Report

Ilter Taha Aktolga

21.12.2020

1 Part 1: K-Nearest Neighbor

1.1 K-fold Cross-validation

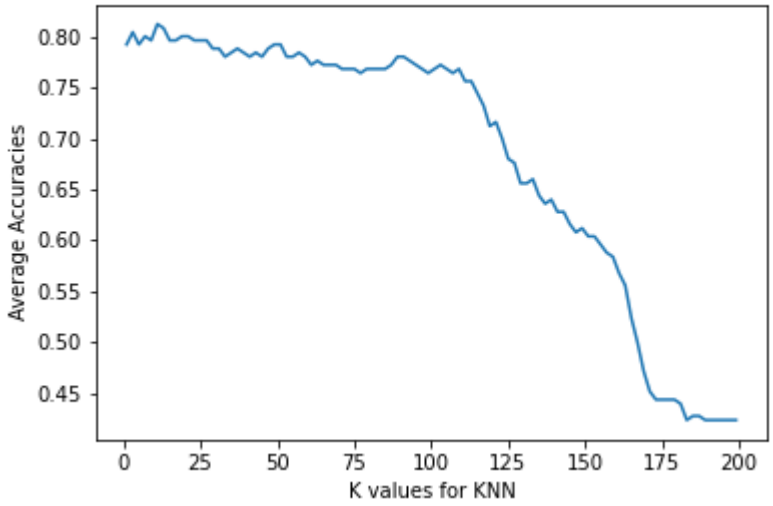

Figure 1: K vs Average Accuracy Plot

1.2 Accuracy drops with very large k values

In KNN, we ask K nearest neighbor to decide the class of the incoming data. When we increase the K values, algorithm is evolving from "deciding with a few neighbor" to "deciding with all dataset". At this point, since we use **majority vote** as a class label, meaning of the *nearest* starts to disappear. In other words, although the distances between some training samples are far away from the test data, we still consider these "far away samples" while deciding the labels. In our data, when we increase K, we see that newcomer data takes the majority label of the train set which reduces the accuracy drastically.

1.3 Accuracy on test set with the best k

I have tested K values for KNN with odd values from 1 to 199. I have got the best cross validation accuracy with K=11. Cross validation(10 fold) accuracy was 0.8119999. The test accuracy of KNN with K=11 was 0.825.

2 Part 2: K-means Clustering

2.1 Elbow method

Plotted k vs final objective values according to the description in the homework on each clustering data can be seen in this section. Note that, for each value of K from 1 to 10, K-means tested restarting 10 times and averaging their objective function values.

For clustering 1, best K values is selected as 2 from the plot below.

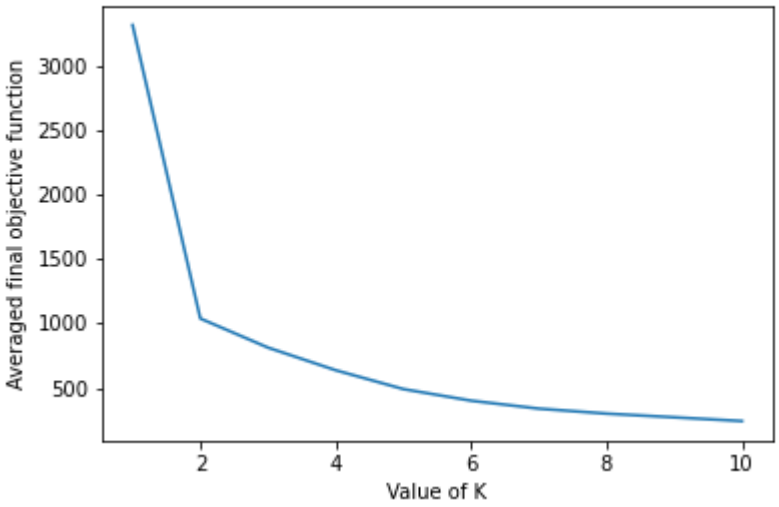

Figure 2: Selected K=2 for elbow method on clustering1

For clustering 2, best K values is selected as 3 from the plot below.

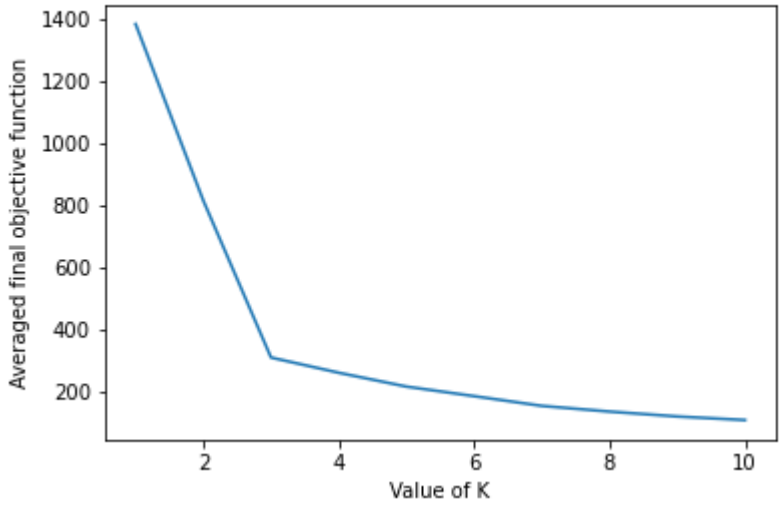

Figure 3: Selected K=3 for elbow method on clustering2

For clustering 3, best K values is selected as 4 from the plot below. However as can be seen that when K=2 , the K-Means also perform very well. Thus, K=2 can be selected also.

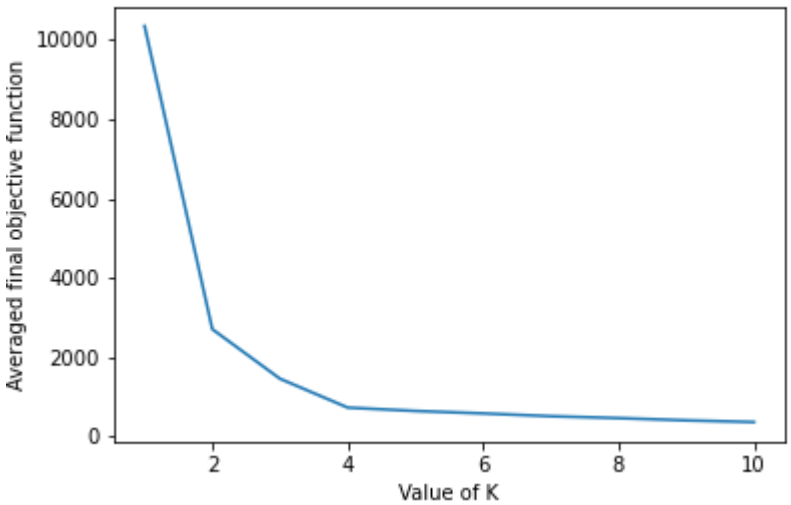

Figure 4: Selected K=4 for elbow method on clustering3

For clustering 4, best K values is selected as 5 from the plot below. Notice that there is no significant difference between 5 and 6, one may select 6.

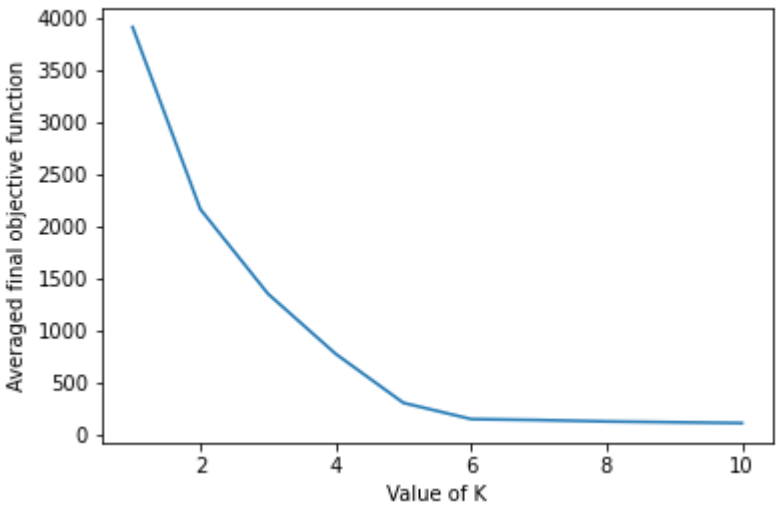

Figure 5: Selected K=5 for elbow method on clustering4

2.2 Resultant Clusters

Plots of the final clusters in each data according to the description in the homework can be seen below.

Result of the clustering with selected k = 2 using the elbow method can be seen below.

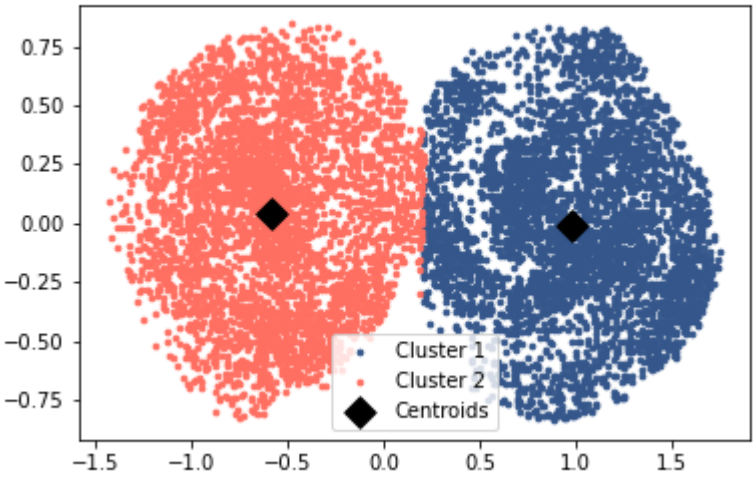

Figure 6: K-means with k=2 on clustering1

Result of the clustering with selected k = 3 using the elbow method can be seen below.

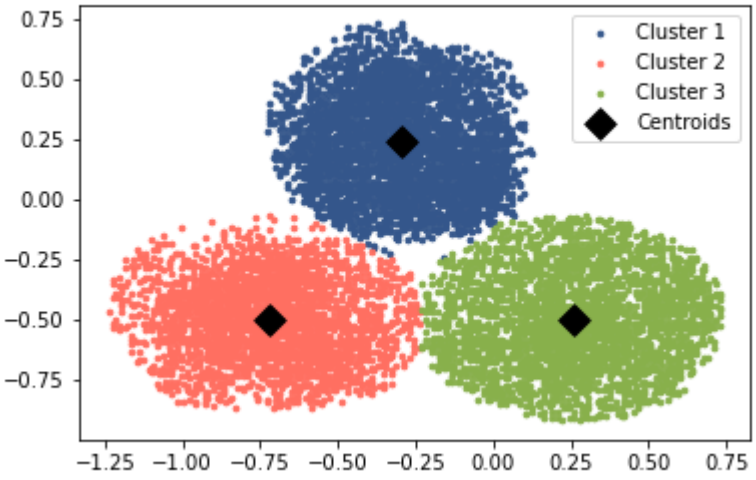

Figure 7: K-means with k=3 on clustering2

Result of the clustering with selected k = 4 using the elbow method can be

seen below.

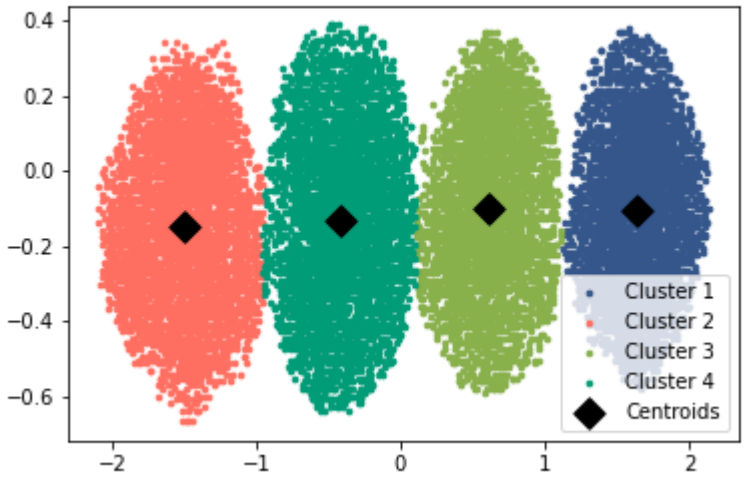

Figure 8: K-means with k=4 on clustering3

Result of the clustering with selected k = 5 using the elbow method can be seen below.

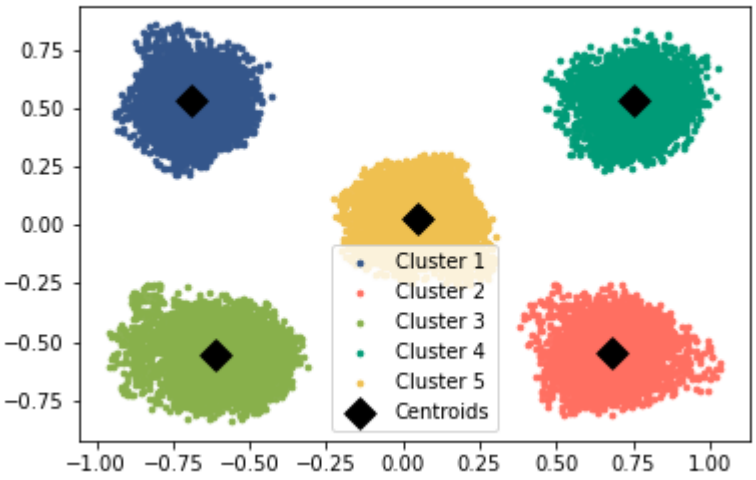

Figure 9: K-means with k=5 on clustering4

3 Part 3: Hierarchical Agglomerative Clustering

Note that related descriptions are written above the each plot.

3.1 data1

Since the single linkage looks at minimum distance between all inter-group pairs, it is expected to get the inside circle as a one cluster and the outside ring as another cluster. We know that for stringy clusters like wheel and round shapes, single linkage works as expected.

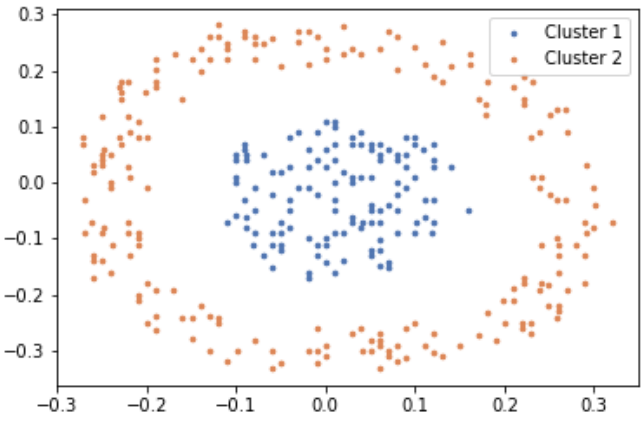

Figure 10: HAC on data1 with single linkage criterion stopped when 2 clusters left

From the plotted figure, it looks like individual clusters started to merge which are colorized as blue created a big cluster that their average distance to other clusters are too far. Therefore the remaining group with orange color created a huge cluster. Unlike in single linkage this behaviour does not serves good for visualization.

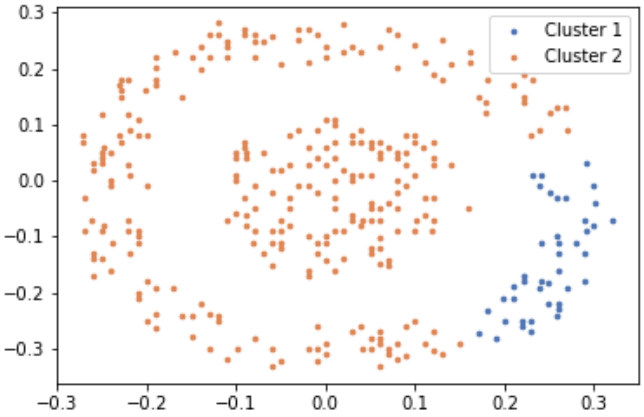

Figure 11: HAC on data1 with average linkage criterion stopped when 2 clusters left

Complete linkage tends to identify more compact objects. As we can see, they don't identify holes and separate objects and defines the inner circle's some of the samples as outer circle's elements. (Some of the elements in the inner circle is blue in the plot below)

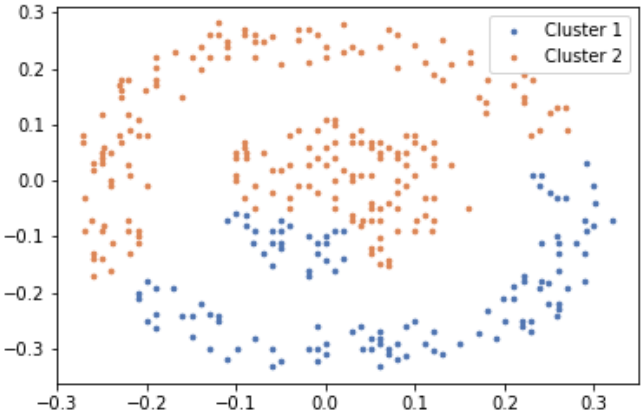

Figure 12: HAC on data1 with complete linkage criterion stopped when 2 clusters left

With each merging item, mean of the clusters are changing. After some time, it may be possible from the plot that orange cluster's center is somewhere

between center circular area and the left of the plot. So that oranges take the majority of the samples. So, the representation is not suitable.

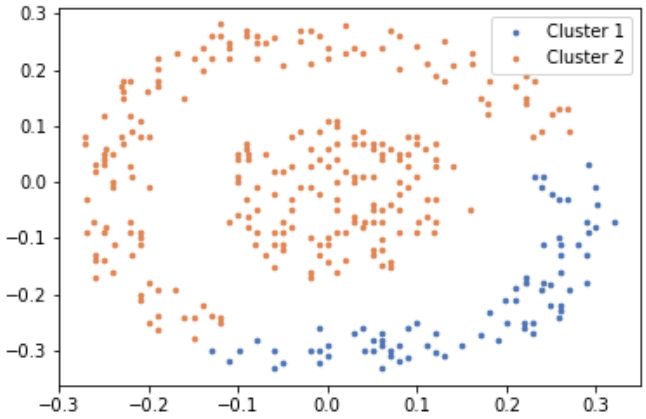

Figure 13: HAC on data1 with centroid linkage criterion stopped when 2 clusters left

3.2 data2

In the beginning, each sample is clustered with it's closest neighbors. Since the shape of two cluster is continuous, data points in the blue cluster are closer to the each other when compared with data points in orange cluster. Therefore there is no color sharing between these two cluster. Each cluster captured separately and colorized separately.

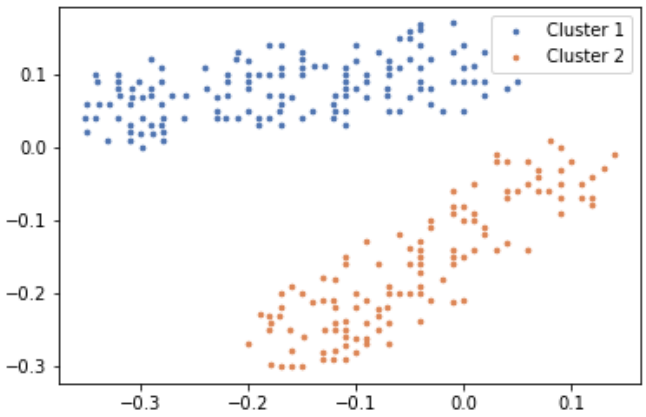

Figure 14: HAC on data2 with single linkage criterion stopped when 2 clusters left

Average linkage at first will behaves like single linkage, since in the beginning each sample is individual cluster. In blue points' region, there is no mid-cluster formed with points from orange cluster, since the average distance to orange clusters are far. Therefore it comes with output similar to single linkage.

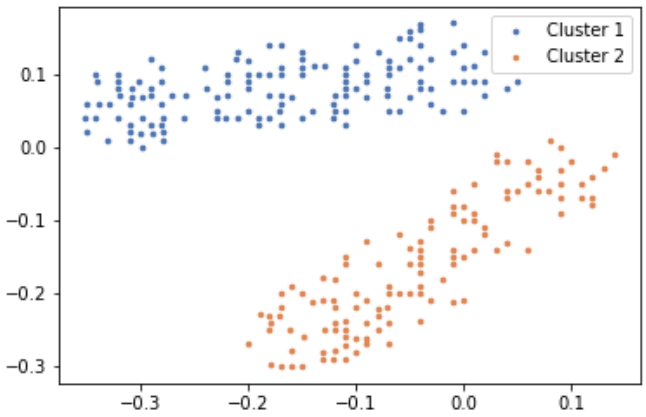

Figure 15: HAC on data2 with average linkage criterion stopped when 2 clusters left

According to complete distance, it may be possible that although the clusters seems like close to each other, since we are considering to merge them in terms

of farthest distance, distance(referring to complete distance criterion) from the top-left orange region to bottom-left orange region might be smaller than the distance between top-left orange region to top-right blue region. The reason why the each cluster doesn't have a unique color can be the result of the starting points of the first merging operations, because this behavior doesn't work like single criterion.

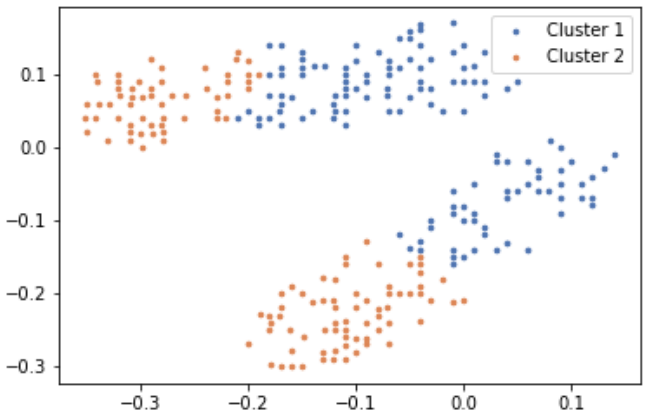

Figure 16: HAC on data2 with complete linkage criterion stopped when 2 clusters left

If the blue points' clustering started from the bottom-left part and if the orange part's centeroid moved to somewhere between top-center and top-right in the data plane, it may not be possible to merge orange points with the blue points in the below cluster. Therefore we have got the following result. This also doesn't give the cluster representation that we are looking for.

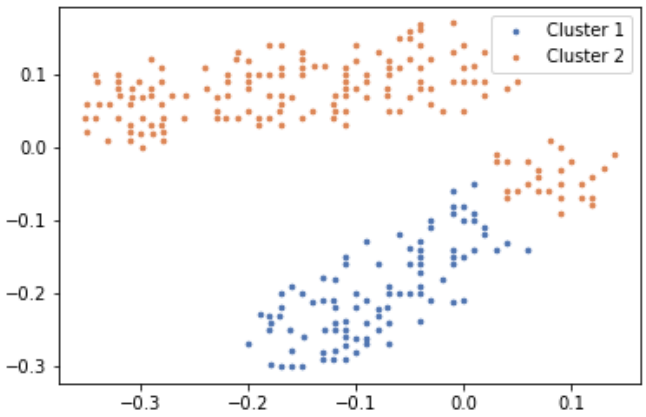

Figure 17: HAC on data2 with centroid linkage criterion stopped when 2 clusters left

3.3 data3

As a result of single linkage and because of the two clusters have a noticeable space between them, it is expected to get two different colored cluster. We can say that single linkage understands the borders of compact clusterings in the data.

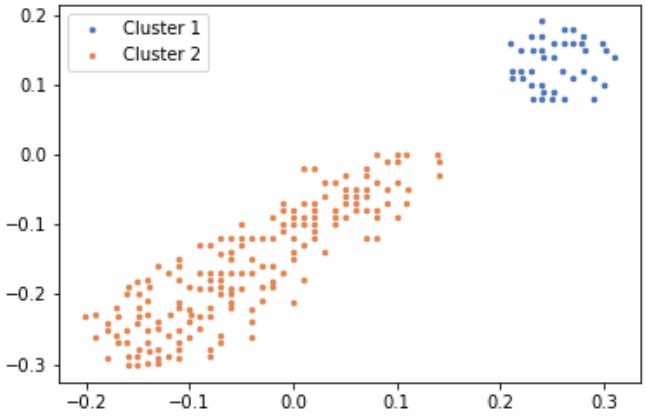

Figure 18: HAC on data3 with single linkage criterion stopped when 2 clusters left

As I mentioned in the previous data's, since each example is single cluster at the beginning, average linkage works like single linkage. It calculates distance and takes minimum and merges. Since two cluster points are compact and there is no holes in them, orange points and blue points are merged in itself. Average linkage performs good as single linkage in this case.

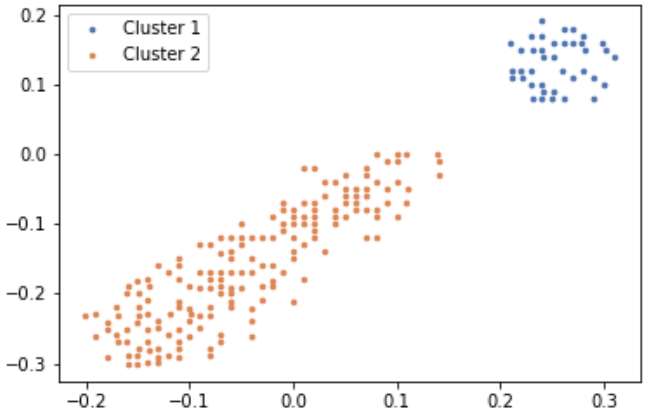

Figure 19: HAC on data3 with average linkage criterion stopped when 2 clusters left

It is possible that orange cluster started merging from top right and the blue cluster started merging from bottom left. During the merging, both are looking for the minimum of the farthest distance. Thus, even there is a noticeable space between orange cluster's data points, it continues to add points to it's cluster which were originally colorized as blue in the single linkage.

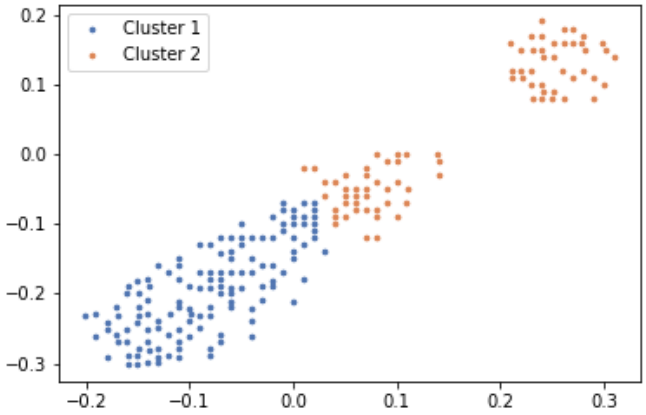

Figure 20: HAC on data3 with complete linkage criterion stopped when 2 clusters left

Since the distance between two cluster is large, during the merging process, points in the orange region cannot be assigned to blue region. When the small mid-clusters are becoming larger they tend to choose the nearest clusters with nearest centroids. If the data of the orange cluster had more space in between them, maybe it would be also possible to merge some orange's to blue cluster.

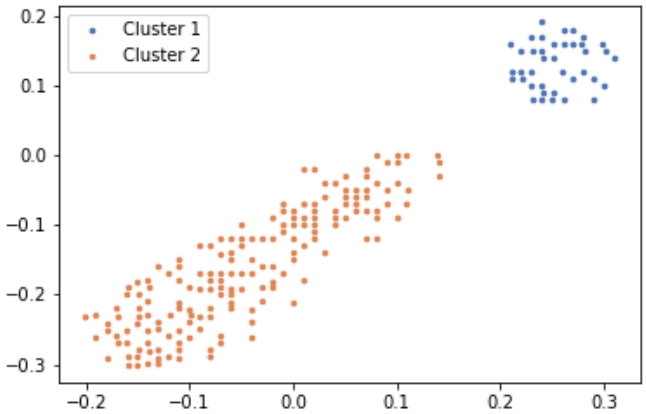

Figure 21: HAC on data3 with centroid linkage criterion stopped when 2 clusters left

3.4 data4

In single linkage, initally selected clusters (yellow,blue and green points) were become too distanced to the rest of the data. Therefore the red cluster dominated the data points. This was not the result we wanted.

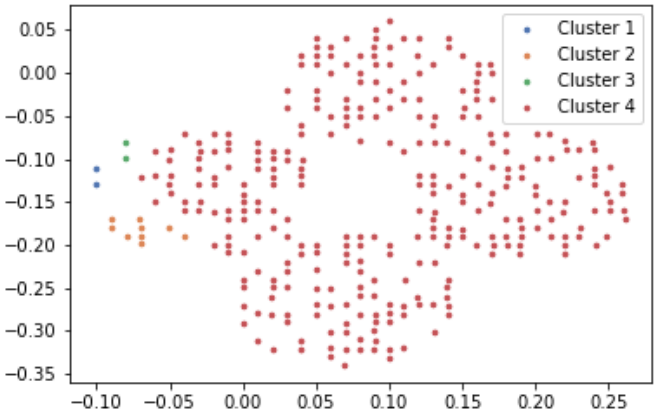

Figure 22: HAC on data4 with single linkage criterion stopped when 4 clusters left

There are four circular data regions in original data. Since the criterion selects to merge by average distance, it will not dominate the whole points in the data plane. When the small clusters are becoming merged to create bigger clusters, average distance between each circular area will be quiet far to each other. For instance green and blue points at the borders of each circular areas is close to each other. However, the criterion does not merge them, since there are other points in both clusters that are far from each other. Criterion is suitable

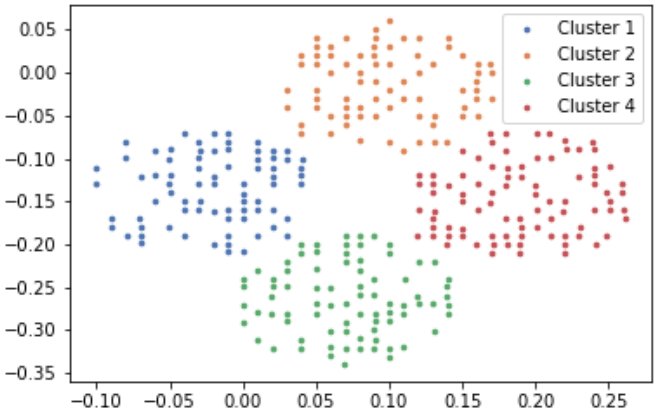

Figure 23: HAC on data4 with average linkage criterion stopped when 4 clusters left

We saw in single linkage that some of the cluster dominates the whole points. And also we noticed that average linkage stopped this domination. Therefore it is quiet possible that farthest distance has an effect on clustering. We can see that the result is better than single linkage and a bit worse than average linkage. The reason why some of the blue points are marked as red may be because of the fact that at the time of selection, farthest distance between big red cluster and that points' cluster was selected. (Unlike in the average criterion)

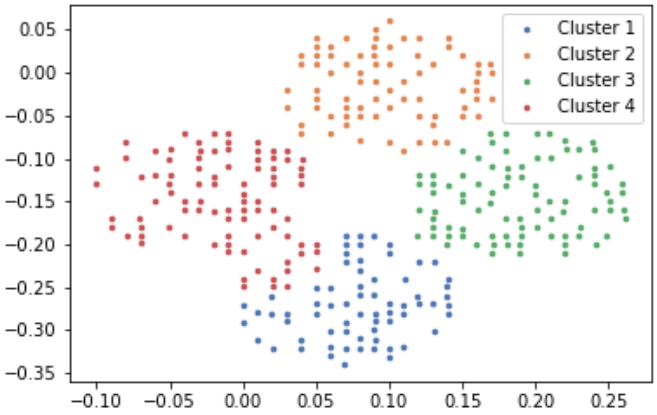

Figure 24: HAC on data4 with complete linkage criterion stopped when 4 clusters left

Distances between centroids behave like average in this data, since there are 4 different and spaced clusters. When clusters becoming bigger during the clustering process, they are less tend to choose points far from their centers. It is the result of the following plot. It is the desired plot. This criterion is suitable.

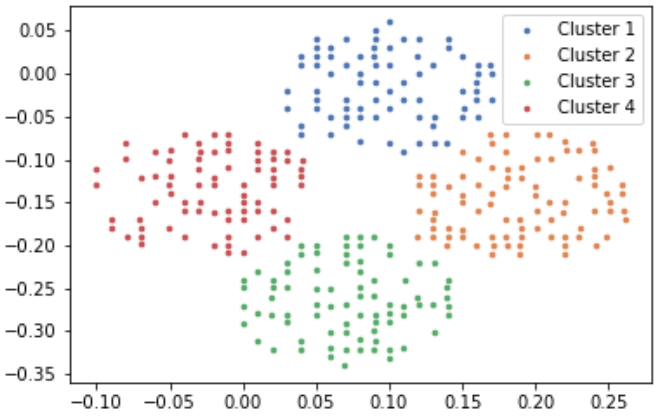

Figure 25: HAC on data4 with centroid linkage criterion stopped when 4 clusters left