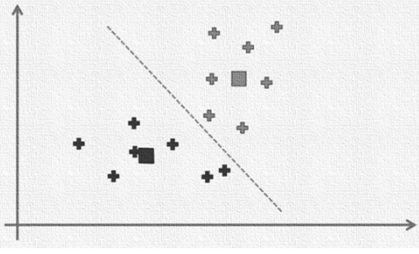
STEP 4: Compute and place the new centroid of each cluster



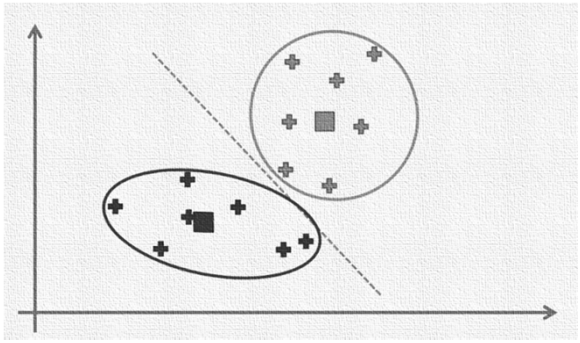
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.

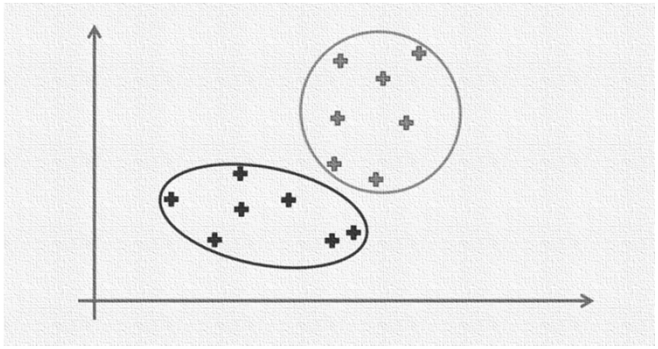


STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



K MEANS MODEL





THANK YOU