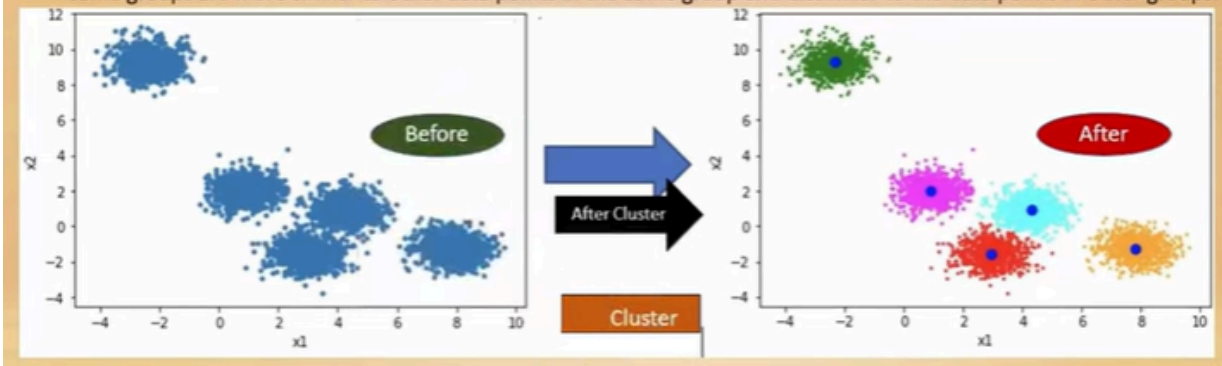
Unsupervised Learning: Clustering

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 and dissimilar to the data points in other groups.



Applications of Clustering: Real-World Scenarios

Clustering is a widely used technique in the industry. It is actually being used in almost every domain, ranging from banking to recommendation engines, document clustering to image segmentation.

- Customer Segmentation
- Document Clustering
- Image Segmentation
- Recommendation Engines

Perform K-Means Cluster

Tasks:

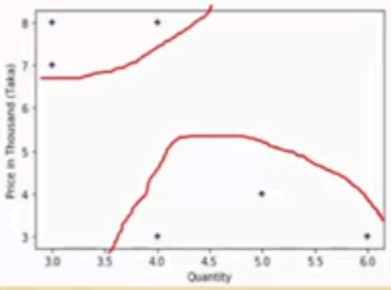
The k-means cluster algorithm mainly performs two important tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center (also called centroid). Those data points which are near to the particular k-center, create a cluster.

Let's see an EXAMPLE

Products	Quantity	Price(K)
FaceWash	3	7
Cream	5	4
Shoes	4	3
Bags	4	8
Jacket	6	3
Shirt	3	8

```
In [6]: plt.xlabel('Quantity')
        plt.ylabel('Price in Thousand (Taka)')
        plt.scatter(dataframe['Quantity'], dataframe['Price(K)'],marker='+',color='blue')
Out[6]: <matplotlib.collections.PathCollection at 0x2a9e524a048>
```



Products	Quantity	Price(K)
FaceWash	3	7
Cream	5	4
Shoes	4	3
Bags	4	8
Jacket	6	3
Shirt	3	8

$C1 = (3,7)$ and $C_2 = (5,4)$

★ For First data point (3,7) , Facewash :

Distance from $C_1 = 0$ ★ (C1)

Distance from $C_2 = \sqrt{(5-3)^2 + (4-7)^2}$

$= \sqrt{4+9}$

$= 4.24$

Products	Quantity	Price(K)
FaceWash	3	7
Cream	5	4
Shoes	4	3
Bags	4	8
Jacket	6	3
Shirt	3	8

$C1 = (3,7)$ and $C_2 = (5,4)$

★ For First data point (3,7) , Facewash :

Distance from $C_1 = 0$ ★ (C1)

Distance from $C_2 = \sqrt{(5-3)^2 + (4-7)^2}$

$= \sqrt{4+9}$

$= 4.24$

★ For Second data point (5,4) , Cream :

Distance from $C_1 = \sqrt{(5-3)^2 + (4-7)^2}$

$= \sqrt{4+9}$

$= \sqrt{13} = 3.60$

Distance from $C_2 = 0$ ★ (C2)

Products	Quantity	Price(K)
FaceWash	3	7
Cream	5	4
Shoes	4	3
Bags	4	8
Jacket	6	3
Shirt	3	8

$c_1 = (3,7)$ and $c_2 = (5,4)$

* For First data point (3,7) Facewash :

Distance from $c_1 = 0$ * (c1)

Distance from $c_2 = \sqrt{(5-3)^2 + (4-7)^2}$

$= \sqrt{9+9}$

$= 4.24$

* For Second data point (5,4) Cream :

Distance from $c_1 = \sqrt{(5-3)^2 + (4-7)^2}$

$= \sqrt{4+9}$

$= \sqrt{13} = 3.60$

Distance from $c_2 = 0$ * (c2)

* For third data point (4,3) shoes :

Distance from $c_1 = \sqrt{(4-3)^2 + (3-7)^2}$

$= \sqrt{1+16}$

$= 4.123$

Distance from $c_2 = \sqrt{(4-5)^2 + (3-4)^2}$

$= \sqrt{1+1}$

$= 1.41$ * (c2)

So new centroid $= \left(\frac{5+4}{2}, \frac{4+3}{2}\right)$

$c_2 = (4.5, 3.5)$

Products	Quantity	Price(K)
FaceWash	3	7
Cream	5	4
Shoes	4	3
Bags	4	8
Jacket	6	3
Shirt	3	8

$c_1 = (3,7)$ and $c_2 (4.5, 3.5)$

For 4th data point (4,8) bags :

Distance from $c_1 = \sqrt{(4-3)^2 + (8-7)^2}$

$= \sqrt{1+1}$

$= 1.41$ * (c1)

Distance from $c_2 = \sqrt{(4-4.5)^2 + (8-3.5)^2}$

$= 0.25 + 20.25$

$= 20.50$

∴ New centroid $= \left(\frac{3+4}{2}, \frac{7+8}{2}\right)$

$c_1 = (3.5, 7.5)$

Products	Quantity	Price(K)
FaceWash	3	7
Cream	5	4
Shoes	4	3
Bags	4	8
Jacket	6	3
Shirt	3	8

$c_1 = (3.5, 7.5)$ and $c_2 = (5, 3.33)$

For 6th data point (3,8) shirt :

Distance from $c_1 = \sqrt{(3-3.5)^2 + (8-7.5)^2}$

$= \sqrt{.25 + .25}$

$= 0.70$ * c_1

Distance from $c_2 = \sqrt{(3-5)^2 + (8-3.33)^2}$

$= \sqrt{4 + 2.16}$

$= 2.48$

New centroid $= \left(\frac{3+4+3}{3}, \frac{7+8+8}{3}\right)$

$c_1 = (3.33, 7.67)$

$c_2 = (5, 3.33)$

- The working of the K-Means algorithm is explained in the below steps:
  - Step-1: Select the number K to decide the number of clusters.
  - Step-2: Select random K points or centroids.
  - Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.
  - Step-4: Calculate the variance and place a new centroid of each cluster.
  - Step-5: Repeat the third steps, which means reassign each data point to the new closest centroid of each cluster.
  - Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.
  - Step-7: The model is ready.

Elbow Method for optimal value of k in