Student ID: 1204599352

In Strategy 1 of this project, we were provided with the initial centroids and asked to determine the final centroids using the Kmeans algorithm. This was done using k = 3 and k = 5 clusters for our dataset of 300 (x,y) combinations.

For k = 3, my initial centroids were (6.2091503, 6.16038763), (1.91568768, 6.83080871), (3.04743588, 1.55807635). Using these centroids, I assigned the 300 data points to clusters using the shortest Euclidean distance to a given centroid. Euclidean distance =

Next, the centroids were recomputed by determining the average of the points contained within each centroid

This process was repeated until the clusters assigned to the 300 points did not change.

The final centroids were calculated as (6.49724962, 7.52297293), (2.56146449, 6.08861338), (5.47740039, 2.25498103).

After determining the final centroids, I calculated the loss based on the following objective function with the overall goal of the Kmeans algorithm being to minimize this value.

The loss was calculated to be 1293.777.

This process was repeated for k = 5 clusters and for another set of 3 and 5 initial centroids. The final results for each of these tests can be found below.

In Strategy 2 of this project, we were provided with one initial centroid, and the remaining initial centroids were determined by finding the maximum average distance from a point to the set of centroids. This point was added to the set of initial centroids until the number of initial centroids was equal to k. This was done for k = 4 and k = 6 clusters for our dataset of 300 (x,y) combinations.

For k = 4, the given initial centroid was (1.51180219, 7.48293717). The point with the greatest distance to this initial centroid was added to the set. The average distance to the points in the centroid set was calculated for each point, and the point with the greatest distance was added to the set. The process outlined in Strategy 1 was then followed for determining the final centroids and loss. The results can be seen below.

From the results, we can see the loss function has been greatly reduced with the use of more clusters. After a certain point, the reduction in the loss function will be minor, and we will no longer be able to produce greater improvements by increasing the number of clusters without overfitting.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | K = 3 Set 1 (ID = 9352) | K = 3 Set 2 (ID = 1111) | K = 5 Set 1 (ID = 9352) | K = 5 Set 2 (ID = 1111) |
| Loss | 1293.777 | 1338.17 | 653.9568 | 649.8721 |
| Initial Centroids | (6.2091503, 6.16038763), (1.91568768, 6.83080871), (3.04743588, 1.55807635) | |  | | --- | | (4.74625798, 3.54661053) | | (9.21069612, 4.5106493) | | (6.2153903, 6.26139225) | | (2.61234619, 8.39116666), (7.89366657, 3.58341277), (3.04101702, -0.36138487), (5.60001917, 3.02332715), (5.40840468, 2.96754178) | |  | | --- | | (2.40998489, 7.99174945) | | (2.36057145, 6.00796623) | | (1.9311184, 6.93692984) | | (7.44472802, 2.41849318) | | (4.74625798, 3.54661053) | |
| Final Centroids | |  | | --- | | (6.49724962, 7.52297293) | | (2.56146449, 6.08861338) | | (5.47740039, 2.25498103) | | |  | | --- | | (3.36812267, 2.62265154) | | (7.41014308, 2.369049) | | (4.84461158, 7.30111158) | | |  | | --- | | (3.006688, 7.03481503) | | (6.80421029, 7.6562222) | | (3.14506148, 0.90770655) | | (7.41419243, 2.32169114) | | (3.5545081, 3.62813062) | | |  | | --- | | (6.7194294, 7.74332931) | | (3.9151203, 4.04166689) | | (2.81432195, 7.08096483) | | (7.41419243, 2.32169114) | | (2.9587703, 1.43276437) | |

Strategy 1

Strategy 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | K = 4 Set 1 (ID = 9352) | K = 4 Set 2 (ID = 1111) | K = 6 Set 1 (ID = 9352) | K = 6 Set 2 (ID = 1111) |
| Loss | 805.1166 | 792.7235 | 469.1317 | 476.1188 |
| Initial Centroid | (6.65537695, 1.30451652) | (1.51180219, 7.48293717) | (4.32239695, 0.33088885) | (9.26998864, 9.62492869) |
| Final Centroids | (6.7837460879035447, 2.8501999866465444), (3.3426476922451784, 6.9260280301001709), (7.1792862118615073, 8.0520790997296903), (2.8523514931105352, 2.2818648297203246)) | (3.1693274926251656, 6.8871389244710501), (7.2526268312565767, 2.4001582635520533), (6.8565833302289008, 7.6614341998064841), (3.2125746077046626, 2.4965808657995252)) | (7.5561678223977262, 2.235167959857534), (7.9143099778183137, 8.5199098077000759), (2.5000777598402069, 6.8398080230756451), (2.6977678104318801, 2.0545644767702576), (5.3321424077429915, 4.444481111506235), (4.9237310901500182, 7.8337315427726981)) | (7.7564832491464841, 8.5566892790634146), (3.4955665791995627, 3.5661123157286907), (2.5633381461259046, 6.9782248006066236), (7.414192434680615, 2.3216911383868664), (5.4642773567278944, 6.8377135364358912), (3.1450614829591448, 0.90770654865881528) |

