CENG499 Hw2 Report

Ali Dogan

December 19th 2020

1 Part 1: K-Nearest Neighbor

1.1 K-Fold Cross-Validation

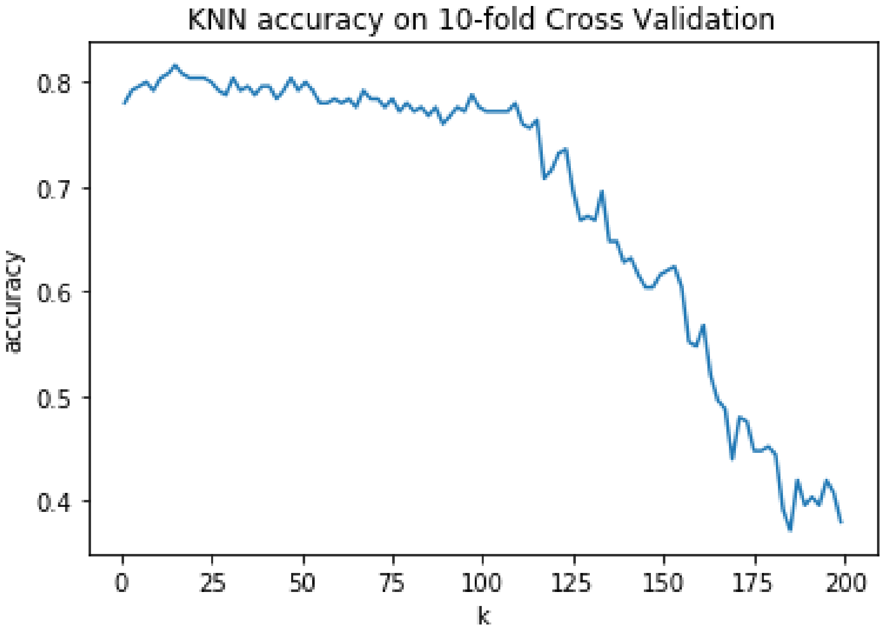

Figure 1: KNN cross validaton

1.2 Accuracy drops with very large k valuse

KNN algorithm works by the locality principle of the data. In other words, we get an insight about the given data looking by its neighbours assuming that there is some sort of clustering in the data. So, as the number of the neighbours increase, we are looking more data around it and after some point, we are far from that locality or cluster, i.e, we are losing the information we gain from that address, point. And when the number k reaches at 200, we are basically looking at all the data except the furthest 50, so it gives very few information.
I found the best k value to be 7 using the 10-fold cross validation. With k being 7 and using training data, I achieved 0.78 accuracy on test data.

2 Part 2: K-means Clustering

2.1 Elbow Method

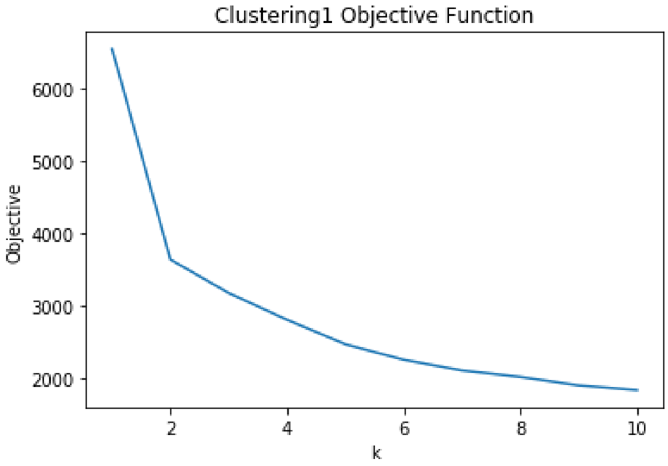

Figure 2: Cluster1 Obj function shows k = 2 is the appropriate choice

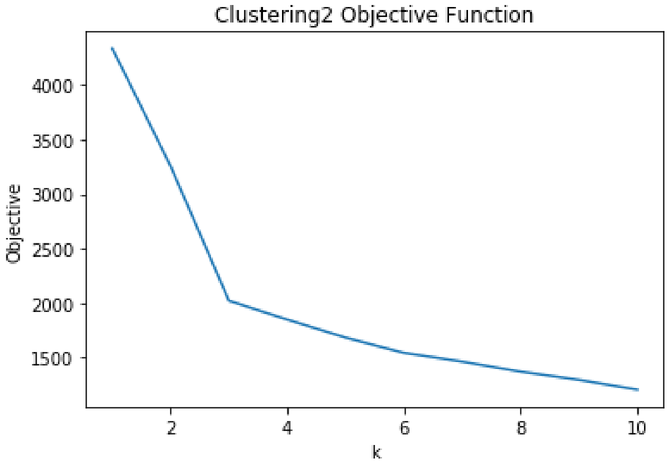

Figure 3: Cluster2 Obj function shows k = 3 is the appropriate choice

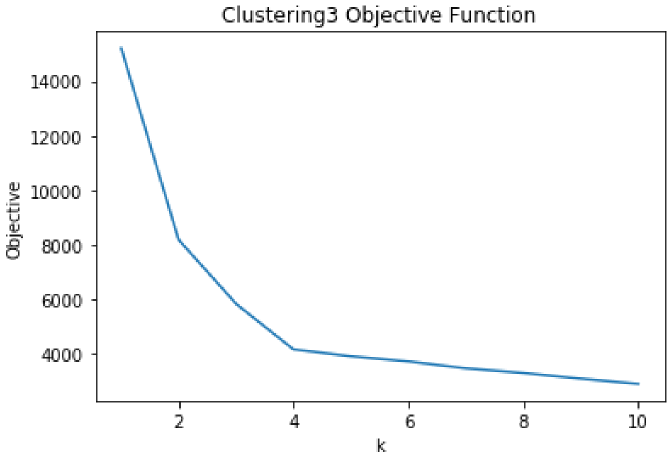

Figure 4: Cluster3 Obj function shows k = 4 is the appropriate choice

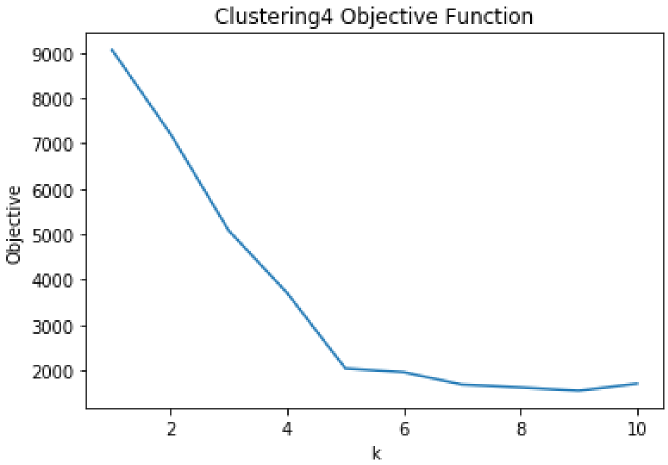

Figure 5: Cluster4 Obj function shows k = 5 is the appropriate choice

2.2 Resultant Clusters

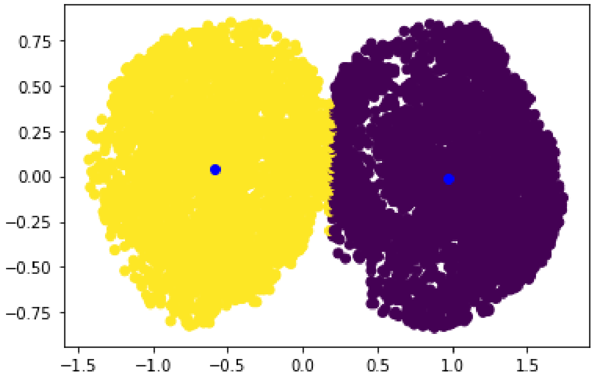

Figure 6: Cluster1 with k = 2

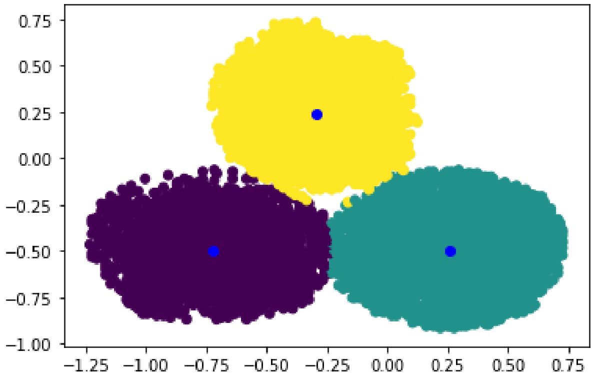

Figure 7: Cluster2 with k = 3

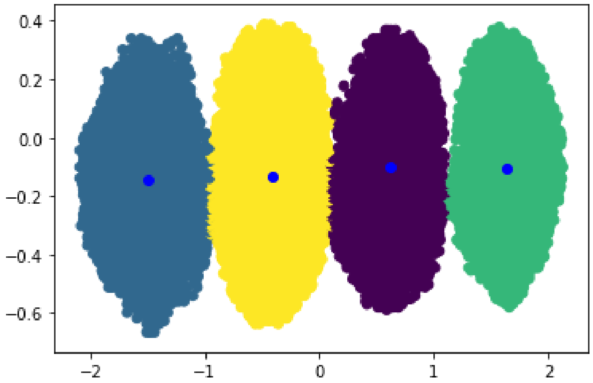

Figure 8: Cluster3 with k = 4

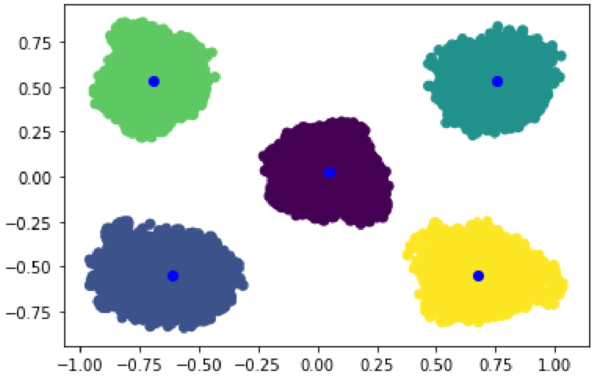

Figure 9: Cluster4 with k = 5

3 Part 3: Hiearchical Agglomerative Clustering

3.1 data1

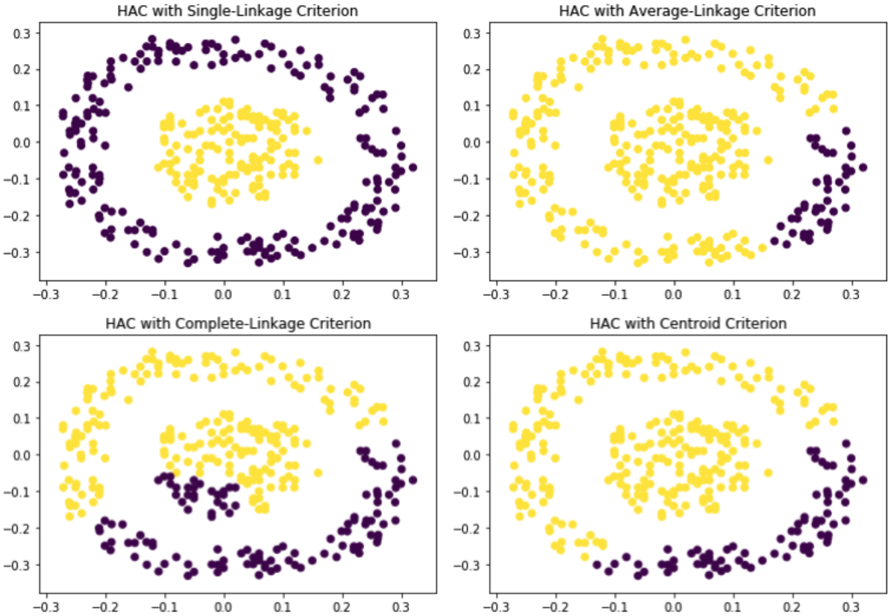

Figure 10: Clusterings with Single Linkage

The Single-linkage worked as expected. Others however did not give the best results. The main issue is the fact that one cluster is surrounding the other. Because of this as we merge the clusters with centroid, the centroid of the outer cluster will be nearly as same as the inner one. In the average linkage, the problem is that as the clusterings occur the average distance from one cluster located in out and another one again located in out tend to increase even though they are neighbours because the cluster at the outside has a rectangle(when consider only a part of circle) -circle shape. In the complete-linkage, that problem also applies.

-

3.2 data2

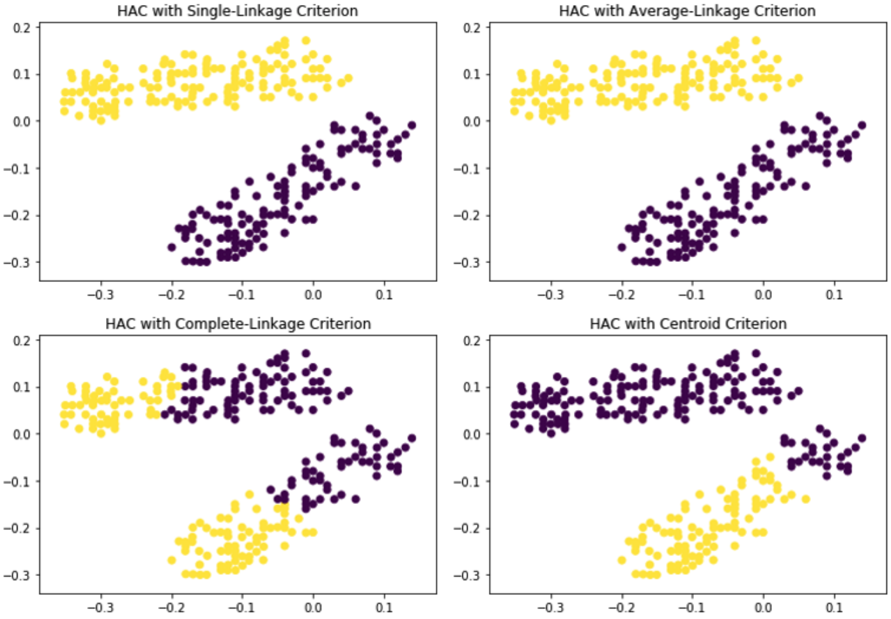

Figure 11: Clusterings with Complete Linkage

The behaviour of complete linkage criterion causes problems when the clusters do not have a compact, more dense clustering especially when the clusters distance from each other are not far enough. In other words, the problem can be explained better below. This is the clusters before couple of steps in the data2:

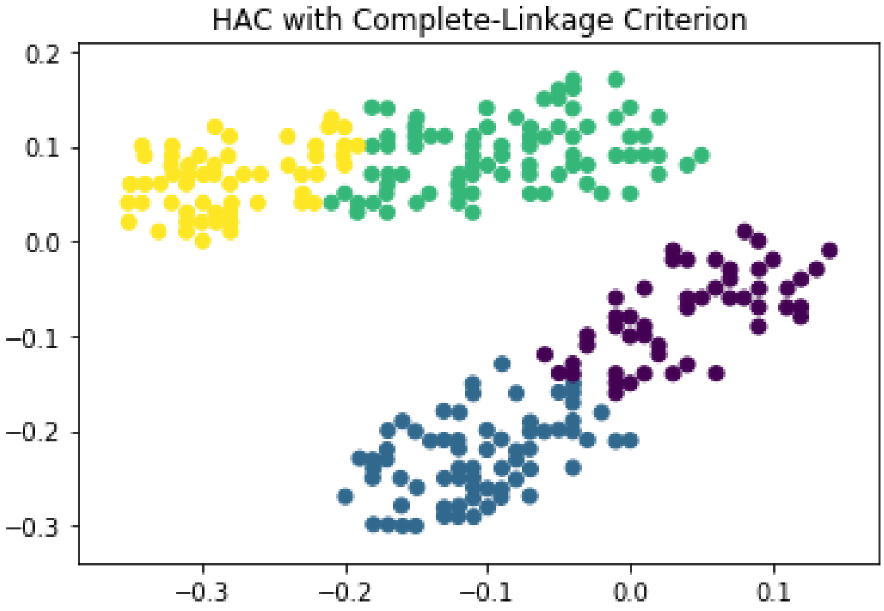

Figure 12: Clusterings with Complete Linkage

In the plot above, there are four clusters. With complete linkage, in the next step, the green and purple clusters are merged because their furthest two points are the smallest among any other two cluster pair.

3.3 data3

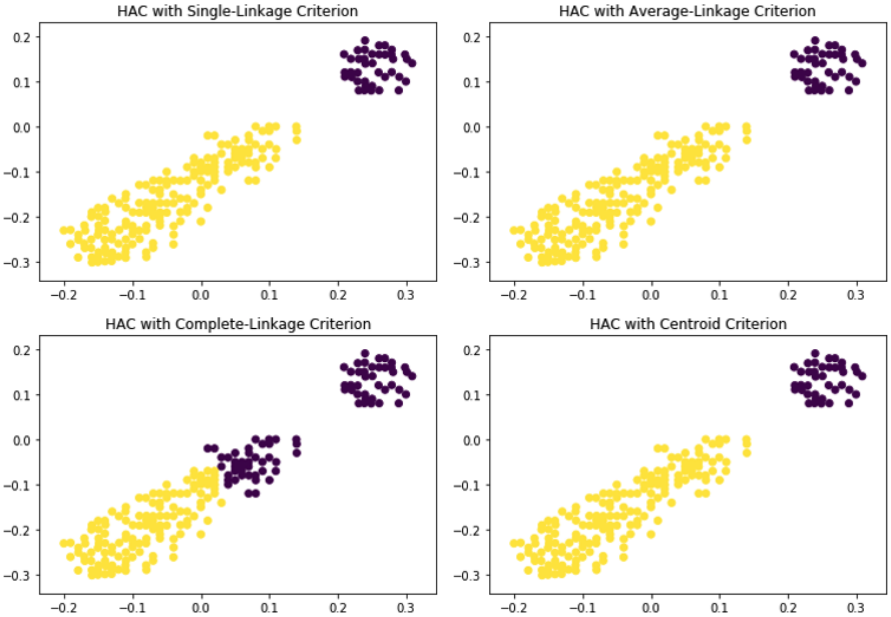

Figure 13: data3 Dlusterings

data3 is suitable for all criterions except the complete linkage. The problem with complete linkage is that the complete linkage do not depend on the size of the cluster, we only look at the furthest two points between clusters and select the minimum one. At the last steps, when using complete-linkage, the distance between the circle cluster and upper body of the rectangle cluster is smaller than between the upper body of the rectangle and bottom of the rectangle cluster.

3.4 data4

Centroids have the same problem as in the average-linkage.

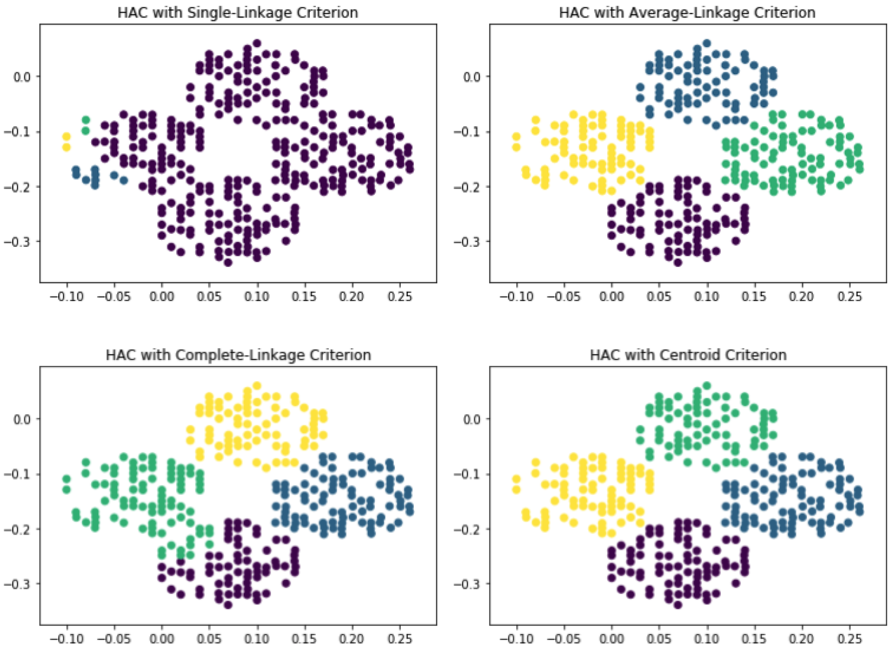

Figure 14: data 4 Clusterings

Single-linkage criterion resulted in a very poorly clustering. No matter how big is the two clusters in terms of its members, single linkage criterion only look for the closest two points between the clusters. This is not a problem when the clusters have large distances with each other and there is no outliers in clusters. This is the case for the first 3 data. However, the clusters in data4 are close to each others in a way that some two points from different clusters has a smaller distance than two point at the same cluster.