Jinsan Kim
CSE575 Statistical Machine Learning
Project 2 Summary

K-Means Project

1. Introduction

In this project, we are going to use two K-Means strategies on an unsupervised data set. In the first strategy, number k from 2 to 10 and their initial centroids are given to begin with. We are supposed to find the last centroids and the loss. In the second strategy, number k from 2 to 10 and their first initial centroid are given. The rest of the initial centroids need to satisfy the following: "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" (Li, 2020). We need to find the rest of the initial centroids. Rest of the steps are same as the first strategy, we find the final centroids and the loss.

2. Loss Formula and K-Means Steps

a. Loss Formula
   i. "When clustering the samples into k clusters D_i, with respective center vectors $\mu_1, \mu_2, \dots, \mu_k$, the objective function is defined as $\sum_i^k \sum_{x \in D_i} \|x - \mu_i\|^2$" (Li, 2020).

b. For the first strategy, steps are following:
   i. Set the given k and initial centroids
   ii. Find which centroid is the closest to each sample.
   iii. Calculate the mean of each cluster and assign it as a new centroid
   iv. Repeat step two and three until you get the same centroids again.

c. For the second strategy, steps are following:
   i. Set the given k
   ii. Find the sample with maximum distance from the given point and assign it as the second point
      1. When you calculate the maximum distance, you can use the Euclidean distance.
   iii. Repeat step two until you have k number of initial centroids
   iv. Rest of the steps is same as the first strategy's step two to four

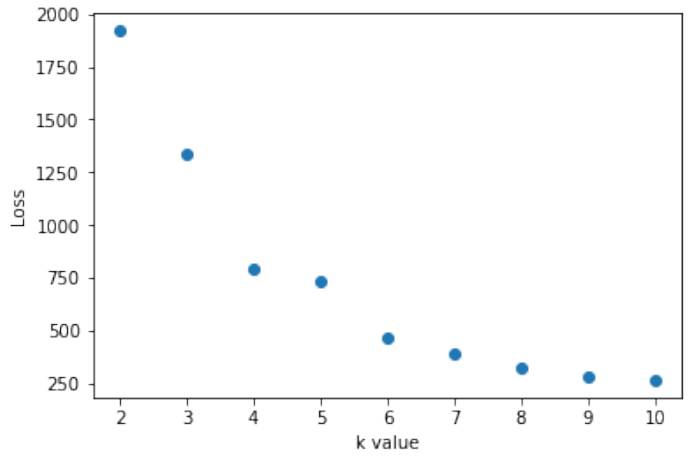
3. Result of Strategy 1 and Strategy 2

Strategy 1

```
k value:  2
final centroids:  [[5.0005623415887621, 2.4854274751531933], [4.852611930287174, 7.27164171120277
47]]
loss:  1921.03348586
k value:  3
final centroids:  [[7.2397511895844495, 2.4820826910731952], [3.2489642305948876, 2.5802769113756
954], [4.8337531753922862, 7.316058236043574]]
loss:  1338.1059838
k value:  4
final centroids:  [[7.2526268312565767, 2.4001582635520533], [3.2428534716041297, 2.5519790549394
479], [2.8337661044723124, 6.9189568989338754], [6.5795764318878449, 7.573335948203881]]
loss:  788.964580664
k value:  5
final centroids:  [[3.1383713351563061, 6.9382047631200781], [3.2047574481774648, 2.6974577504123
82], [6.3952811976054686, 0.73717126051528736], [6.8642851898847095, 7.7731608078896022], [7.3171
227859698824, 3.0475400723829336]]
loss:  737.039913387
k value:  6
final centroids:  [[2.5416525233103449, 7.0026783235364887], [5.2402829638043125, 7.5313102932678
211], [7.5561678223977262, 2.235167959857534], [2.6819863341889292, 2.0946158678008091], [7.91430
99778183137, 8.5199098077000759], [5.2305366674047375, 4.2793424960490993]]
```

```
loss:  462.926355825
k value:  7
final centroids:  [[7.9143099778183137, 8.5199098077000759], [4.8183305760856632, 3.6950232007185
4], [2.6819863341889292, 2.0946158678008091], [4.8593987502311542, 7.9416382089060544], [6.154682
2840552204, 5.7014072056794127], [2.5365010787901392, 6.859419784482168], [7.5561678223977262, 2.
235167959857534]]
loss:  389.925967825
k value:  8
final centroids:  [[2.1832146220312532, 7.7035534091809064], [2.7362500545223889, 1.9751010338399
708], [5.0217765979334894, 7.8240125818092165], [3.0006022905415501, 5.7829578632937819], [4.9125
149733766662, 3.5631409634126254], [7.9143099778183137, 8.5199098077000759], [7.5561678223977262,
 2.235167959857534], [6.1546822840552204, 5.7014072056794127]]
loss:  326.26502937
k value:  9
final centroids:  [[4.9125149733766662, 3.5631409634126254], [2.7362500545223889, 1.9751010338399
708], [2.1832146220312532, 7.7035534091809064], [7.9143099778183137, 8.5199098077000759], [6.1546
822840552204, 5.7014072056794127], [7.9252023260690194, 2.9385093784153717], [5.0217765979334894,
 7.8240125818092165], [7.0193903625122083, 1.2121258965006791], [3.0006022905415501, 5.7829578632
937819]]
loss:  276.712617353
k value:  10
final centroids:  [[7.9252023260690194, 2.9385093784153717], [4.9125149733766662, 3.5631409634126
254], [2.7362500545223889, 1.9751010338399708], [7.521973033096792, 8.1607039991590007], [2.18321
46220312532, 7.7035534091809064], [7.0193903625122083, 1.2121258965006791], [5.0217765979334894,
7.8240125818092165], [8.4112701077989058, 8.9749038318521084], [6.1546822840552204, 5.70140720567
94127], [3.0006022905415501, 5.7829578632937819]]
loss:  264.52659863
```

Plot for Strategy 1

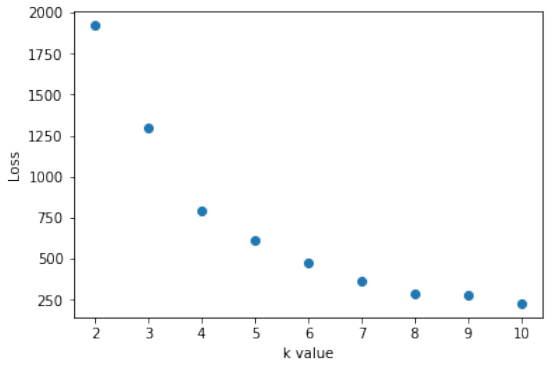


Strategy 2

```
k value:  2
final centroids:  [[5.0005623415887621, 2.4854274751531933], [4.852611930287174, 7.27164171120277
47]]
loss:  1921.03348586
k value:  3
final centroids:  [[2.5614644894663545, 6.0886133828954794], [5.4774003886790341, 2.2549810279847
229], [6.4972496208497086, 7.5229729298951709]]
loss:  1293.77745239
k value:  4
final centroids:  [[7.2526268312565767, 2.4001582635520533], [3.2285300905383707, 2.5240486292057
867], [6.6259253846324615, 7.5761491676226784], [2.9054774114449513, 6.9051227633399481]]
loss:  789.237972218
k value:  5
final centroids:  [[7.2526268312565767, 2.4001582635520533], [7.7564832491464841, 8.5566892790634
146], [2.6012329625686772, 6.9161050575199603], [3.2125746077046626, 2.4965808657995252], [5.4025
250775739151, 6.7363617521879933]]
loss:  613.282439206
k value:  6
```

```
final centroids:  [[3.4955665791995627, 3.5661123157286907], [7.7564832491464841, 8.5566892790634
146], [2.5633381461259046, 6.9782248006066236], [3.1450614829591448, 0.90770654865881528], [5.464
2773567278944, 6.8377135364358912], [7.414192434680615, 2.3216911383868664]]
loss:  476.118751676
k value:  7
final centroids:  [[2.2420475191254021, 3.2510074863184211], [7.9143099778183137, 8.5199098077000
759], [3.1690614508664035, 0.81432514729916761], [5.2402829638043125, 7.5313102932678211], [5.230
5366674047375, 4.2793424960490993], [2.5416525233103449, 7.0026783235364887], [7.556167822397726
2, 2.235167959857534]]
loss:  362.933114045
k value:  8
final centroids:  [[4.8183305760856632, 3.69502320071854], [3.1690614508664035, 0.814325147299167
61], [7.9143099778183137, 8.5199098077000759], [2.5365010787901392, 6.859419784482168], [7.556167
8223977262, 2.235167959857534], [6.1546822840552204, 5.7014072056794127], [2.2420475191254021, 3.
2510074863184211], [4.8593987502311542, 7.9416382089060544]]
loss:  289.932726045
k value:  9
final centroids:  [[4.8183305760856632, 3.69502320071854], [8.3987075293162672, 8.925497059743181
1], [3.1690614508664035, 0.81432514729916761], [4.7884251842441943, 7.8829164588807563], [7.55616
78223977262, 2.235167959857534], [6.1546822840552204, 5.7014072056794127], [2.5365010787901392,
6.859419784482168], [2.2420475191254021, 3.2510074863184211], [7.3419558760390631, 8.237439821229
1986]]
loss:  277.391433977
k value:  10
final centroids:  [[4.8183305760856632, 3.69502320071854], [8.3987075293162672, 8.925497059743181
1], [3.1690614508664035, 0.81432514729916761], [4.7884251842441943, 7.8829164588807563], [7.92520
23260690194, 2.9385093784153717], [6.1546822840552204, 5.7014072056794127], [2.5365010787901392,
6.859419784482168], [2.2420475191254021, 3.2510074863184211], [7.3419558760390631, 8.237439821229
1986], [7.0193903625122083, 1.2121258965006791]]
loss:  227.839021959
```

Plot for Strategy 2



Analysis for Strategy 1 and 2

We can see that as the number k increases the loss decreases. However, the loss of Strategy 1 is little higher than Strategy 2. In Strategy 2, when we find the initial centroids, we removed the samples when they become the centroids. I think by doing that we could eliminate the outliers in the samples and could have better initial centroids to begin with.

Works Cited

Li, B. (2020). Unsupervised Learning The K-Means Algorithm. (p. 4). Phoenix: ASU Ira A.Fulton Schools of Engineering.