**Project Part 2**

Clustering Using

K-means

A**rijit Panda**

28th March 2020

# K-Means

## Introduction

## Clustering is the grouping of similar objects into a group known as cluster. Objects a particular cluster are different compared to objects present in another cluster. Since we do not have the ground truth to compare the result of the clustering algorithm with the actual value, it is an unsupervised learning method. For this part of the project, we have used k-means clustering algorithm to find the clusters. K-means algorithms is an iterative algorithm, which divides the data set into k-predefined centers to form the clusters.

## Objective

## This report summarizes the clustering results for K-means algorithm based on two approaches of selecting the initial clusters for clustering. The approaches are as follows:

## Randomly pick the initial clusters from the given data samples

## pick the first center randomly; for the i-th center (i>1), choose a sample (among all possible samples) such that the average distance of this chosen one to all previous (i-1) centers is maximal.

## Calculations Required

## Mean/Centroid (μ)

## Euclidean Distance :

## We calculate Euclidean distance to measure similarity.

## Objective Function:

## The goal of K-means is to minimize the objective function mentioned below. So we plot objective function value vs number of clusters to show it clearly

## 

## Pseudo Code (Algorithm)

## Specify number of clusters (k)

## Initialize k centroids either by strategy 1 or strategy 2 as mentioned before

## For each point calculate Euclidean distance from all centroids. Assign the point to the nearest centroid.

## For each cluster calculate the new centroid by taking average of all points in that cluster

1. Keep repeating 3 and 4 until there is no change in all centroid values.

## Result:

## Strategy 1:

## Example 1:

## The objective function value for cluster sizes 2-10 is

[2498.1135603167722, 1338.1076016520994, 792.7234630260813, 598.5546443663118, 467.65908827443667, 389.9259678246258, 431.850510233623, 241.37192245730458, 247.32593228077965]

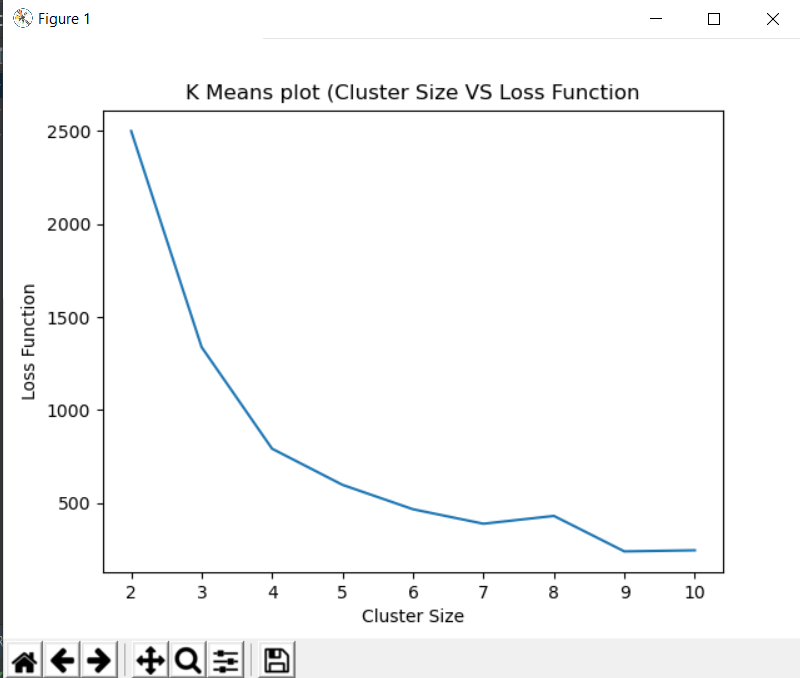


Fig-1

## Example 2:

## The objective function value for cluster sizes 2-10 is

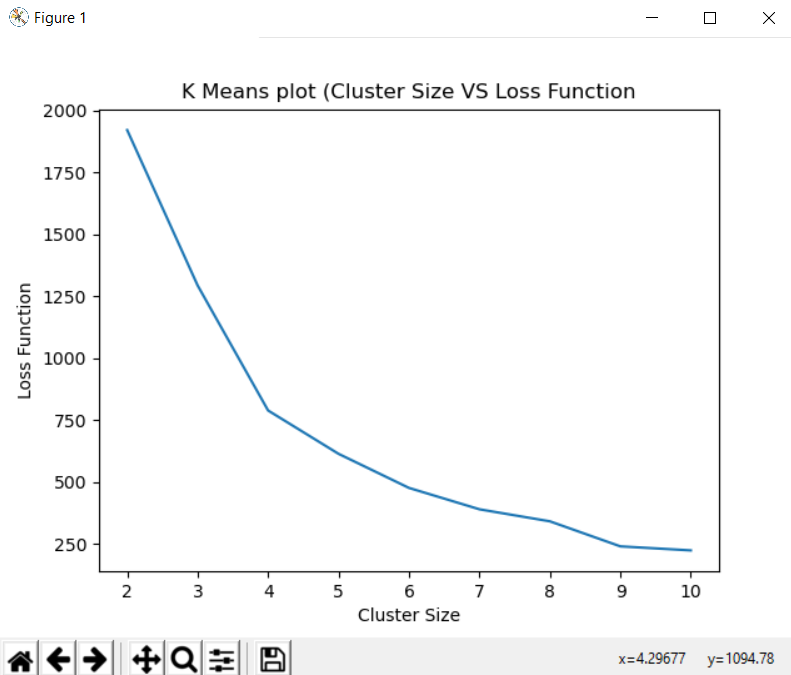
[1921.0334858562053, 1293.777452391135, 788.9645806635206, 613.986628606663, 476.2965705269664, 390.02655879270344, 341.43136465668397, 240.3803140277054, 224.08688719123077]

Fig-2

## Strategy 2:

## Example 1:

## The objective function value for cluster sizes 2-10 is

## [1921.0334858562053, 1293.777452391135, 805.1166457472608, 592.5283842592465, 476.11875167635293, 452.2372575407144, 340.9178565866489, 332.68708044807516, 261.25253133214545]

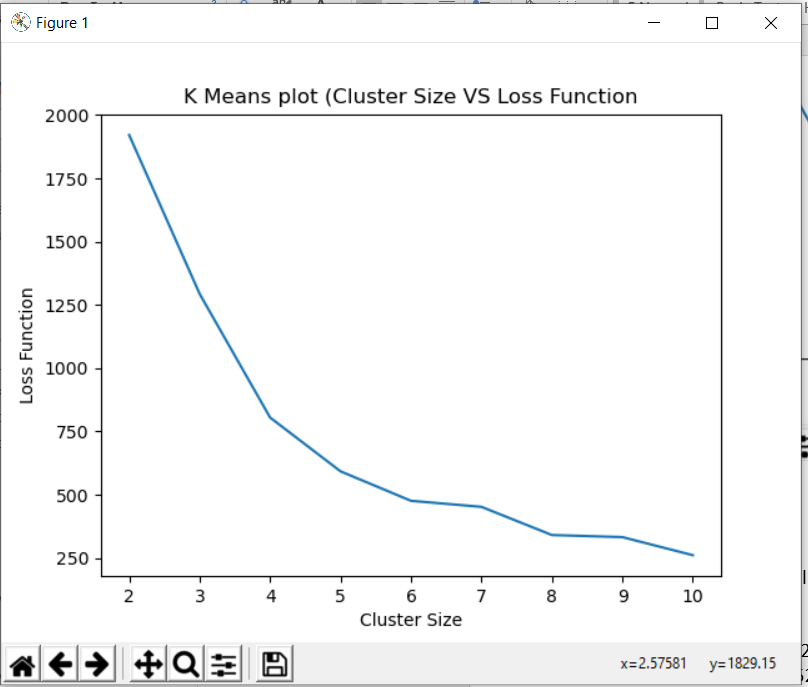


Fig-1

## Example 2:

## The objective function value for cluster sizes 2-10 is

[1921.0334858562053, 1294.298417485318, 803.2167238057566, 613.282439205604, 476.2965705269664, 397.44812244413663, 290.92433447443756, 281.0247981815767, 293.03884409460176]

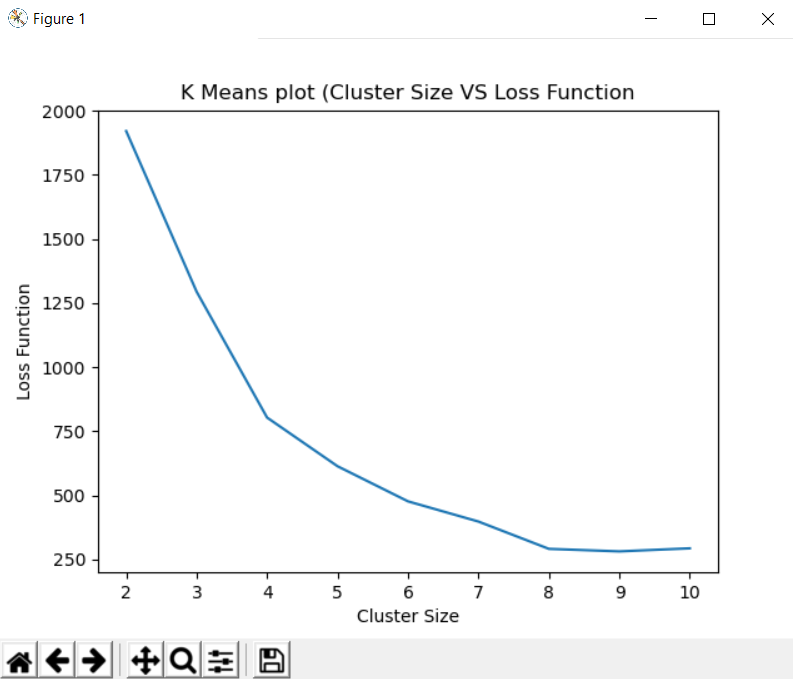


Fig-2