Jinsan Kim
CSE575 Statistical Machine Learning
Project 2 Summary
Prof. Li

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, we are given k number of random centroids to begin with. We are supposed to find the last centroids and the loss. In the second strategy, we are given the first initial centroid and the number k. 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". 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

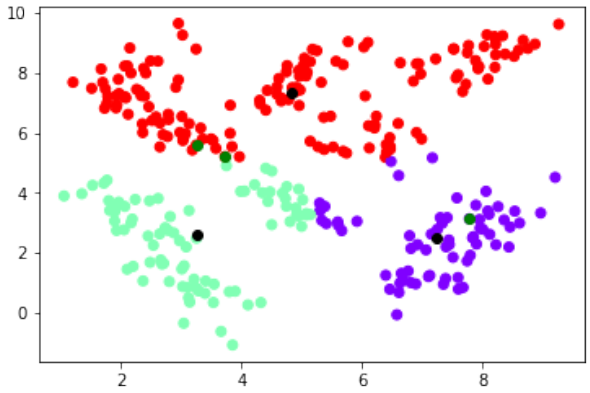
3. Result of Strategy 1 and Strategy 2

Strategy 1 Set 1

```
k:  3

Initial points assigned: [[7.78551305, 3.12724529], [3.72610844, 5.2043243
9], [3.2492998, 5.59125171]]

Final Centroids: [[7.2397511895844495, 2.4820826910731952], [3.24896423059
48876, 2.5802769113756954], [4.8337531753922862, 7.316058236043574]]

Total Cost:  1338.1059838
```

Plot for Strategy 1 Set 1

Green dots are the initial centroids and the black dots represent the final centroids. Different colors show their cluster.

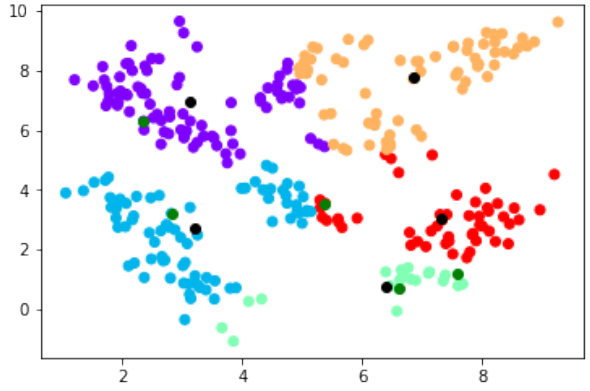


Strategy 1 Set 2

```
k:  5

Initial points assigned: [[2.3537231, 6.29810755], [2.81629029, 3.199972
5], [6.6161895, 0.66750633], [5.38398051, 3.53840433], [7.59731342, 1.1650
4743]]

Final Centroids: [[3.1383713351563061, 6.9382047631200781], [3.20475744817
74648, 2.697457750412382], [6.3952811976054686, 0.73717126051528736], [6.8
642851898847095, 7.7731608078896022], [7.3171227859698824, 3.0475400723829
336]]

Total Cost:  737.039913387
```

Plot for Strategy 1 Set 2

Green dots are the initial centroids and the black dots represent the final centroids. Different colors show their cluster.



Strategy 1 Analysis

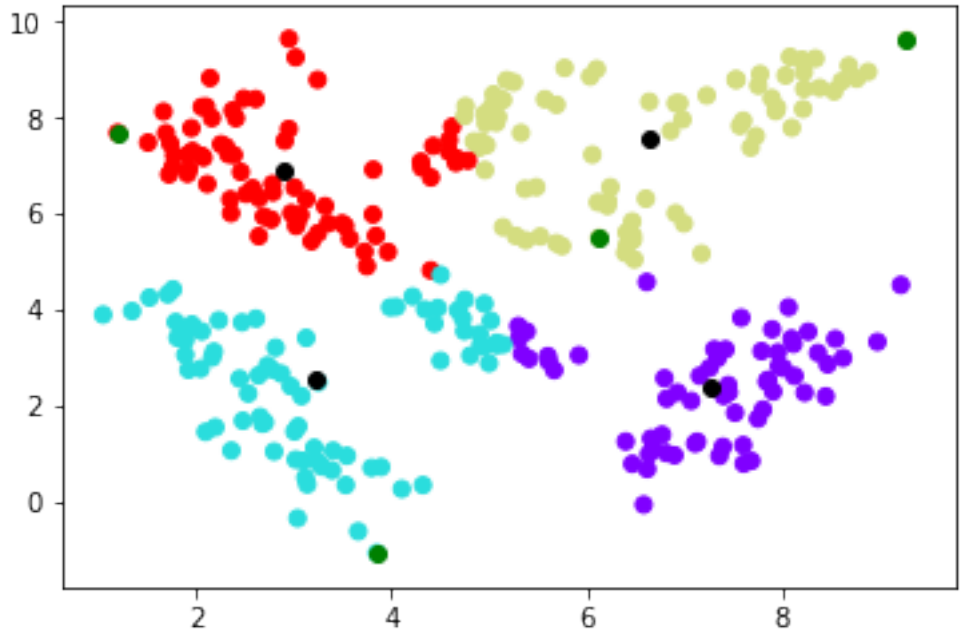
Set 1 has k = 3 and Set 2 has k = 5. The losses are 1338.1059838 and 737.039913387 respectively. The set with bigger k has less loss by looking at their “Total Cost”.

Strategy 2 Set 1

```
k:  4

Initial point assigned: [6.12393256, 5.49223251]

Initial point assigned and other initial centroids: [[6.12393256, 5.492232
51], [3.85212146, -1.08715226], [9.26998864, 9.62492869], [1.20162248, 7.6
8639714]]

Final Centroids: [[7.2526268312565767, 2.4001582635520533], [3.22853009053
83707, 2.5240486292057867], [6.6259253846324615, 7.5761491676226784], [2.9
054774114449513, 6.9051227633399481]]

Total Cost:  789.237972218
```

Plot for Strategy 2 Set 1

Green dots are the initial centroids and the black dots represent the final centroids. Different colors show their cluster.



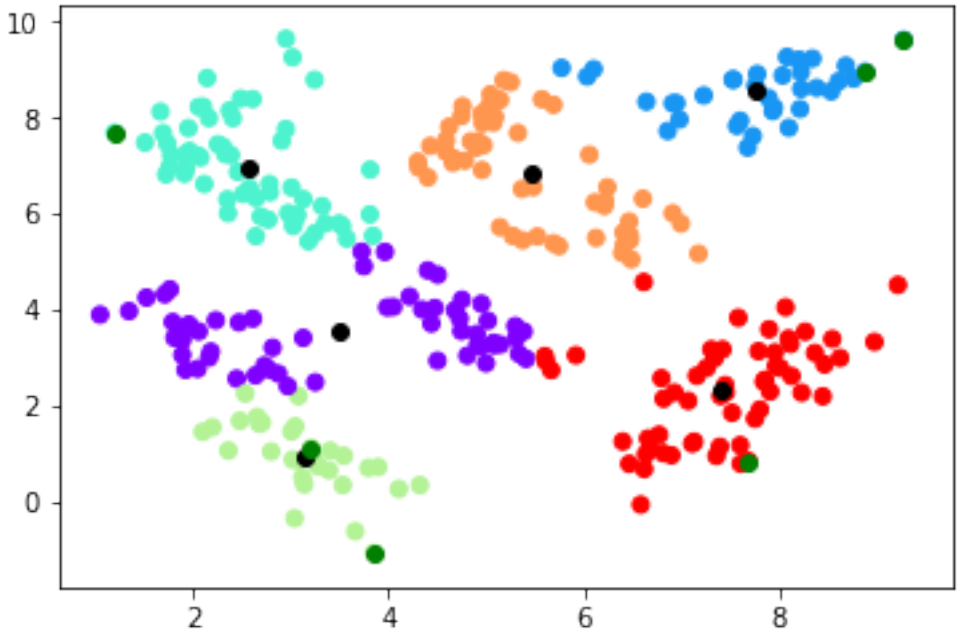
Strategy 2 Set 2

```
k:  6
```

```
Initial point assigned: [3.2115245, 1.1089788]

Initial point assigned and other k-1 centroids: [[3.2115245, 1.1089788],
[9.26998864, 9.62492869], [1.20162248, 7.68639714], [3.85212146, -1.087152
26], [8.87578072, 8.96092361], [7.68097556, 0.83542043]]

Final Centroids: [[3.4955665791995627, 3.5661123157286907], [7.75648324914
64841, 8.5566892790634146], [2.5633381461259046, 6.9782248006066236], [3.1
450614829591448, 0.90770654865881528], [5.4642773567278944, 6.837713536435
8912], [7.414192434680615, 2.3216911383868664]]

Total Cost:  476.118751676
```

Plot for Strategy 2 Set 2

Green dots are the initial centroids and the black dots represent the final centroids. Different colors show their cluster.



Strategy 1 Analysis

Set 1 has k = 4 and Set 2 has k = 6. The losses are 789.237972218 and 476.118751676 respectively. The numbers are from “Total Cost”. The set with bigger k has less loss.

Works Cited

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