Report

Ahmet Berker KOÇ

18.05.2021

1 Part 1: K-Nearest Neighbor

1.1 K-fold Cross-validation

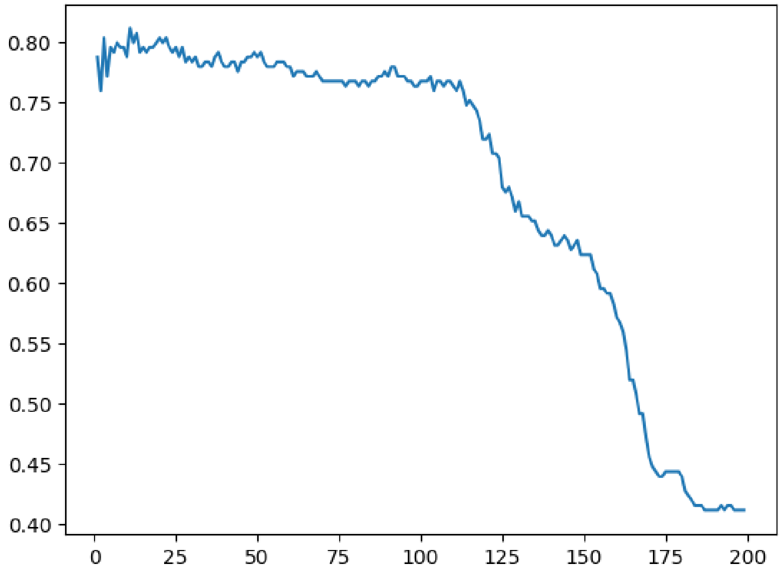


Figure 1: Accuracy vs k graph for K-fold Cross Validation

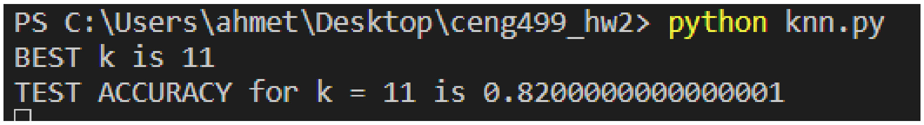
1.2 Accuracy drops with very large k values

High value of k is against the basic principle of knn (that data (points) that are near might have similar classes). Although points are not near, they might have similar class because of the high value of k. To better understand and interpret, let's think about the extreme case k is equal to the number of train data. In this case, all new data are predicted same as most existing label in the data.

Also Choosing a small k values cause an unstable decisions. Effect of noise become high

1.3 Accuracy on test set with the best k

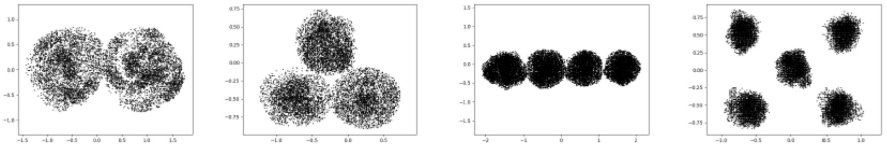
I print the best k value which gave us the best accuracy. Then I test it with test dataset. I obtain the 0.82 test accuracy as shown in Figure2



```
PS C:\Users\ahmet\Desktop\ceng499_hw2> python knn.py
BEST k is 11
TEST ACCURACY for k = 11 is 0.8200000000000001
```

Figure 2: Result for the best k value

2 Part 2: K-means Clustering



(a) clustering1 (b) clustering2 (c) clustering3 (d) clustering4

Figure 3: Clusters from homework document

Before experiment, I examine Figure 3. My expectation for the most suitable k values for each cluester is as follows. This value is seen clearly in the Figure3

1. k=2 for cluster1
2. k=3 for cluster2
3. k=4 for cluster3
4. k=5 for cluster4

2.1 Elbow method

In order to choose a most suitable k value. We will use elbow method which is one of the most populer method to find optimal number of clusters. In order to use elbow method, I find the objective value for each k (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) and I plot the graph. Where the elbow point of this graph appears is right k. Objective converge a value after this k value i.e. we expect the graph to flatten on the horizontal axis after that value. Sometimes, this elbow point is not obvious but the first point where the objective function converge is the right point.

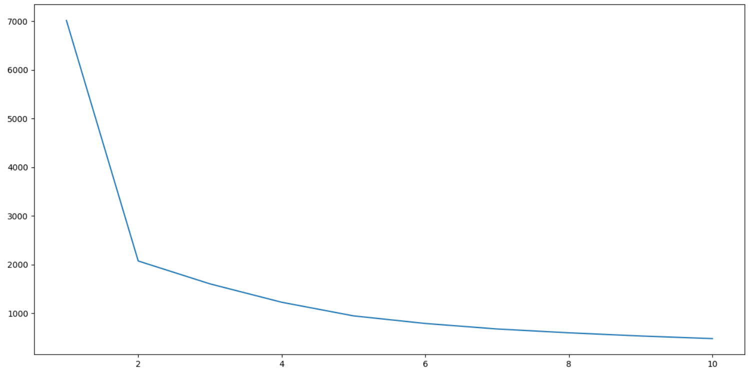


Figure 4: Objecitive vs k Graph for Elbow method for data1

- Clusturing1 objecitive vs k graph is shown in Figure 4. Elbow point is easily seen as k=2 in the figure. This k value is same as our expectation

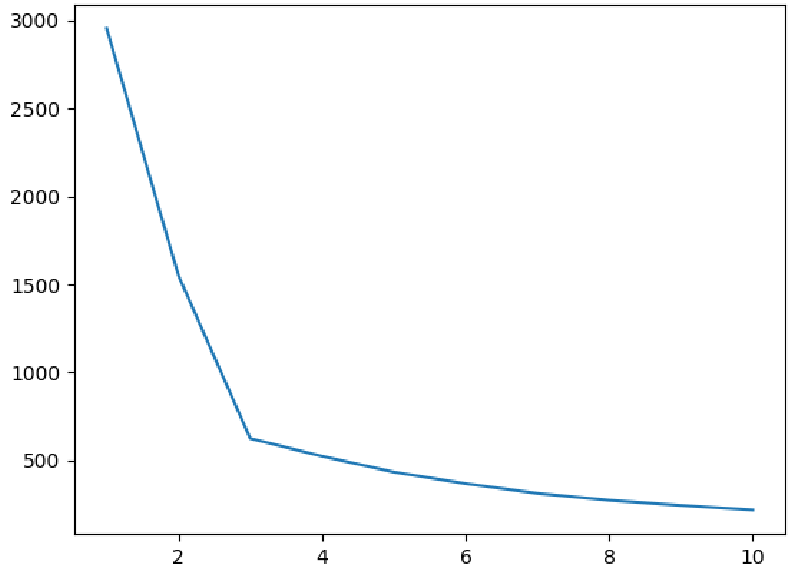


Figure 5: Objecitive vs k Graph for Elbow method for data2

- Clusturing2 objecitive vs k graph is shown in Figure 5. Elbow point is easily seen as k=3 in the figure. This k value is same as our expectation

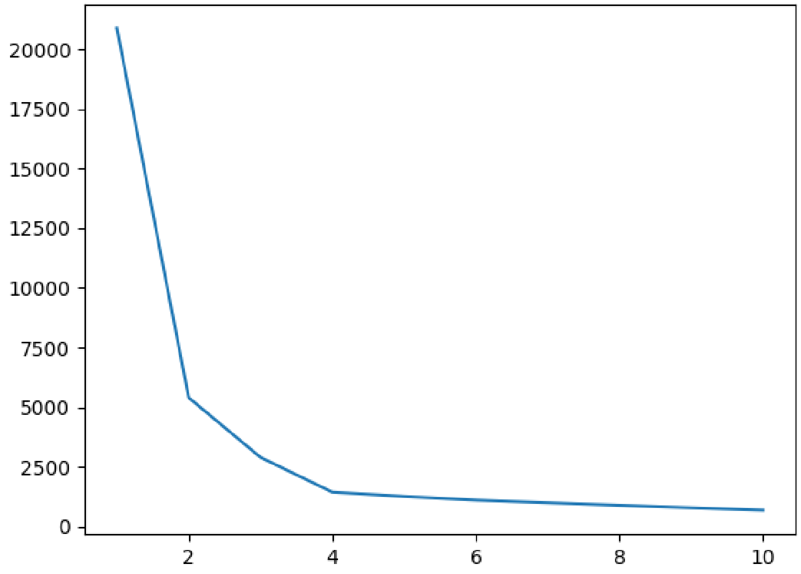


Figure 6: Objecitive vs k Graph for Elbow method for data3

- Clusturing3 objecitive vs k graph is shown in Figure 6. Elbow point is easily seen as k=4 in the figure. This k value is same as our expectation

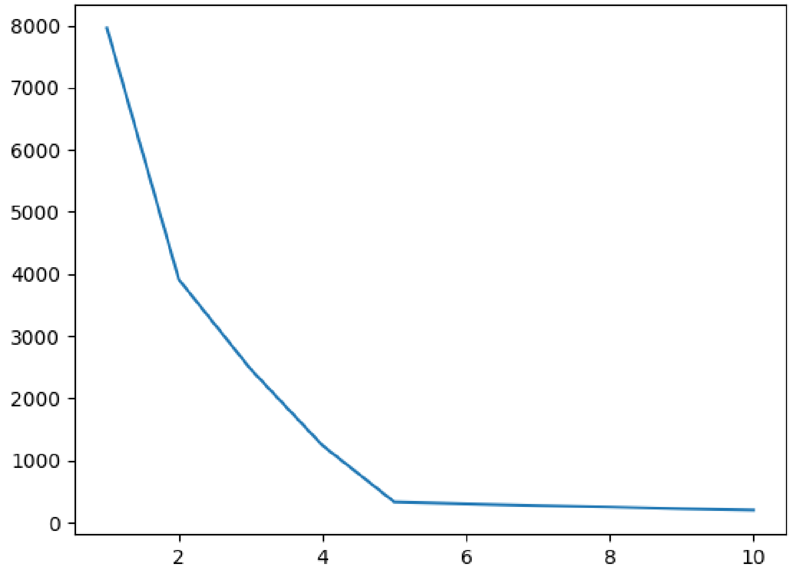


Figure 7: Objecitive vs k Graph for Elbow method for data4

- Clusturing4 objecitive vs k graph is shown in Figure 7. Elbow point is easily seen as k=5 in the figure. This k value is same as our expectation

2.2 Resultant Clusters

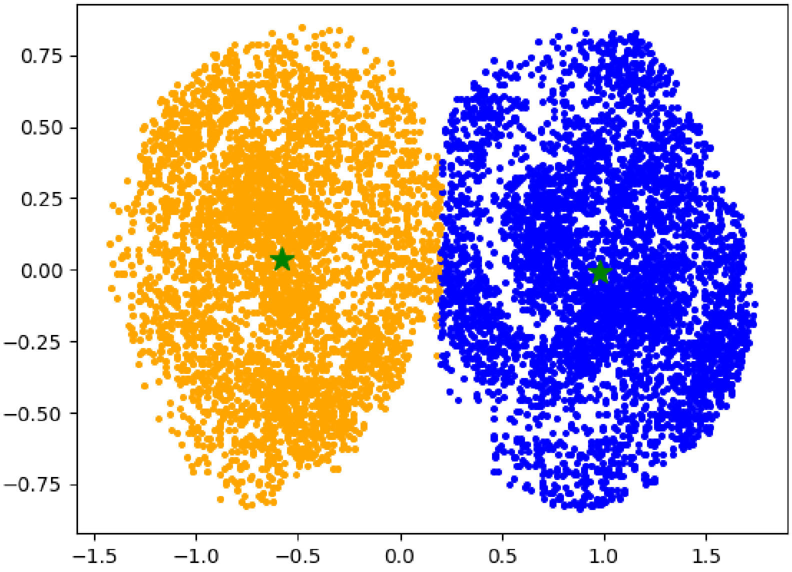


Figure 8: Final clusters and center points for data1

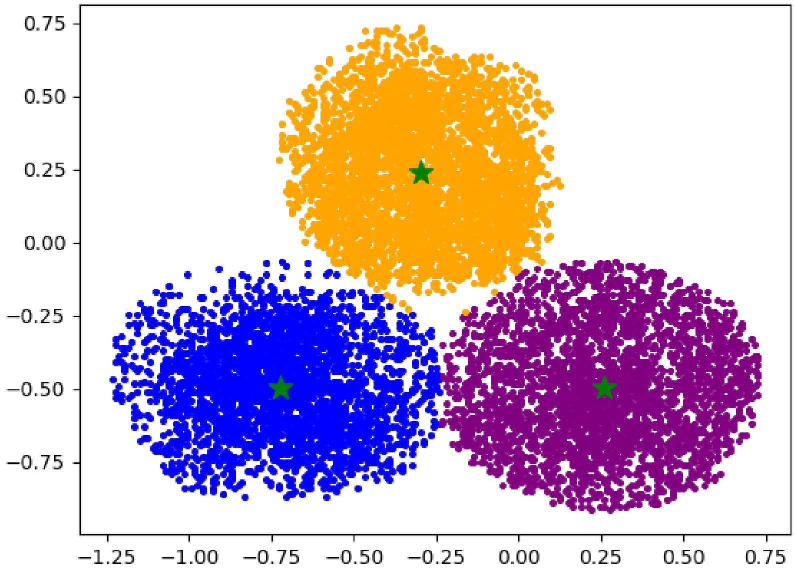


Figure 9: Final clusters and center points for data2

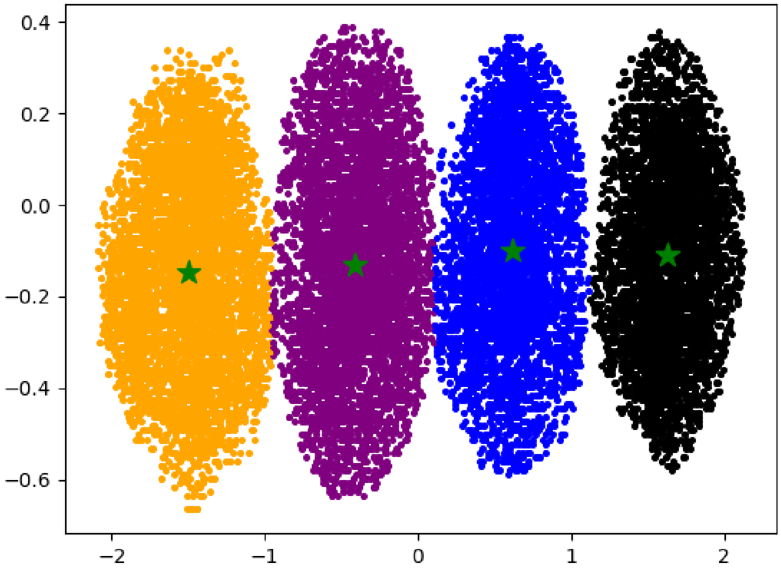


Figure 10: Final clusters and center points for data3

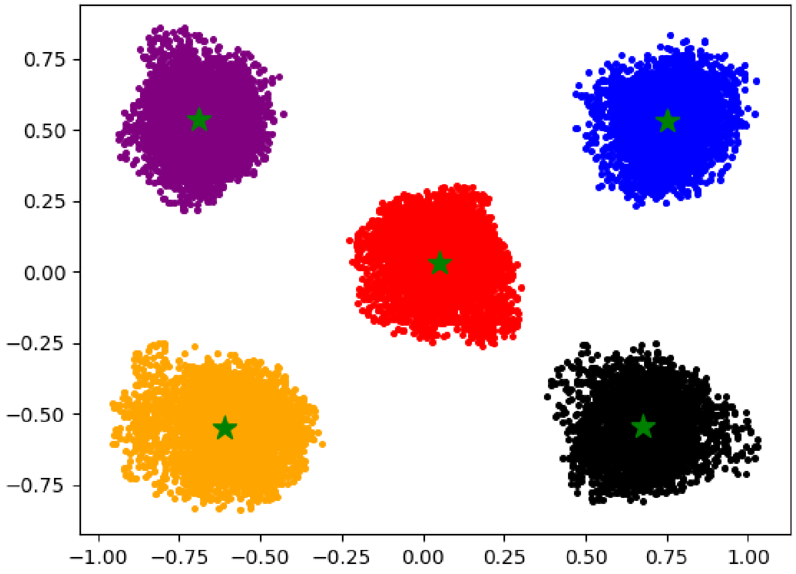


Figure 11: Final clusters and center points for data4

3 Part 3: Hierarchical Agglomerative Clustering

In this part, our aim is to take two closest clusters and make them one cluster. The distance between clusters are calculated for diffrent criterion as follows. The most popular two are complete linkage and average linkage. According to the distribution of data, most suitable criterion is changed.

1. Single-Linkage
   Single linkage or nearest neighbor looks at minimum distance between all inter-group pair

2. Complete-Linkage
   Complete-linkage or farthest neighbor looks at maximum distance between all inter-group pair

3. Average-Linkage
   Average linkage uses the average distance between all inter-group pairs

4. Centroid-Linkage
   Centroid linkage first computes the centroid of each group and then looks at the distance between them.

3.1 data1

In this part final number of cluster is chosen as 2

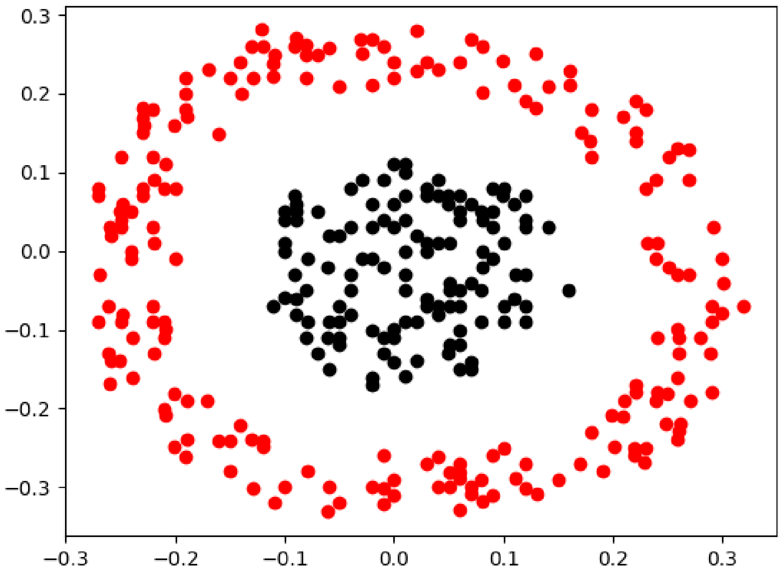


Figure 12: single linkage for data1

Comment1: Single linkage criterion is suitable for dataset1. As seen in Figure 12, the data is clustered in a logical way. Since data that are close to each other are clustered, the data in the middle and other data distant to them are clustered in two groups.

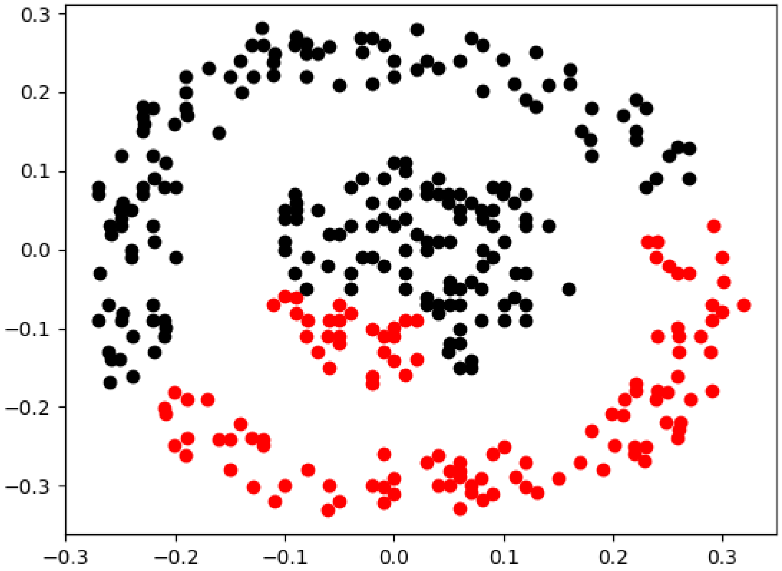


Figure 13: complete linkage for data1

Comment2: Complete linkage criterion is not suitable for dataset1. As seen in Figure 13, clustring is not very meaning full. Since complete linkage works with minimum of most far elements of clusters, it creates more compact cluster comparing with single linkage. It requires all of the distances to be small.

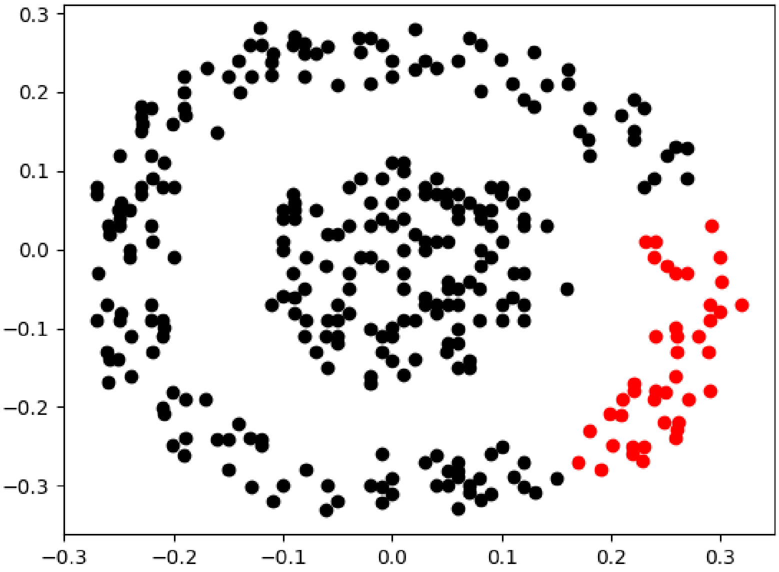


Figure 14: average linkage for data1

Comment3: Average linkage criterion is not suitable for dataset1. As seen in Figure 14, there is cluster in cluster and this causes a bad result for average linkage criterion on dataset1

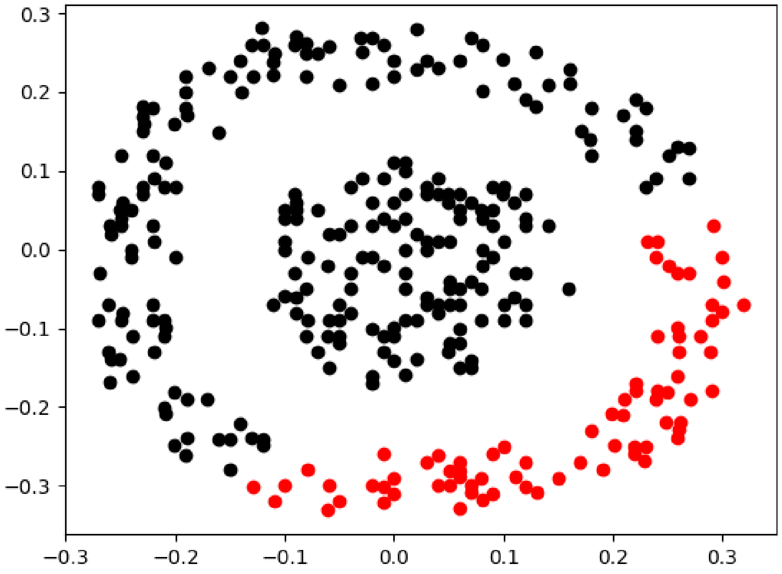


Figure 15: centroid linkage for data1

Comment4: Centroid linkage criterion is not suitable for dataset1. The reason is same as the average linkage. As seen in Figure 15, there is cluster in cluster and this causes a bad result for centroid linkage criterion on dataset1

3.2 data2

In this part final number of cluster is chosen as 2

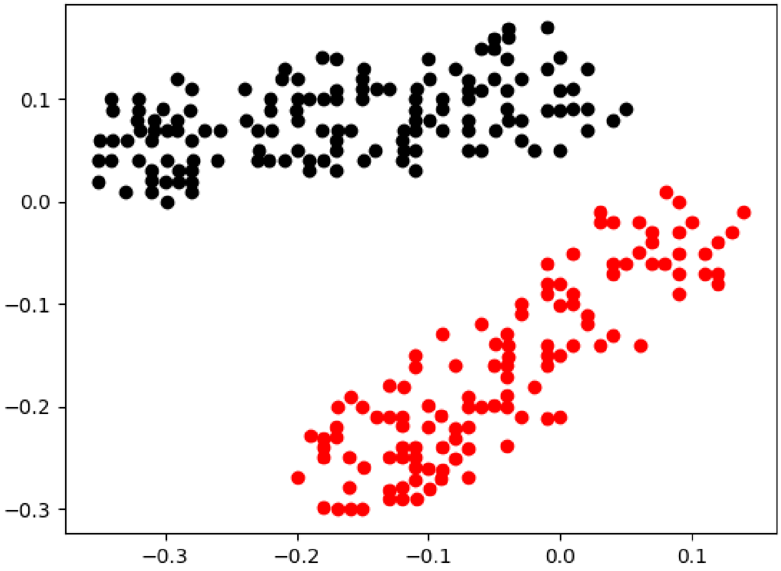


Figure 16: single linkage for data2

Comment1: Single linkage criterion is suitable for dataset2. As seen in Figure 16, the data is clustered reasonably. Since data that are close to each other are clustered, data in the right bottom and data in the left top are clustered

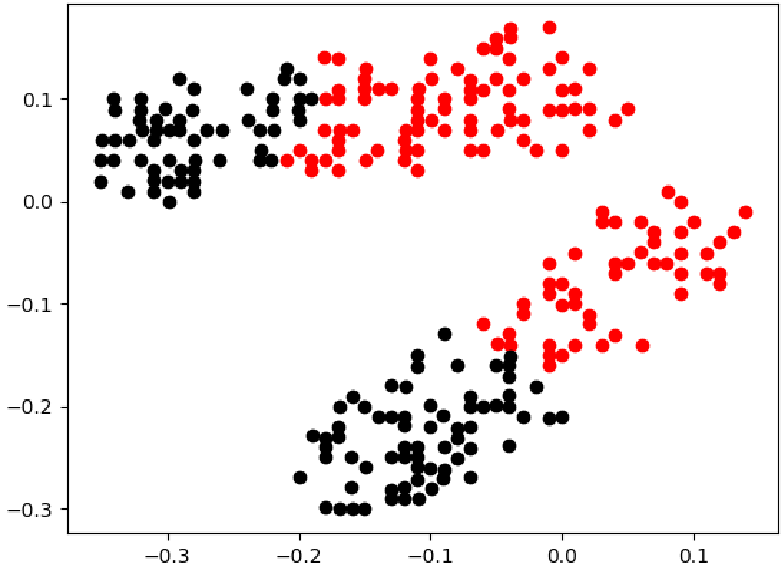


Figure 17: complete linkage for data2

Comment2: Complete linkage criterion is not suitable for dataset2. it creates more compact cluster comparing with single linkage. It requires all of the distances to be small. Therefore, instead of long thin rod shaped cluster, it create tigter clusters as seen in Figure 17

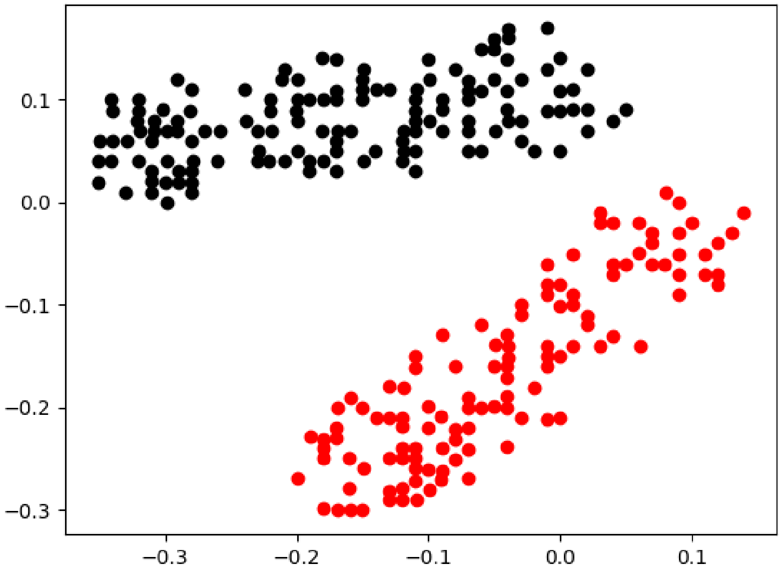


Figure 18: average linkage for data2

Comment3: Average linkage criterion is suitable for dataset2. As seen in Figure 18, like single linkage clusters, the data is clustered reasonably. Average distance between all inter-group pairs is a good approach for this dataset. There is no cluster in cluster unlike dataset1.

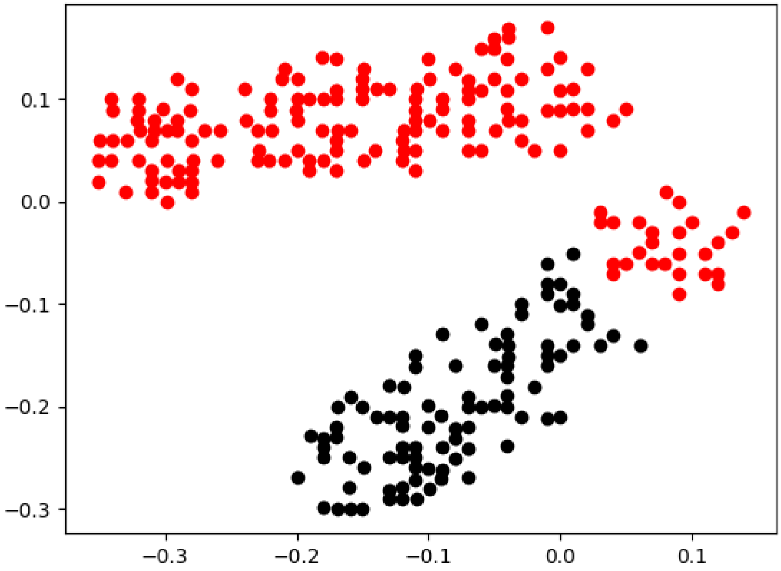


Figure 19: centroid linkage for data2

Comment4: Centroid linkage criterion is not suitable for dataset2. As seen in Figure 19, unlike single linkage clusters, center point of the top of the right part of the data (single linkage red cluster) is closer to top left data (single linkage black cluster). The space in the area where red data and black data are separated caused the distance between center point to be high. This high distance leads to bad clustering

3.3 data3

In this part final number of cluster is chosen as 2

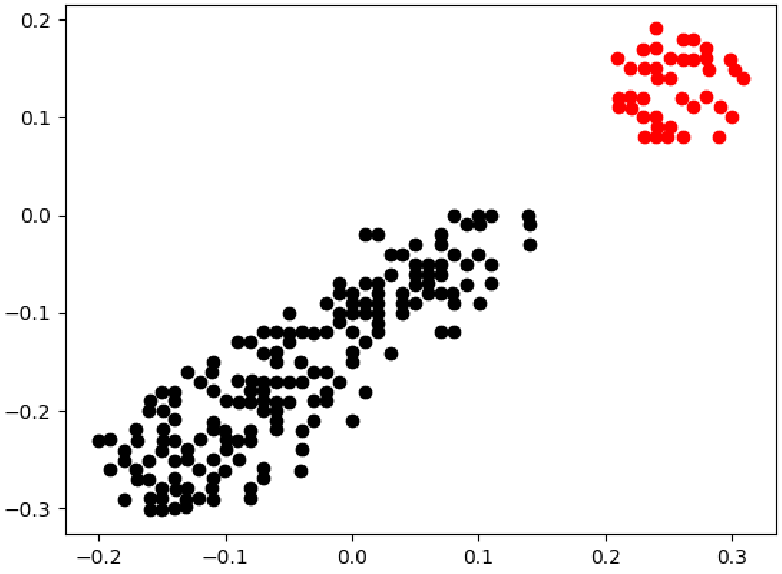


Figure 20: single linkage for data3

Comment1: Single linkage criterion is suitable for dataset3. As seen in Figure 20, the data is clustered reasonably. There is a big gap between the two data groups in the dataset3. This enable single linkage criterion to give good 2 clusters. Data for each clusters are very close to each other.

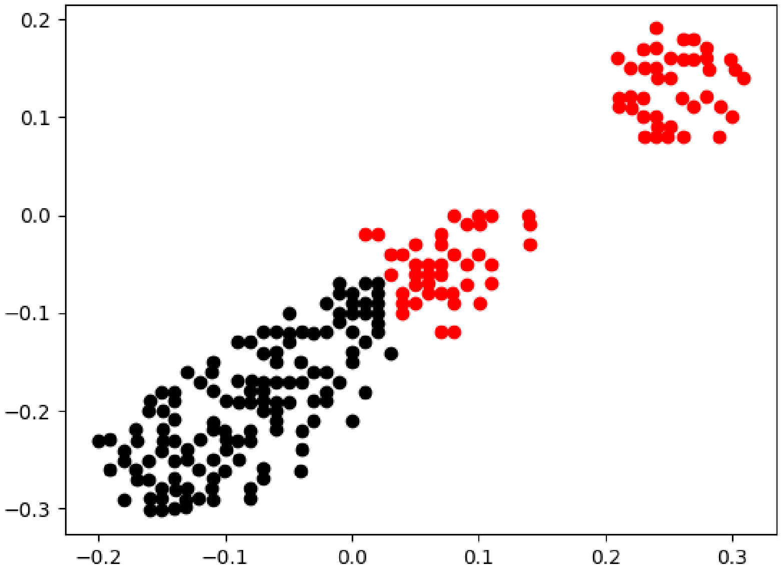


Figure 21: complete linkage for data3

Comment2: Complete linkage criterion is not suitable for dataset3. The farthest distance of black cluster is farther than that of red cluster (at the right top). Therefore, unlike single, average and complete linkage, the right end of the data remaining on the left belong to red cluster. The result is expected however, complete linkage criterion is not suitable for dataset3

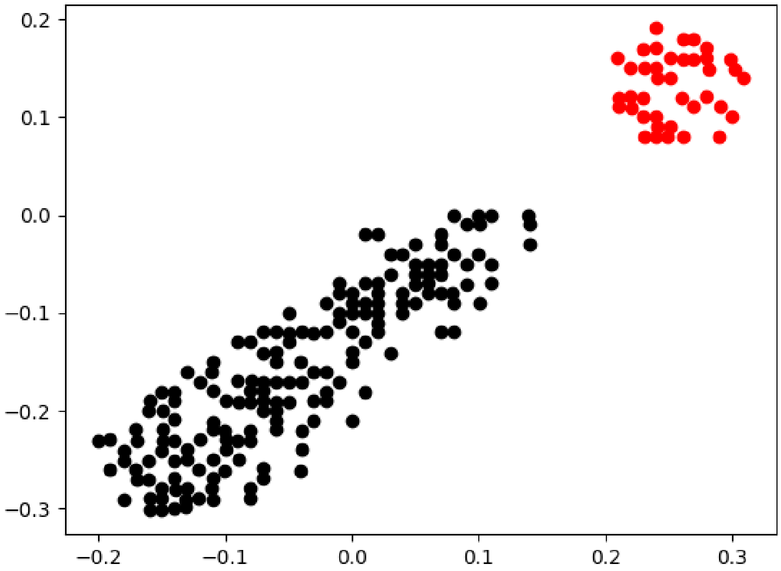


Figure 22: average linkage for data3

Comment3: Average linkage criterion is suitable for dataset3. As seen in Figure 22, the data are clustered reasonably. Average distance between all inter-group pairs is a good approach for this dataset. There is no cluster in cluster unlike dataset1. Data of two clusters are far from each other. There is a big gap between black and red data. Therefore, when taking average distance, this result is expected.

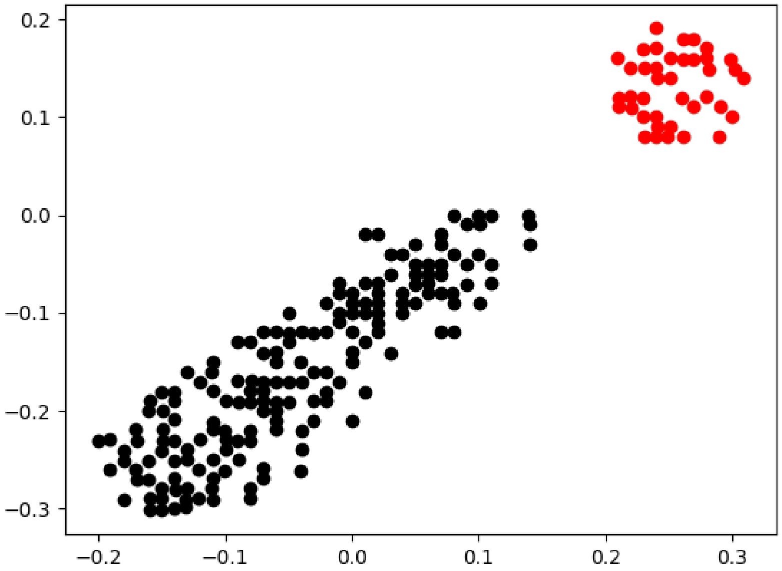


Figure 23: centroid linkage for data3

Comment4: Centroid linkage criterion is not suitable for dataset3. As seen in Figure 23, there are two logical cluster. Clusters are far from each other. Center point of the clusters also far from each other. This distrubition of the data is appropriate for centroid linkage criterion. So, centroid linkage give a good result for clustering.

3.4 data4

In this part final number of cluster is chosen as 4

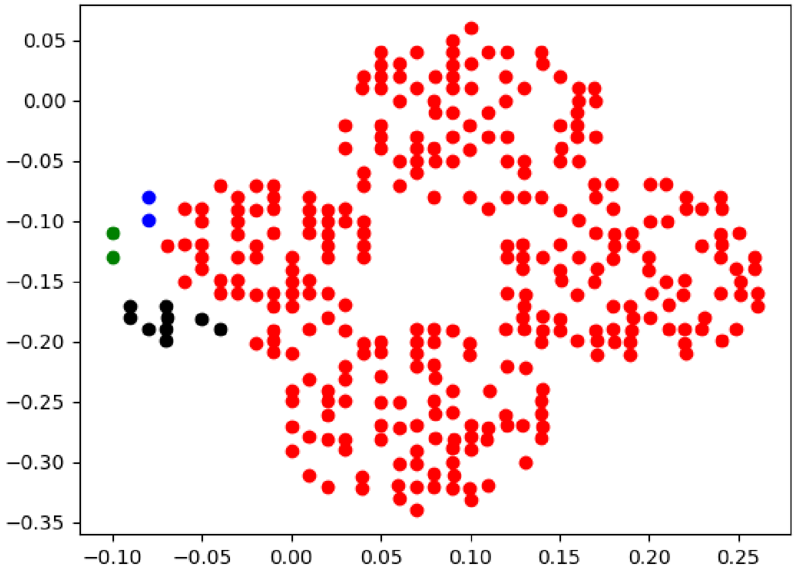


Figure 24: single linkage for data4

Comment1: Single linkage criterion is not suitable for dataset3. As seen in Figure 24, distance between data are so close. Therefore, almost all data belong to same cluster which represent as red color. Green, blue and black are a little far from red data. This enable them to belong different clusters from red cluster.

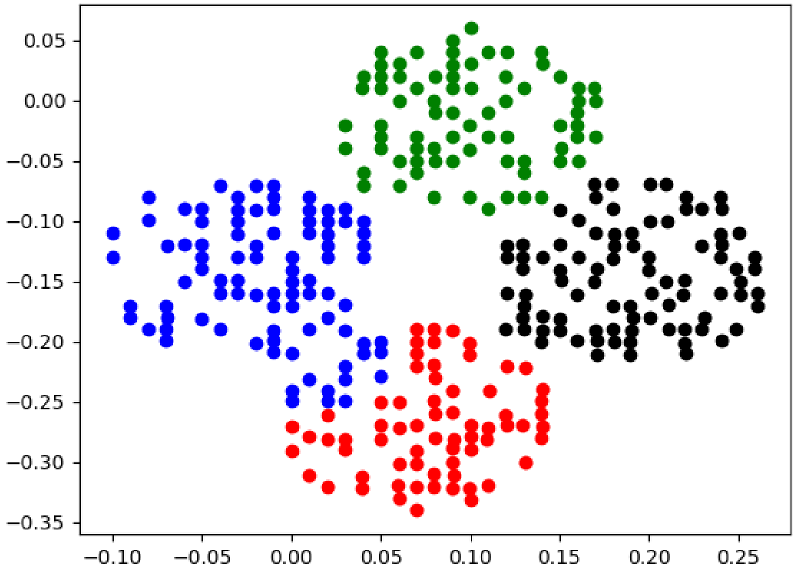


Figure 25: complete linkage for data4

Comment2: Among all of the dataset, complete linkage give the best clustering result for its own performance. The more tightly distributed structure of the data enable to give logical result. Although centroid and average linkage give more logical results (considering the red and blue clusters), complete linkage also give a good result.

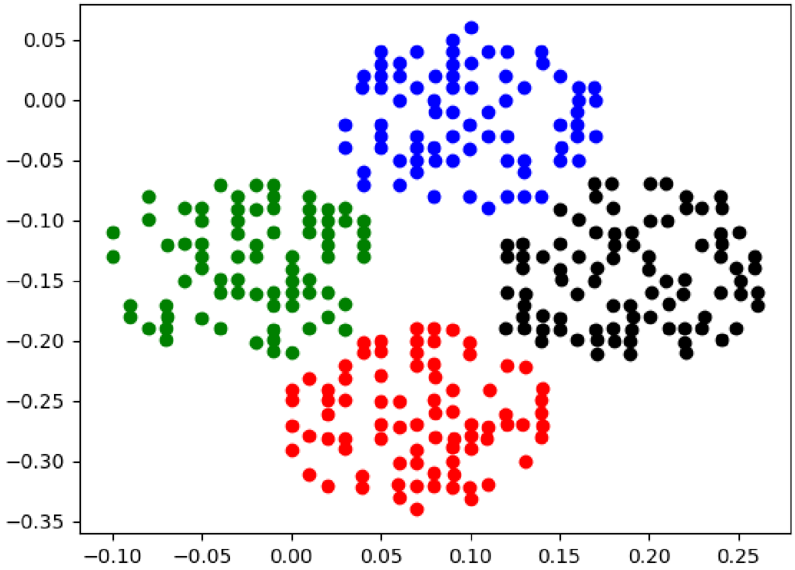


Figure 26: average linkage for data4

Comment3: Average linkage criterion is suitable for dataset4. Average linkage criterion tends to create a compact clusters. In dataset4, 4 circular clusters is seen obviously. If we think, compactness of average linkage and this distrubition of data we expect clusters as Figure26. Also, there is no cluster in cluster unlike dataset1 ;therefore, for dataset4 average linkage give a good result.

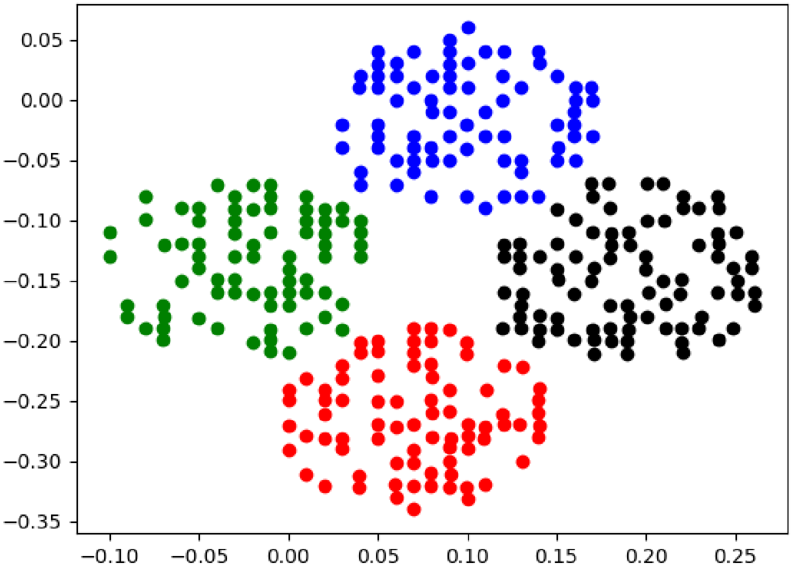


Figure 27: centroid linkage for data4

Comment4: Centroid linkage criterion is not suitable for dataset4. As seen in Figure 27, there are two logical cluster. Clusters are centers are from each other. When we look at the data, 4 center point can be seen easily. Therefore, data are gathered around these 4 centers. This enable the centroid linkage criterion is a criterion for dataset4.