Certificate Course in Machine Learning using Python [6 Weeks]

Clustering Problem
What Is A Good Clustering?

A good clustering will produce high quality clusters in which:

- The intra-class (that is, intra-cluster) similarity is high.
- The inter-class similarity is low
- The measured quality of a clustering depends on both the document representation and the similarity measure used.

K-Means Program Customer Segments

As the name itself suggests, this algorithm will regroup n data points into K number of clusters. So given a large amount of data, we need to cluster this data into K clusters.

- Our goal is to categorise the customers to prepare better strategy.
- We don't know how many types of customers are present in the dataset (like rich, poor, who loves shopping, who doesn't love shopping etc).
- We don't know exactly how many number of clusters that need to chosen
- Let's assume for a moment that we are going to segment our data into 3 clusters (or may be 5 cluster).

Some of the mathematical terms involved in K-means clustering are centroids, euclidian distance.

- On a quick note centroid of a data is the average or mean of the data.
- **Euclidian distance** is the distance between two points in the coordinate plane. Given two points A(x1,y1) and B(x2,y2), the euclidian distance between these two points is :

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

- We can use other method also to measure the distance.

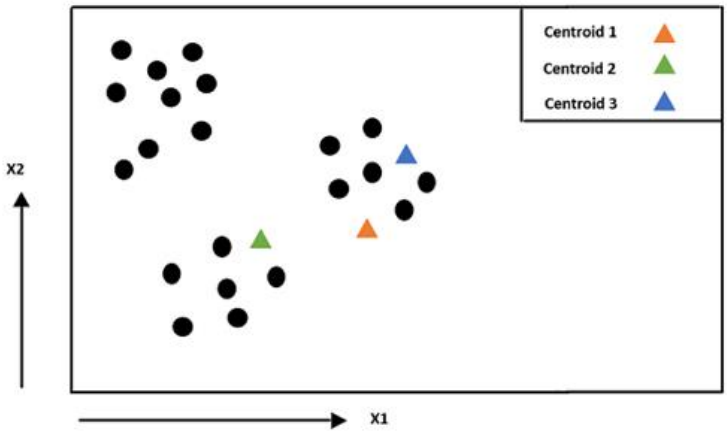
Algorithm:

Now let's start talking about the implementation steps. We shall be using either cluster centers or centroids words to describe the cluster centers.

Step 1:

- Randomly initialize the cluster centers of each cluster from the data points.
- Let's assume K=3, so we choose randomly 3 data points and assume them as centroids.

Random Initialization of Centroids



- Here three cluster centers or centroids with the green, orange, and blue triangle markers are chosen randomly.

Step 2:

2a.

- For each data point, compute the euclidian distance from all the centroids (3 in this case) and assign the cluster based on the minimal distance to all the centroids.

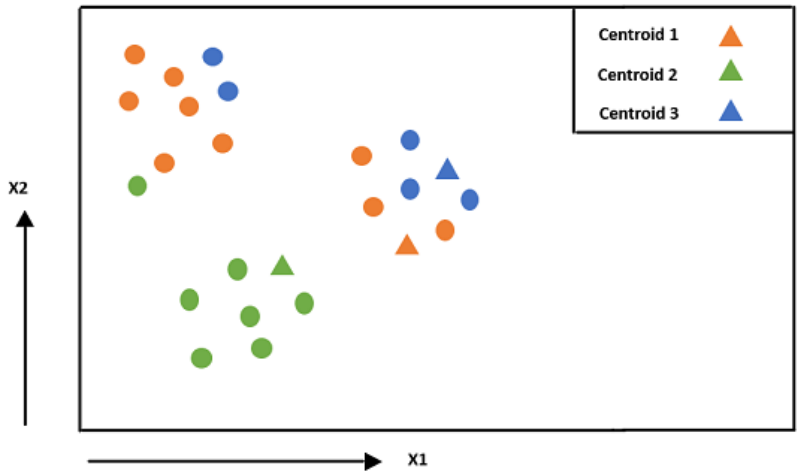
In our example, we need to take each black dot, compute its euclidian distance from all the centroids (green, orange and blue), and finally color the black dot to the color of the closest centroid.

2b.

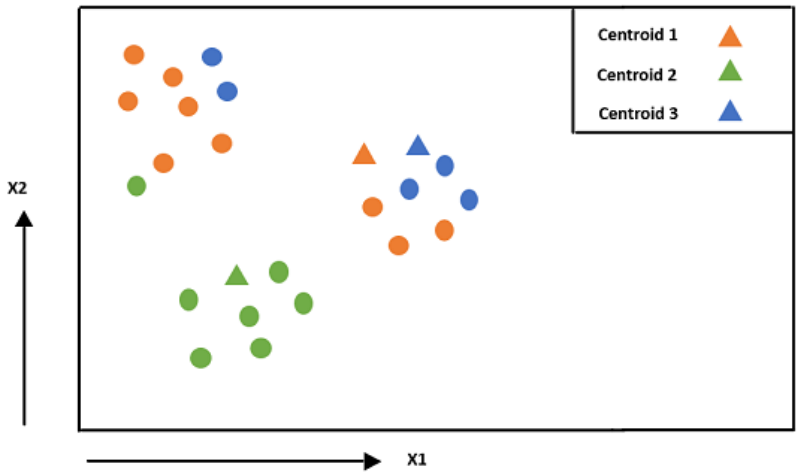
- Adjust the centroid of each cluster by taking the average of all the data points which belong to that cluster on the basis of the computations performed in step 2a.

In our example, as we have assigned all the data points to one of the clusters, we need to calculate the mean of all the individual clusters and move the centroid to calculated mean.

Repeat this process till clusters are well separated or convergence is achieved.

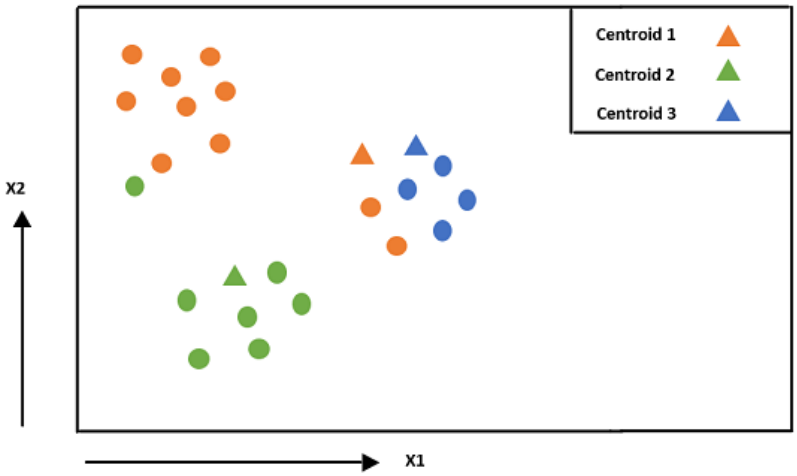


2.a Assign the centroids to each data point based on computed euclidian distances

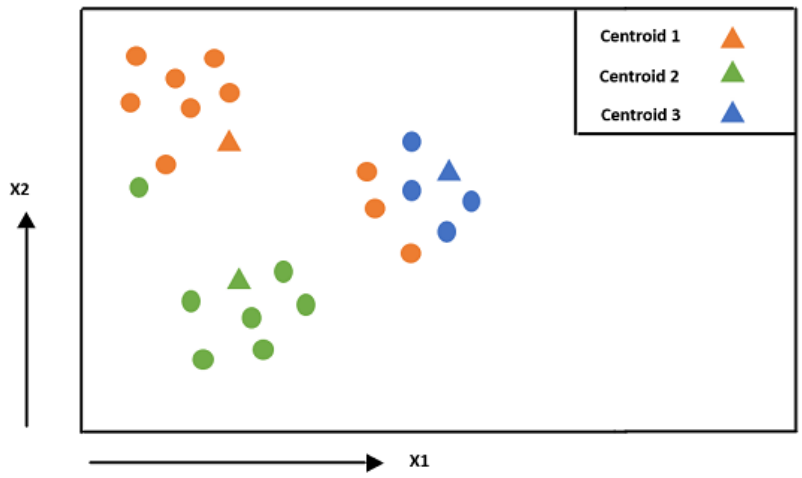


2b. Adjust the centroids by taking the average of all data points which belong to that cluster

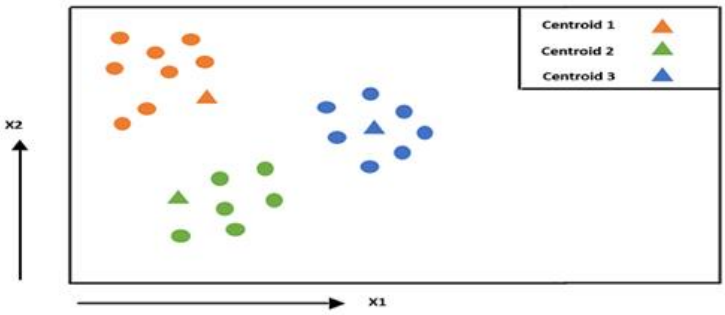
Repeat these steps until convergence is achieved:



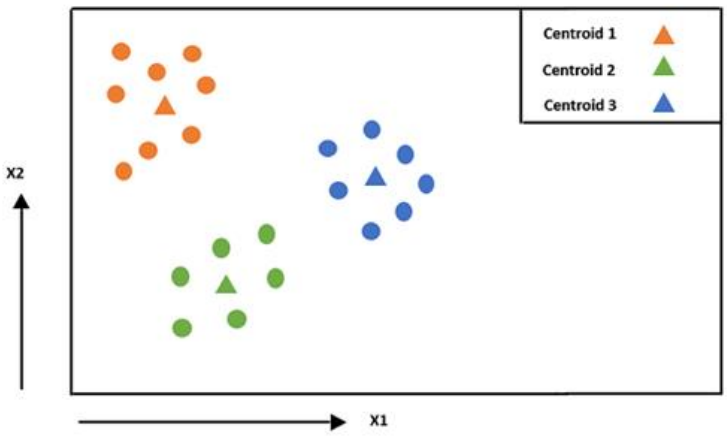
Step 2.a



Step 2.b



Step 2.a



Step 2.b

Previous Next

PREVIOUS ACTIVITY

◄ Cross Validation: Python code

Jump to...

NEXT ACTIVITY

Feature Importance & Correlation Matrix ►

Stay in touch

Contact Us

- http://nielit.gov.in/gorakhpur/
- abhinav@nielit.gov.in or ajay.verma@nielit.gov.in