CSE 575 - PROJECT 2 - K Means - Report

ABHISHEK PATIL - Student ID: 1216126522

INTRODUCTION - The project implements two strategies of K means algorithm, an unsupervised clustering technique. Numerical results of these strategies are included in the results section below.

RESULTS (Given points data for My ID: 1216126522)

Strategy 1 - k Means Random Initialization of initial centroids

1. **k1** = 3, **i_point1** = (5.77144223,9.04075394), (1.96633923,7.30845038), (2.97097541,2.39669382)

Finial Centroids-
(6.49724962,7.52297293),(2.56146449,6.08861338),(5.47740039,2.25498103)
Loss / Objective Function- 1293.7774523911348

2. **k1** = 5, **i_point2** = (3.75004647,4.90070114), (2.10606162,8.23183769), (2.81629029,3.1999725), (4.30228618,7.08489147), (2.69511302,5.93967352)

Finial Centroids-
(7.55616782,2.23516796), (3.10305616,7.09676725), (2.68198633,2.09461587), (7.22084656,8.44524898), (5.36049077,4.49678809)
Loss / Objective Function- 592.87792926547263

Strategy 2 - k Means++ Initialization of initial centroids

1. **k1** = 4, **i_point1** = (1.52668895,4.24557918)

Calculated Initial Centroids-
(1.52668895,4.24557918),(9.26998864,9.62492869), (3.85212146,-1.08715226), (2.95297924,9.65073899)
Finial Centroids-
(3.30296804,2.55443267), (6.85658333,7.6614342), (7.34802851,2.35222497), (3.153427,6.9129207)
Loss / Objective Function- 792.53781044133041

2. **k1** = 6, **i_point2** = (2.87448907,2.657599)

Calculated Initial Centroids-
(2.87448907,2.657599), (9.26998864,9.62492869), (3.85212146,-1.08715226), (2.95297924,9.65073899), (7.68097556,0.83542043), (8.87578072,8.96092361)
Finial Centroids-
(3.49556658,3.56611232), (7.75648325,8.55668928), (3.14506148,0.90770655), (2.56333815,6.9782248), (7.41419243,2.32169114), (5.46427736,6.83771354)
Loss / Objective Function- 476.11875167635299

OBSERVATION AND ANALYSIS (Given points data for My ID: 1216126522)

For each strategy, I ran the algorithm with the given initial centroid points for different k values ranging from 2 to 10. The plots are as follows:

Strategy 1 Initial centroids for k = 2,4,6,7,8,9,10 are chosen randomly. Given to us for k = 3,5.

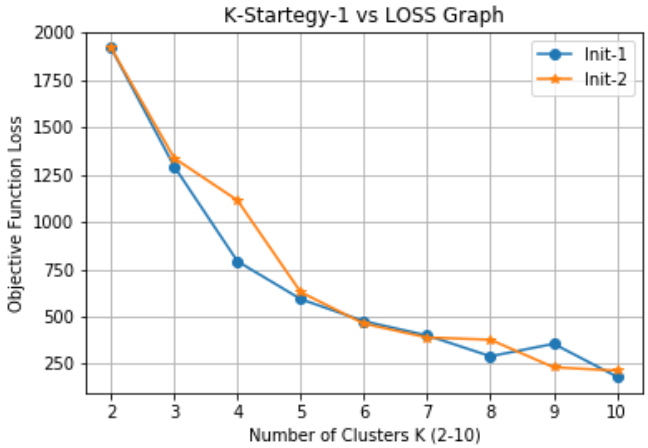

Fig.1. K-Means-Strategy 1 Objective Function graph for two different initial values

Strategy 2 Initial first centroid for k = 2,3,5,7,8,9,10 are chosen randomly. Given to us for k = 4,6. The rest of the k-1 centroids for each k are calculated using the strategy 2 (and *NOT random*).

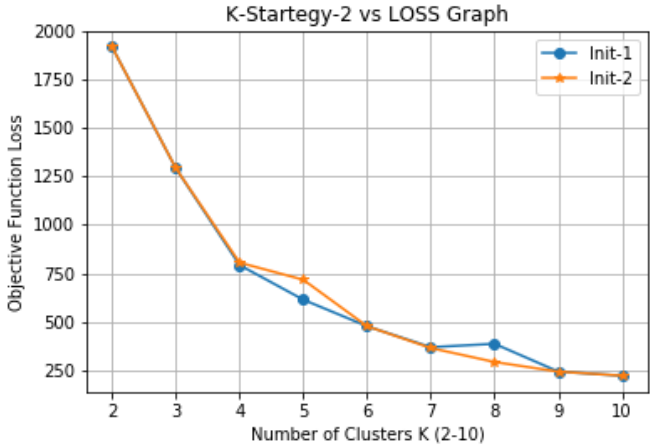

Fig.2. K-Means-Strategy 2 Objective Function graph for two different initial values

ANALYSIS

We studied K-means using two different strategies. In the first strategy, we used random centroids from the data. However, in the second strategy we used only one random initial centroid and built other reasonably further away centroids from it. The second strategy feels intuitively better to find clusters that are well separated from each.

Based on the 2 strategies the plots in fig.1 and fig.2 are created. The elbow graph of K versus loss is easily seen here. Strategy 2 seems to show a better curve because of smartly chosen initial centroids. It has a more smooth decrease of loss in comparison to strategy one.

As is shown in the graphs, strategy 2 helps remove the problem of centroid initialization which is present in strategy 1. Having the initial centroids as far as possible gives more chances to create good well spaced clusters using normal K means.