Experiment – 5 Implementation of Clustering algorithm(Kmeans)

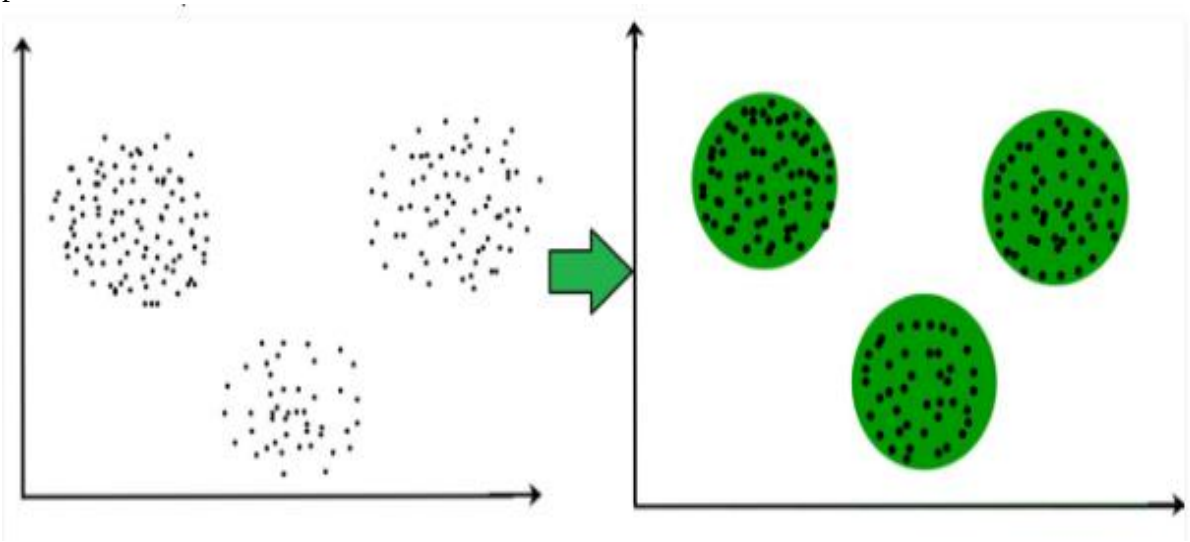
Name: **Shaikh Shadab** **Rollno** **:** **17DCO74**
Class : **TE.CO** **Batch** **:** **B3**

#Theory

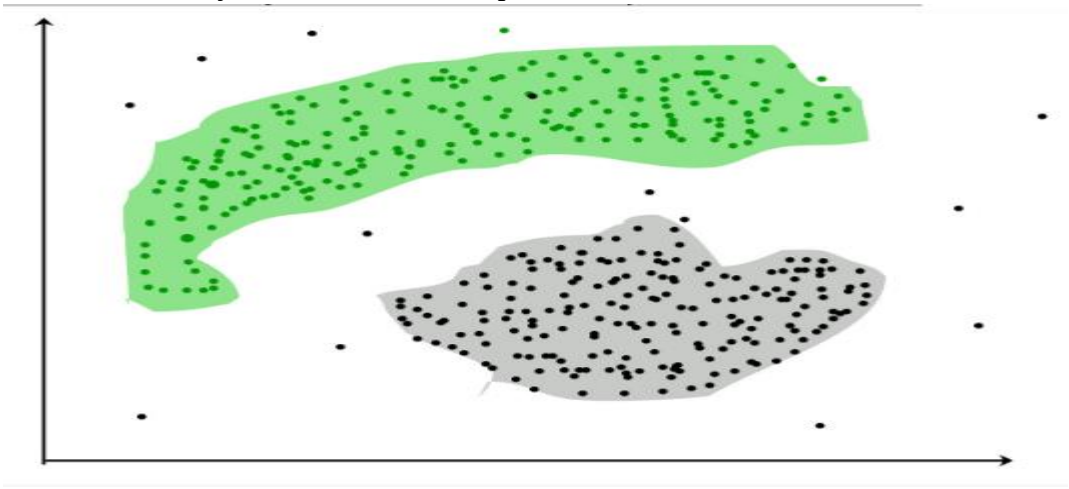
It is basically a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

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. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

For ex– The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture



It is not necessary for clusters to be a spherical. Such as :

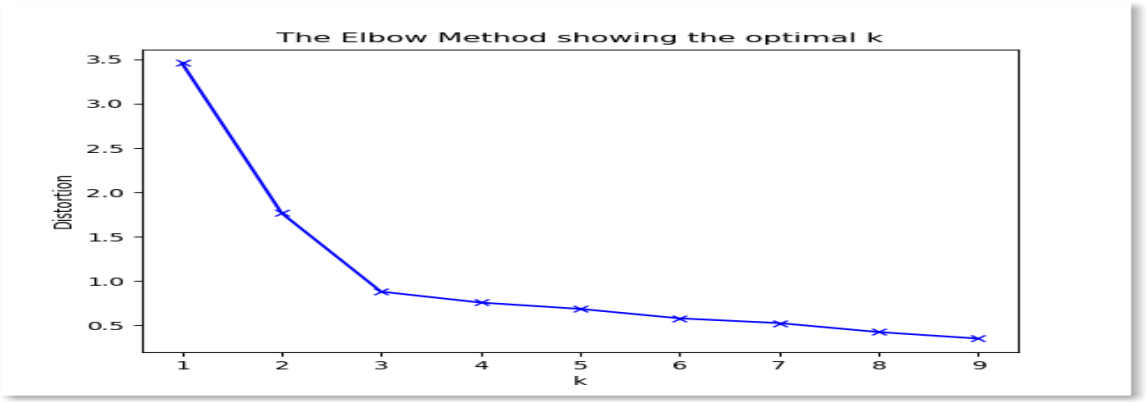


- **K-means clustering and elbow method**

  It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partition n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster .we need to always specify the number of clusters that we need the data set clustered into. So the most easiest way of doing this is the use of Elbow method.

  Elbow method to find optimal value for k (#clusters).

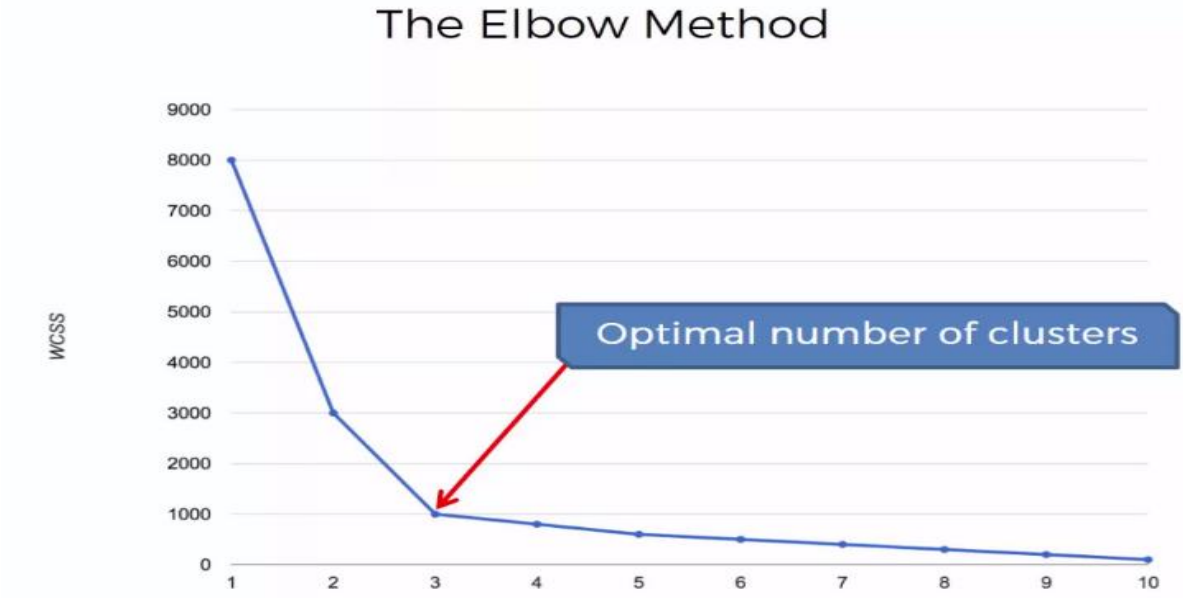
  - values for K on the horizontal axis
  - The distortion on the Y axis (the values calculated with the cost function).
    This results in:



  - When K increases, the centroids are closer to the clusters centroids.
  - The improvements will decline, at some point rapidly, creating the elbow shape.
  - That point is the optimal value for K. In the image above, K=3.

  Most of the time, Elbow method is used with either squared error(sse) or within cluster sum of square(wcss). Just like the name suggests, wcss is the summation of the each clusters distance between that specific clusters each points against the cluster centroid. Look at the below image to understand, how to calculate the wcss value for 3 cluster data set,

  So, if we plot the wcss value against the number of clusters that we tried to get that wcss value, normally we end up getting a graph similar



#Algorithm

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1. Randomly select ‘c’ cluster centers.
2. Calculate the distance between each data point and cluster centers.
3. Calculate the distance between each data point and cluster centers.Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. Recalculate the new cluster center using:

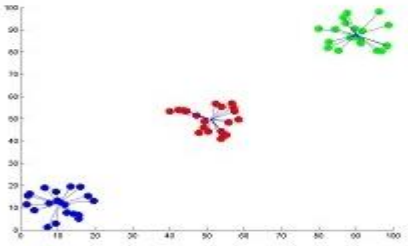
$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, ‘ci’ represents the number of data points in ith cluster.

5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data point was reassigned then stop, otherwise repeat from step 3).

Advantages

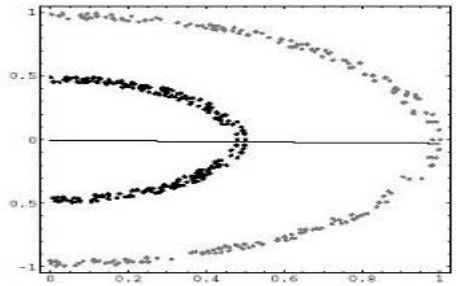
1. Fast, robust and easier to understand.
2. Relatively efficient: O(tknd), where n is # objects, k is # clusters, d is # dimension of each object, and t is # iterations. Normally, k, t, d << n.
3. Gives best result when data set are distinct or well separated from each other.



Showing the result of k-means for 'N' = 60 and 'c' = 3

Disadvantages

1. The learning algorithm requires apriori specification of the number of cluster centers.
2. The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.
3. Algorithm fails for non-linear data set.

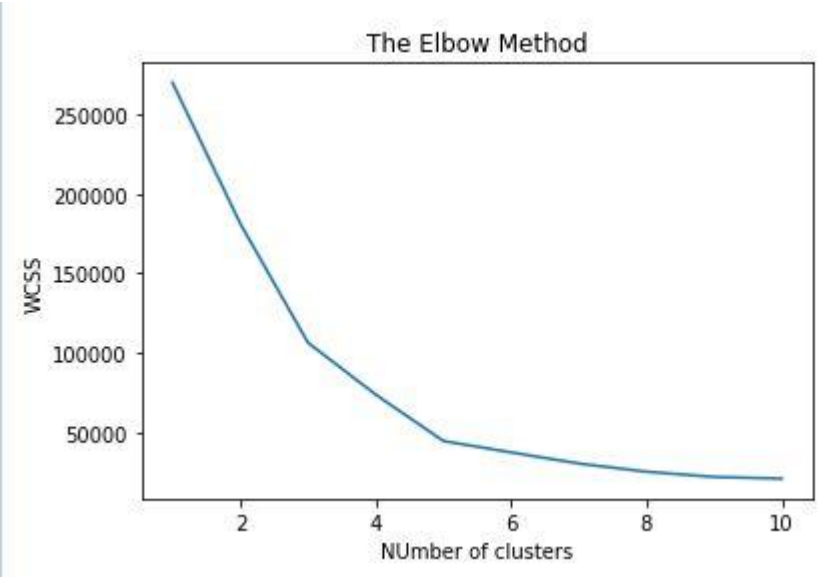


: Showing the non-linear data set where k-means algorithm fails

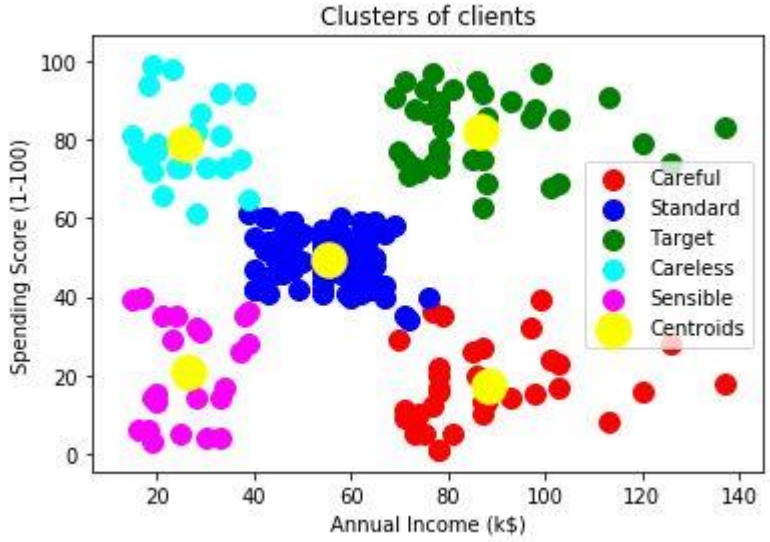
#Source Code

```
import matplotlib.pyplot as plt
import pandas as pd                          #Importing necessary libraries

# Importing the dataset
dataset = pd.read_csv('Mall_Customers.csv')
X = dataset.iloc[:,[3,4]].values             #Getting all rows and third and fourth column

# Using the elbow method to find the optimal number of clusters
from sklearn.cluster import KMeans
wcss = []
for i in range (1,11):
    kmeans = KMeans(n_clusters = i,init = 'k-means++',max_iter = 300,n_init =
10,random_state = 0)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
plt.plot(range(1,11),wcss)
plt.title('The Elbow Method')
plt.xlabel('NUmber of clusters')
plt.ylabel('WCSS')
plt.show()

# Appying to mall dataset
kmeans = KMeans(n_clusters = 5,init = 'k-means++',max_iter = 300,n_init =
10,random_state = 0)
y_kmeans = kmeans.fit_predict(X)

# Visualization of the clusters
plt.scatter(X[y_kmeans == 0,0],X[y_kmeans == 0,1],s = 100,c = 'red', label = 'Careful')
plt.scatter(X[y_kmeans == 1,0],X[y_kmeans == 1,1],s = 100,c = 'blue', label = 'Standard')
plt.scatter(X[y_kmeans == 2,0],X[y_kmeans == 2,1],s = 100,c = 'green', label = 'Target')
plt.scatter(X[y_kmeans == 3,0],X[y_kmeans == 3,1],s = 100,c = 'cyan', label = 'Careless')
plt.scatter(X[y_kmeans == 4,0],X[y_kmeans == 4,1],s = 100,c = 'magenta', label = 'Sensible')
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],s = 300, c =
'yellow',label = 'Centroids')
plt.title('Clusters of clients')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```

#Visualisation the elbow method to find the optimal number of clusters



Visualization of the clusters



#Conclusion

In this experiment we have seen a type of unsupervised clustering algorithm that is k-means. We have also seen a method named as elbow to determine the number of clusters and also we have gone through the mechanism of how to find the optimum number of clusters using k-means. We took a simple example of such a customer of shopping mall based on their annual income and their expenses.