Lucy Zhang
02/06/23
ID: 1229576766

Project 2: K-means-Strategy

Introduction:

In this project, I'm required to implement the K-means algorithm and apply my implementation on the given dataset (AllSamples.npy), which contains a set of 2-D points. I'm required to implement two different strategies for choosing the initial cluster centers with function below:

$$\Sigma_{i=1}^{k}\Sigma_{x\in D_i}\left|\left|x-\mu_i\right|\right|^2$$

K-means Strategy:

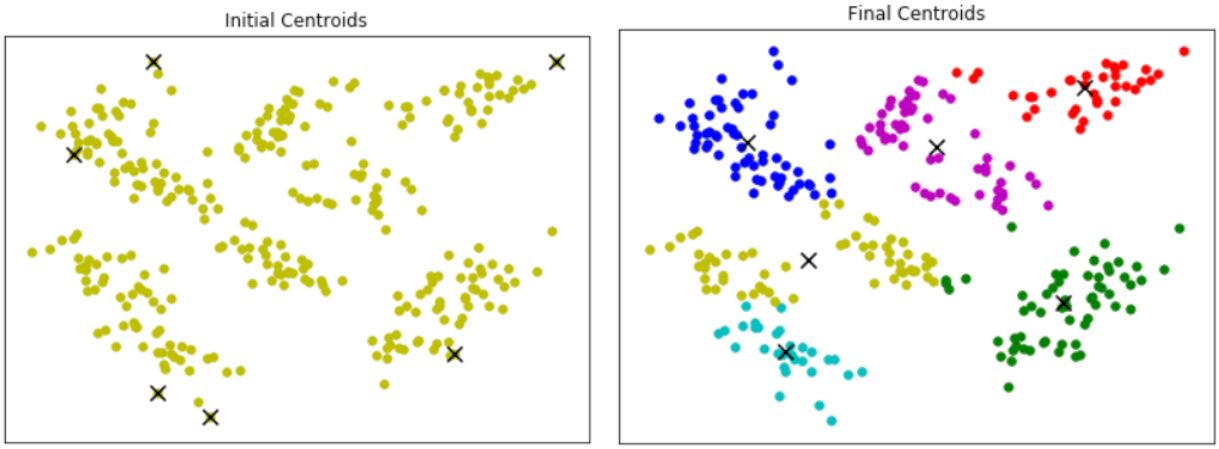
With given number k of clusters and initial cluster centers, I have computed the final coordinate of the centroids and compute the loss based on the objective function.

Clusters	Initial Cluster Centers	Final Centroids	Loss
3	[6.09952696 9.0178614] [8.21925014 9.11712554] [6.2396717 6.55049457]	[2.56146449 6.08861338] [6.49724962 7.52297293] [5.47740039 2.25498103]	1293.78
4	[3.57542555 5.47748903] [9.26998864 9.62492869] [3.85212146 -1.08715226] [2.95297924 9.65073899]	[3.33995748 2.59215224] [6.60345839 7.57042104] [7.38076264 2.33245532] [2.85859235 6.93136525]	788.27
5	[5.36626615 6.51434231] [3.0226944 0.86402039] [8.07641652 9.27162002] [4.4280969 7.41377907] [7.57805025 3.82487017]	[5.37514379 4.53101654] [2.68198633 2.09461587] [7.49365367 8.52417952] [3.22202355 7.15937996] [7.55616782 2.23516796]	592.07
6	[1.72614408 6.81819407] [7.68097556 0.83542043] [9.26998864 9.62492869] [3.85212146 -1.08715226] [2.95297924 9.65073899] [3.04101702 -0.36138487]	[2.56333815 6.9782248] [7.41419243 2.32169114] [7.75648325 8.55668928] [3.14506148 0.90770655] [5.46427736 6.83771354] [3.49556658 3.56611232]	476.12
10	…	…	229.05
20	…	…	103.65

Implementation:

For K-means strategy, I wrote a function implements the k-means algorithm to compute the centroids and the loss. The algorithm updates the centroids and the loss until the difference between the old and new centroids is below 0.0001.

For K-means ++ strategy, I wrote a function for choosing initial centroids for the k-means clustering algorithm. It calculates the distances between each sample in the data and the centroids that have been selected so far and finds the sample with the maximum mean distance to the selected centroids.



(Example of centroids position with 6 clusters)

Observation and Analysis:

While working on this project, I have practiced K-means with large data and observed that the loss gets smaller when clusters increase. Also, K-Means++ is an improvement of the K-Means algorithm that tries to address the issue of poor initial centroid selection in K-Means. In K-Means++, the initial centroids are selected in a way that spreads them out as far as possible from each other, instead of being randomly chosen. This increases the chances of obtaining better clustering results compared to K-Means.