CSE 575

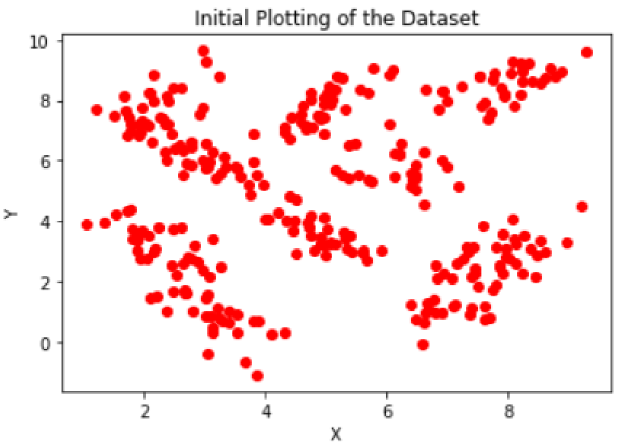
Statistical Machine Learning

Project 2: Unsupervised Machine Learning Algorithm Implementation (KMeans)

Name: Sai Teja Vishal Jangala

ASU ID: 1218332037

In the given dataset, we have 300 points that are plotted as below:



We have now use Unsupervised Machine Learning algorithm, KMeans algorithm to cluster these 300 datapoints. I have implemented KMeans algorithm using 2 strategies. And for each of the strategy, we have varied the cluster size K from 2 to 10.

For every strategy, for every cluster size, I have plotted the scatter-plot using Matplotlib and the results will be displayed here.

Each Strategy, the results are displayed below:

Strategy 1:

In this strategy, we have used randomly picked initial centers from the dataset given to us, and calculated the distance between these centers to each data points.

This done using the formulae below,

$D_x = ||x_i – \mu_{xj} ||^2$, where 1<=i<=n and 1<=j<=k, μ is the centroid

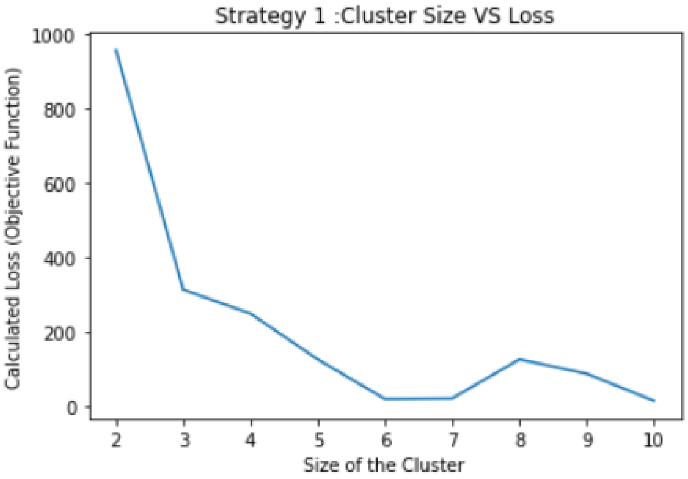
$D_y = ||y_i – \mu_{yj} ||^2$, where 1<=i<=n and 1<=j<=k, μ is the centroid

The distances above are calculated and each datapoint is assigned to it nearest centroid.

The Objective function is calculated and the formula used for objective function is as below:

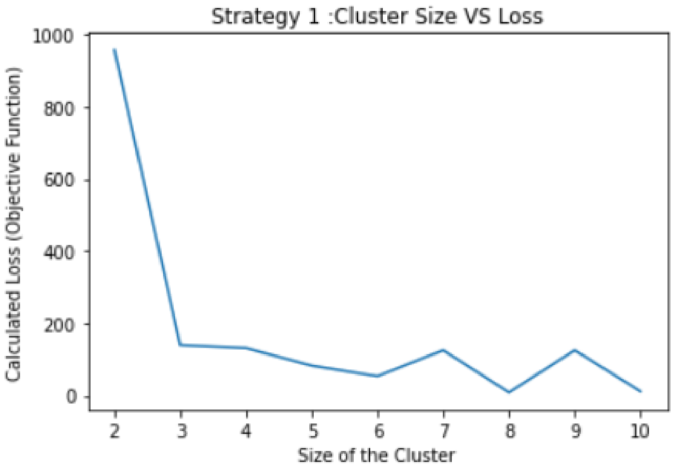
Formula: $[\ \Sigma_{1 \text{ to } K} \Sigma_{Di} \ ||\ x - \mu_i\ ||^2\ ]$

The plot obtained using the above Objective function for Strategy 1 is below:



This plot is obtained after the first initialization with Random centroid as planned to perform in Strategy 1.

After the second initialization of the centroids, with random centrois, the objective function looks as below:



Observations:

1. **Plot 1:**
   In this plot, we can observe that the Loss calculate by object function decreased from centroids K=3 to k=6.
   It then increases from k=6 to k=8.
   It then finally decreased from k=8.

2. **Plot 2:**
   In this plot, the object loss decreases slightly from k=3 to k=6 and then decreases at k=7 and decreases at k=9
   When comes to increase in the loss, the loss increases at k=6 and k=8 as observed in the above graph.

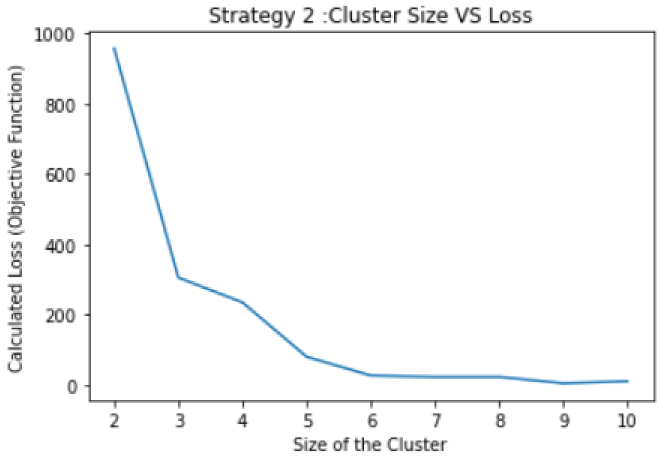
Strategy 2:

In this strategy, the first centroids are taken in random by the random-initialization within the dataset and the second initialization is taken that the average distance from previous initialization is maximum.

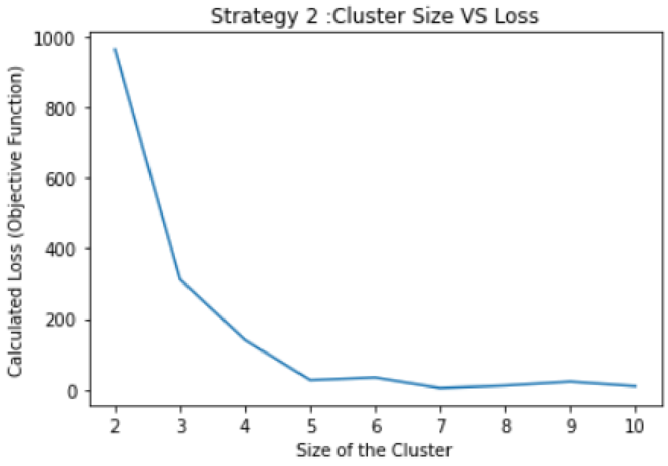
The Objective function is calculated, and the formula used for objective function is as below:

Formula: $[\Sigma_{1 \text{ to } K} \Sigma_{D_i} || x- \mu_i ||^2]$ is the Objective Function.

Calculating the loss for the Objective function using the strategy 2 is as below:



After the second initialization, computing the loss from the Objective function and plotting is will give a line chart as below:



Observation:

1. **Plot 1:**
   From K=3 to K=6, the loss is steadily decreasing, and from k=6, there is no significant increase in the loss.
2. **Plot 2:**
   From k=3 to k=5, the loss is steadily decreasing and this in an improvement showed over the first initialization. After k=5, there is no sudden increase in the Loss function.

Conclusion:

In the conclusion, I would like to say that Strategy 2 performs way better than the Strategy 1. Instead of taking random centroids in every step, it is not efficient compared to take random centroids in the first step and then updating centroids by considering the farthest point.

Strategy 2 outperforms strategy 1.