Asmt 4: Hierarchical and Assignment-based Clustering

Vai Suliafu, u0742607
Tuesday, February 18

1. Hierarchical Clustering

0.1 Part A

For each measure of cluster distance (Single-Link, Complete-Link, Mean-Link), the results can be shown below, where the colors represent a point's assignment to a particular cluster.

0.2 Part B

By visual assessment, we can conclude that the "Mean-Link" distance between clusters provided the most favorable clustering, as it resulted in the most optimal combination of maximizing distances between clusters, and minimizing distances within clusters.

1 2. Assignment-based Clustering

1.1 Part A

For both the Gonzalez Algorithm and k-Means++, the results of the clustering can be viewed below, where the colors represent a point's assignment to a particular cluster.

3-Center Cost for Gonzalez Algorithm: 59.46920197567593
3-Mean Cost for Gonzalez Algorithm: 32.621576568155625

3-Center Cost for k-Means++ Algorithm: 72.83151707784722
3-Mean Cost for k-Means++ Algorithm: 9.509593092651711

We remark that from these comparisons, the Gonzalez algorithm seems to outperform the k-Means++ algorithm in minimizing the maximum distance from a point to its center. In contrast, the k-Means++ algorithm outperforms the Gonzalez algorithm in minimizing the average distance between a point and its center.

1.2 Part B

To better understand the behavior of the 3-Mean Cost for the k-Means++ algorithm, the algorithm was ran 100 times to cluster the same data. For each trial, the 3-Mean Cost was recorded.

A plot the recorded 3-Mean costs can be seen below. Thus, we can conclude that slightly over 30 percent of all simulations had a 3-Mean cost less than or equal to our observed cost from the Gonzalez Algorithm earlier.

1.3 Part C

To understand the behavior of Loyd's Algorithm, which attempts to optimize the placement of centers, a few experiments were performed. First, we arbitrarily selected our first three data points (ie, the points at index 0, 1, and 2) to be our centers.

Then, we let Loyd's Algorithm optimize these centers. The results of the clustering both before and after optimization can be seen at the bottom.

3-Means Cost before optimization = 38.3
3-Means Cost after optimization = 10.3

Moreover, we also wanted to see if Loyd's Algorithm could further optimize our earlier clustering results from the Gonzalez Algorithm.

The results are shown below.

3-Means cost before optimization = 34.62
3-Means cost after optimization = 17.1

Finally, we wanted to see Loyd Algorithms effect on the average 3-Means cost of k-Means++. Thus, we ran the k-Means++ algorithm 100 times, each

time optimizing the center selection with Loyd's Algorithm and recording the 3-Means cost. The distribution of the optimized 3-Means cost can be seen below (significantly lower than the distribution shown earlier).

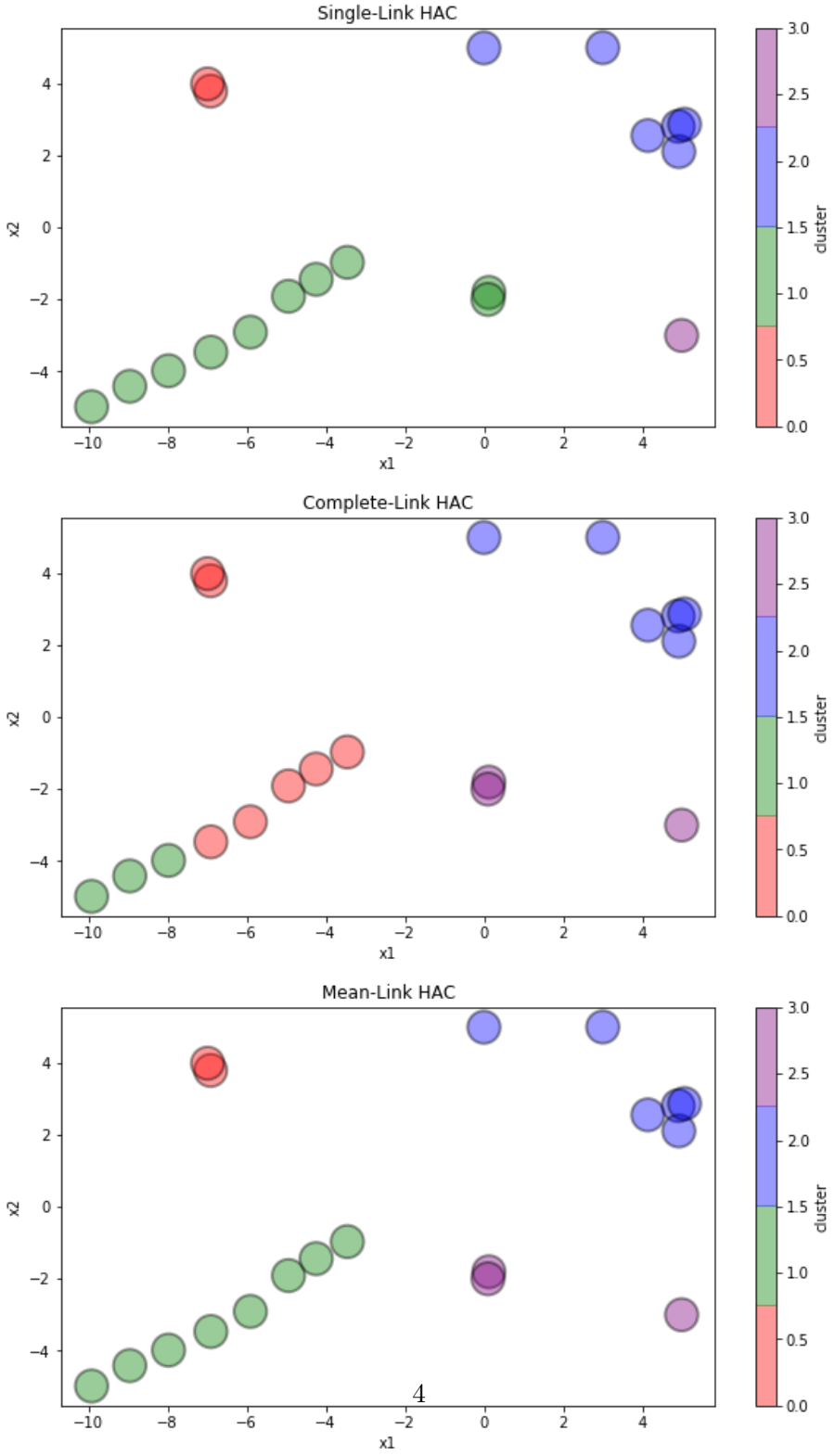

Figure 1: Various Distance Measures for HAC

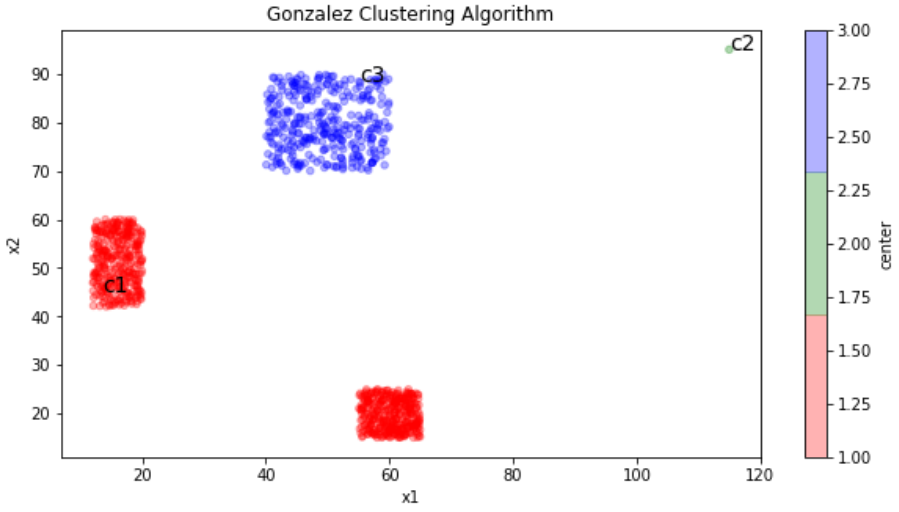

Figure 2: Gonzalez Algorithm Clustering

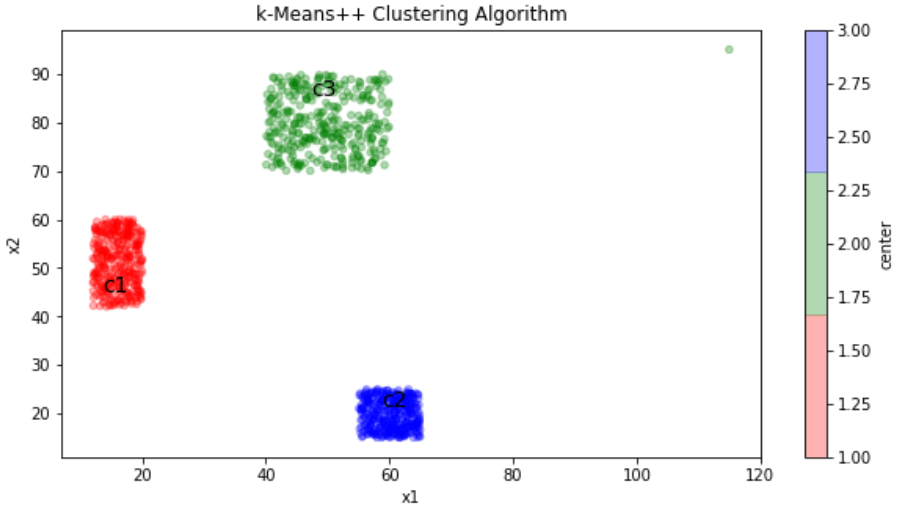

Figure 3: k-Means++ Algorithm Clustering

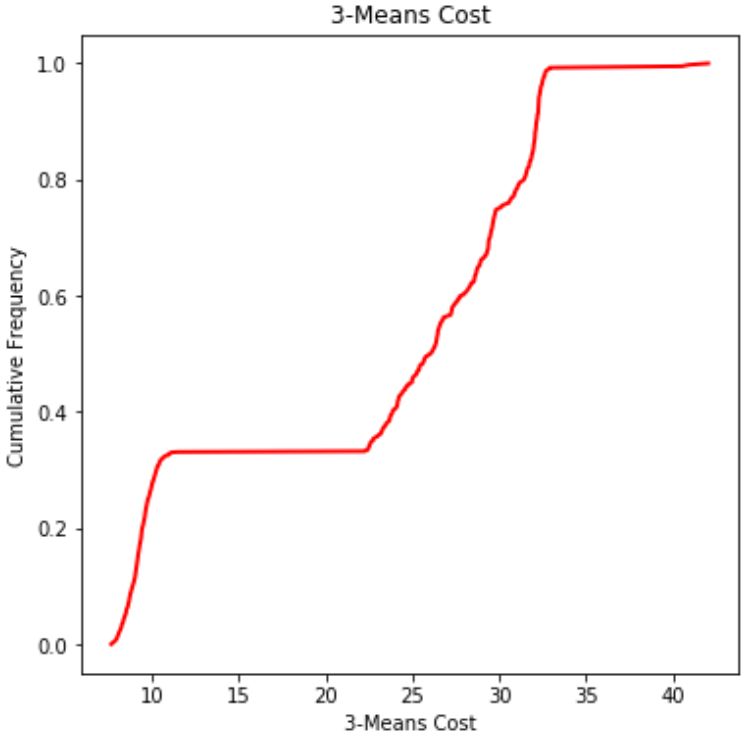

Figure 4: k-Means++ 3-Means Cost

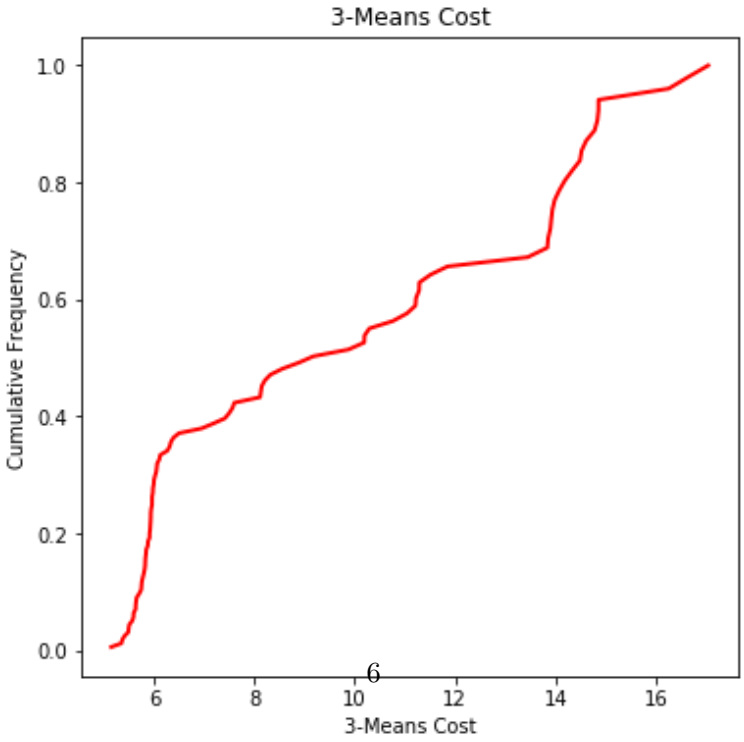

Figure 5: k-Means++ 3-Mean Cost with Loyd Algorithm Optimization

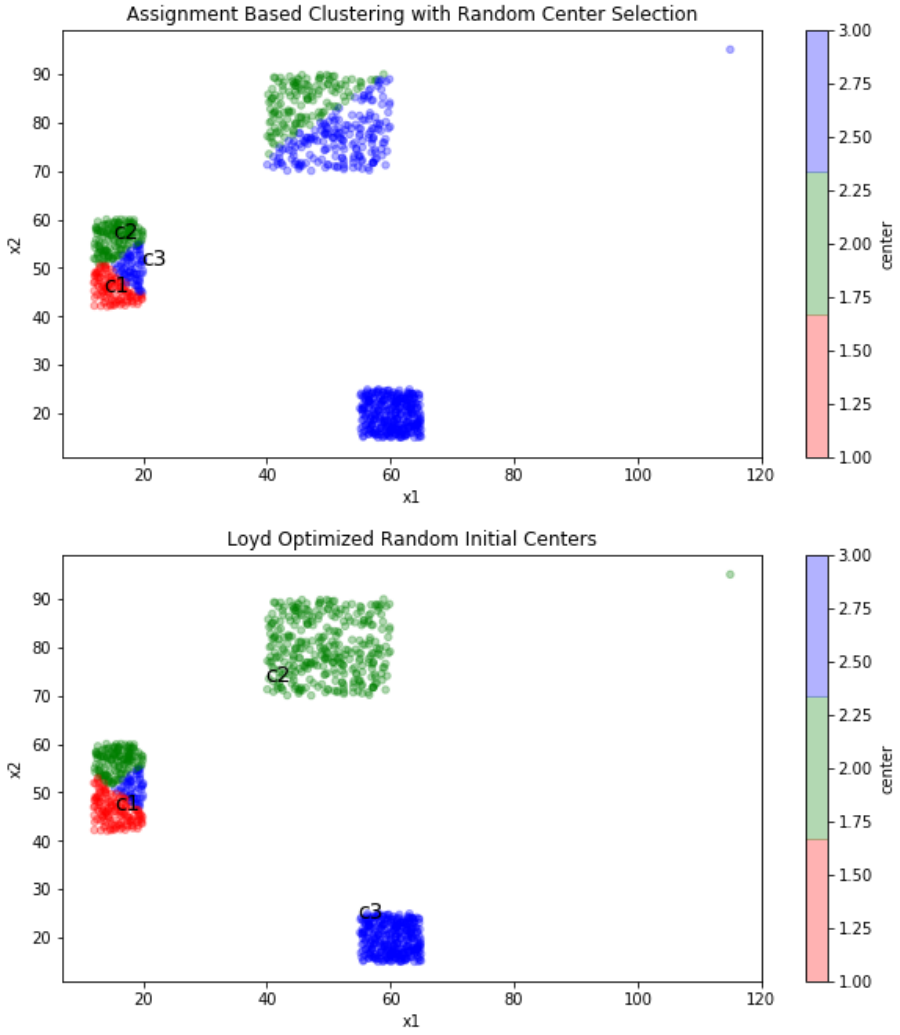

Figure 6: Randomly Selected Centers Pre-Post Loyd Algorithm Optimization

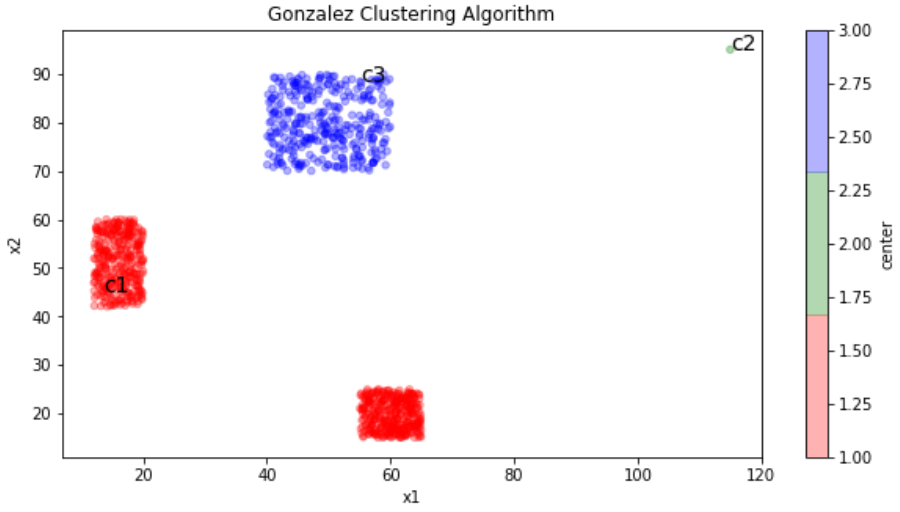

Figure 7: Gonzalez Algorithm Clustering

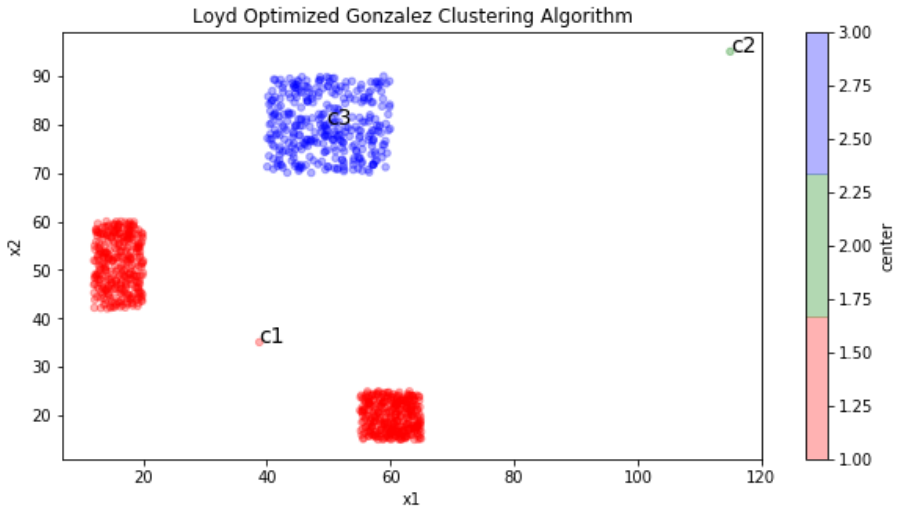

Figure 8: Gonzalez Clustering Algorithm with Loyd Algorithm Optimization