**PRACTICAL NO –6**

**TITLE- K-Means clustering.**

**THEORY- -**K- Means *K*-means clustering is a type of unsupervised learning, which is used when you have un-labelled 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:

1. The centroids of the *K* clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyse the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.

Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

**PROBLEM STATEMENT**- To develop an algorithm for K-Means clustering.

**ALGORITHM-**

**1.Importing the libraries.**

**2.Import the datasets.**

**3.Using the elbow method to find the optimal number of clusters**

**4. Fitting K-Means to the dataset**

**5. Visualising the clusters**

**6.End**

**CODE-**

# -\*- coding: utf-8 -\*-

"""

Created on Tue Jul 10 10:34:51 2018

K-Means Clustering

@author: RoKu

"""

#import libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

#import dataset

dataset = pd.read\_csv('Mall\_Customers\_v.csv')

X = dataset.iloc[:,[3,4]].values

"""

Here we are not predicting anything so there is no X and y here we only have X to clus

plot also we are not using certain columns of the dataset

No encoding needed

No splitting needed

"""

#Using elbow method to find optimal number of clusters

from sklearn.cluster import KMeans

wcss = []

for i in range(1,11):

kmeans = KMeans(n\_clusters=i,init='k-means++',n\_init=10,max\_iter=300,random\_state=42)

kmeans.fit(X)

wcss.append(kmeans.inertia\_)

#use plot method to plot score and cluster

plt.plot(range(1,11),wcss)

plt.title('The Elbow Method')

plt.xlabel('Number of cluster')

plt.ylabel('wcss-values')

"""

Use the number of cluster such that the slope is less for further K and

is changing insignificantly...

Observe the 5-6 slope on this graph..

here 5 is the optimal value of the clusters

"""

##fitting the K-Means to the dataset

kmeans = KMeans(n\_clusters=5,init='k-means++',n\_init=10,max\_iter=300,random\_state=42)

y\_kmeans = kmeans.fit\_predict(X)

"""

here the fitting method is used with fit-pedict |^|^

"""

#visualize

"""

get to know how these clusters are stored in y\_kmeans

"""

plt.figure()

plt.scatter(X[y\_kmeans==0,0],X[y\_kmeans==0,1],s=100,c='red', label='Standard')

plt.scatter(X[y\_kmeans==1,0],X[y\_kmeans==1,1],s=100,c='blue', label='Target-1')

plt.scatter(X[y\_kmeans==2,0],X[y\_kmeans==2,1],s=100,c='green',label='Exclude')

plt.scatter(X[y\_kmeans==3,0],X[y\_kmeans==3,1],s=100,c='cyan', label='Careless-lure them')

plt.scatter(X[y\_kmeans==4,0],X[y\_kmeans==4,1],s=100,c='black',label='They will spend..target')

plt.scatter(kmeans.cluster\_centers\_[:,0],kmeans.cluster\_centers\_[:,1],s=100,c='yellow',label='Centroids')

plt.title('Clusters of customers')

plt.xlabel('Annual Income (in thousand rupees)')

plt.ylabel('Spending Score (range-1:100)')

plt.legend()

plt.show()

**RESULTS**-

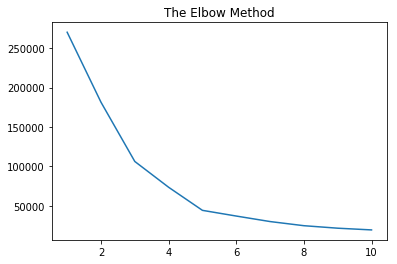
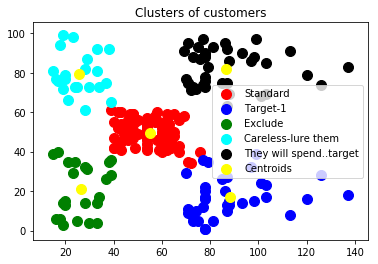


Fig.1-Elbow Method Visualization.

Hence, we can see that the elbow point is value ‘3’ on the X- axis. Thus we take the number of clusters equal to 3 in our python code. Below figure shows the clusters :--



**DISCUSSIONS-**

**We have our data set in which we have the murder and assault rates of the population of several cities in the world. We have to thus classify the population of these cities based on the people who have committed those crimes. Hence we have used the K Means Clustering algorithm as a solution to our problem statement.**

**We have first used the Elbow Method with the help of WCSS in order to find the optimal number of clusters to be used in our K Means Clustering algorithm.**

**Later we got to know that we need only 3 clusters and so we updated the python code and visualised the results.**

**CONCLUSIONS-**he K-means clustering algorithm that generates partitions of large datasets may provide a better characterization of the population who committed crimes and may be of additional benefit in distinguishing criminals and disciplined civilians compared with whole mixed mean of population alone.

|  |  |
| --- | --- |
| **SUBMISSION DATE**-  15/03/2019 | **SIGN OF COURSE INSTRUCTOR**- |
| **ROLL NO OF THE STUDENT-**  **I-01** | **NAME OF THE STUDENT-**  **Rohit Kulkarni** |