CENG 499

Special Topics: Introduction to Machine Learning

HOMEWORK 2 REPORT

ARDAN YILMAZ

2172195

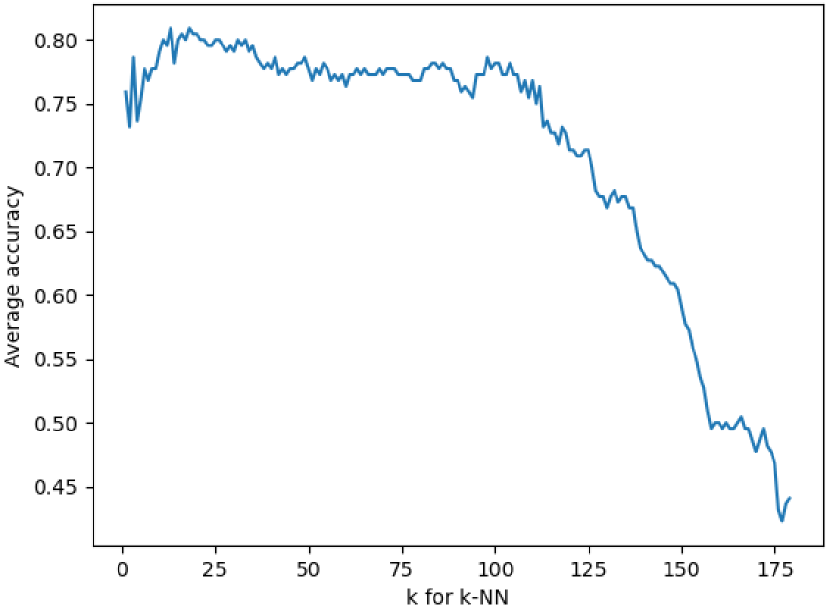
K-NN

As the number of neighbors, k, is a hyperparameter for the KNN algorithm, different numbers of k values from 1, ...179 have been tried with 10-fold cross-validation, and the corresponding accuracy values have been recorded. And the KNN algorithm is applied on the test with the k value to produce the highest accuracy, and the resulting accuracy for the test set is recorded, which can be shown below.

The change in the average accuracy as k increases is due to majority voting. Simply, clusters with a larger number of instances have more votes; that's why with larger k values, the algorithm is biased towards the cluster with more samples.

The same algorithm is applied using Manhattan and Euclidian distances, and the resulting graphs are as follows:

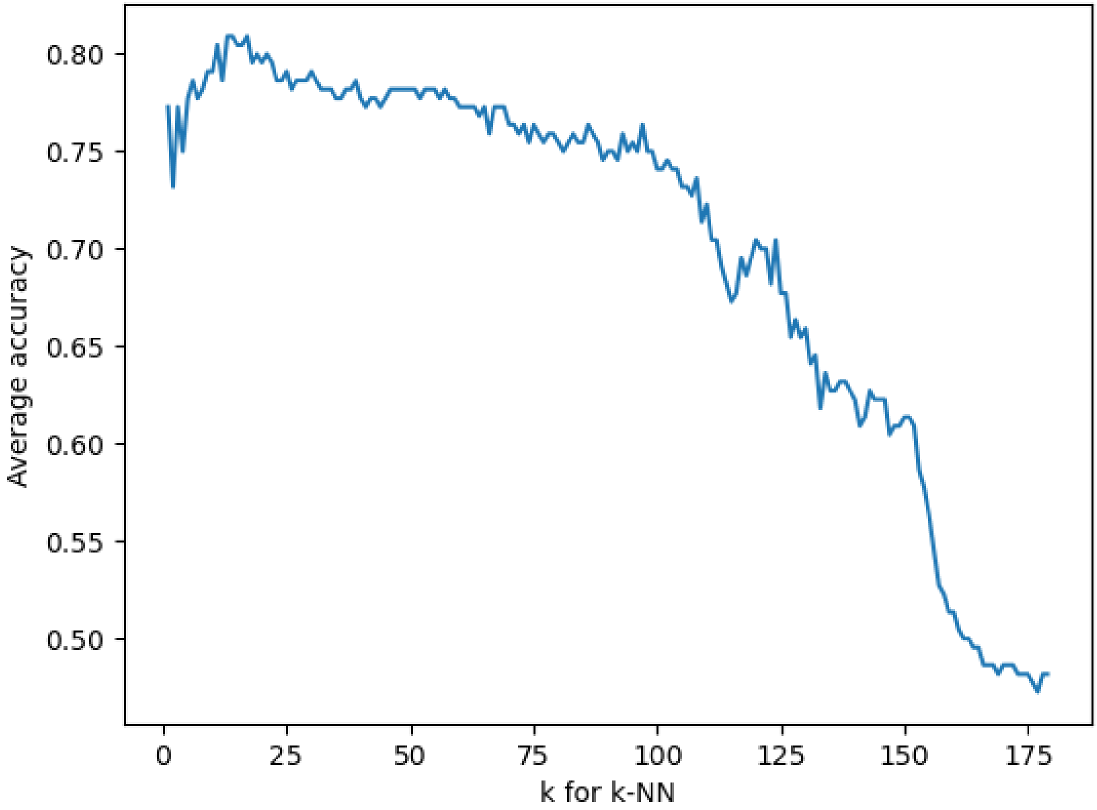
1. Using Manhattan (L1) Distance



Optimal k value = 13 (with train accuracy: 0.8090909090909092)

Test accuracy with k=13: 0.8388888888888889

2. Using Euclidian (L2) Distance



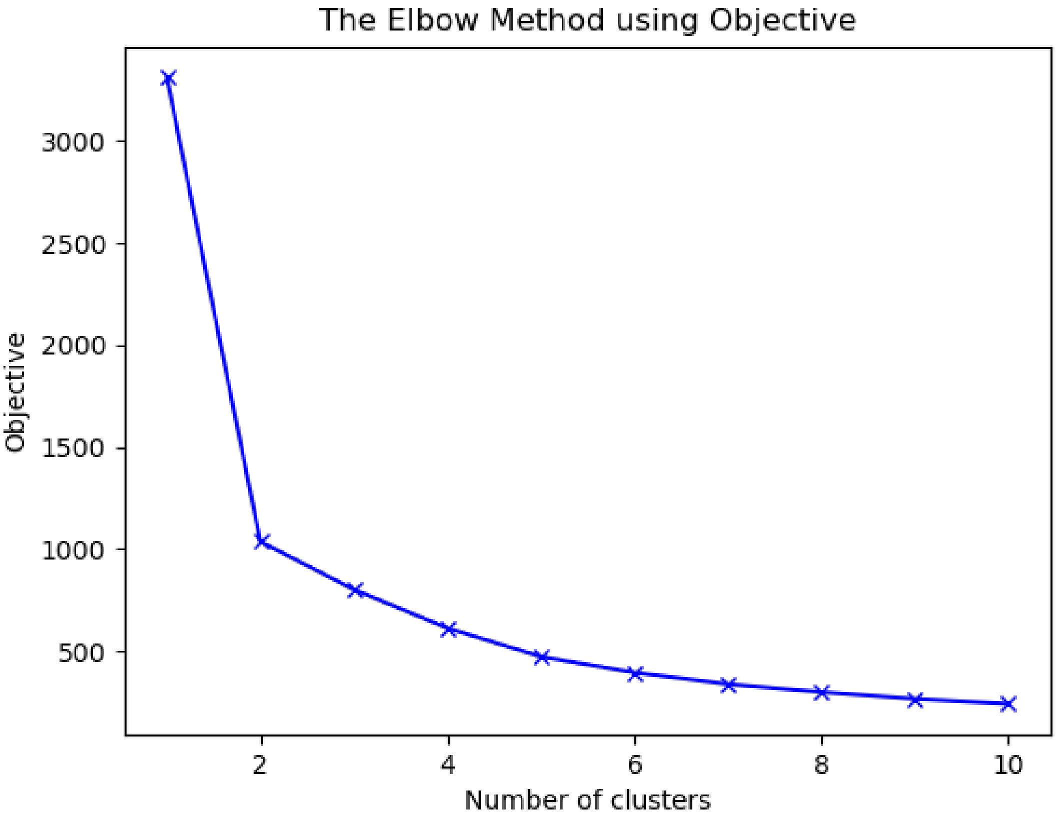
Optimal k value = 13 (with train acc 0.8090909090909092)

Test accuracy with k=13: 0.8111111111111111

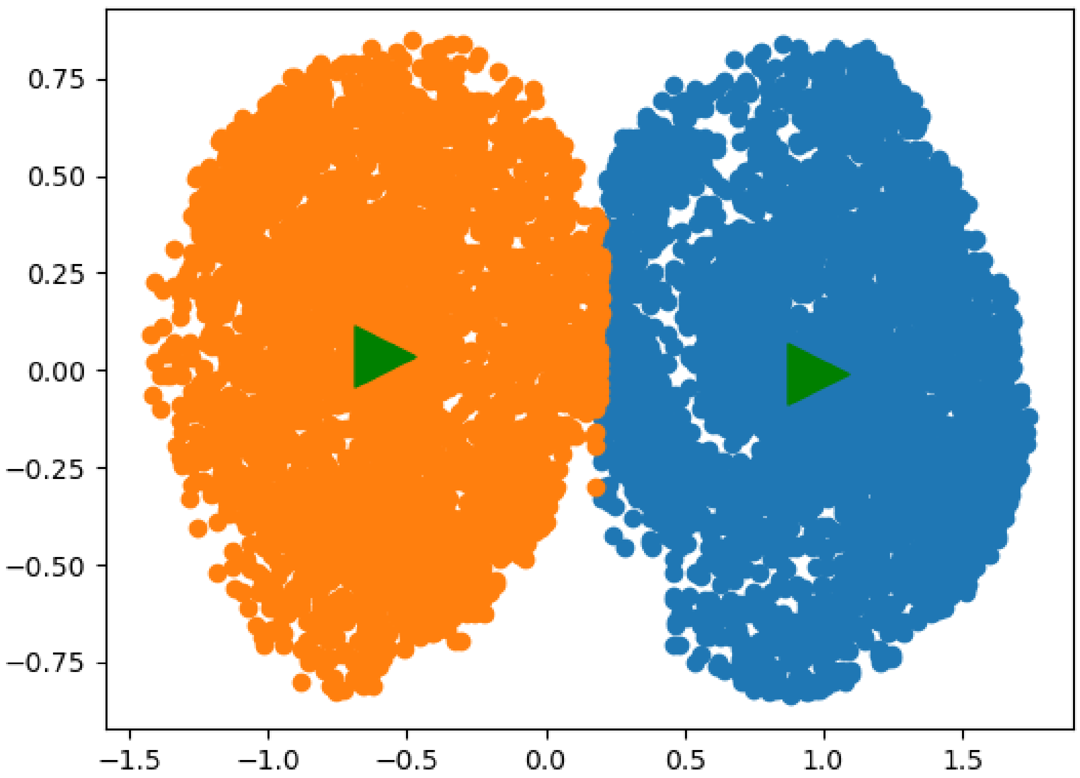
KMEANS

Different k values (1,...,10) and 10 different initial configurations for each k have been tried, and the objective value for each k with the best configuration is plotted for each dataset, where the optimal cluster number is chosen using the elbow method. Then, the resulting clusters are shown for each dataset after their corresponding objective function vs k graphs, where the green arrowheads indicate the cluster centers.

Dataset 1

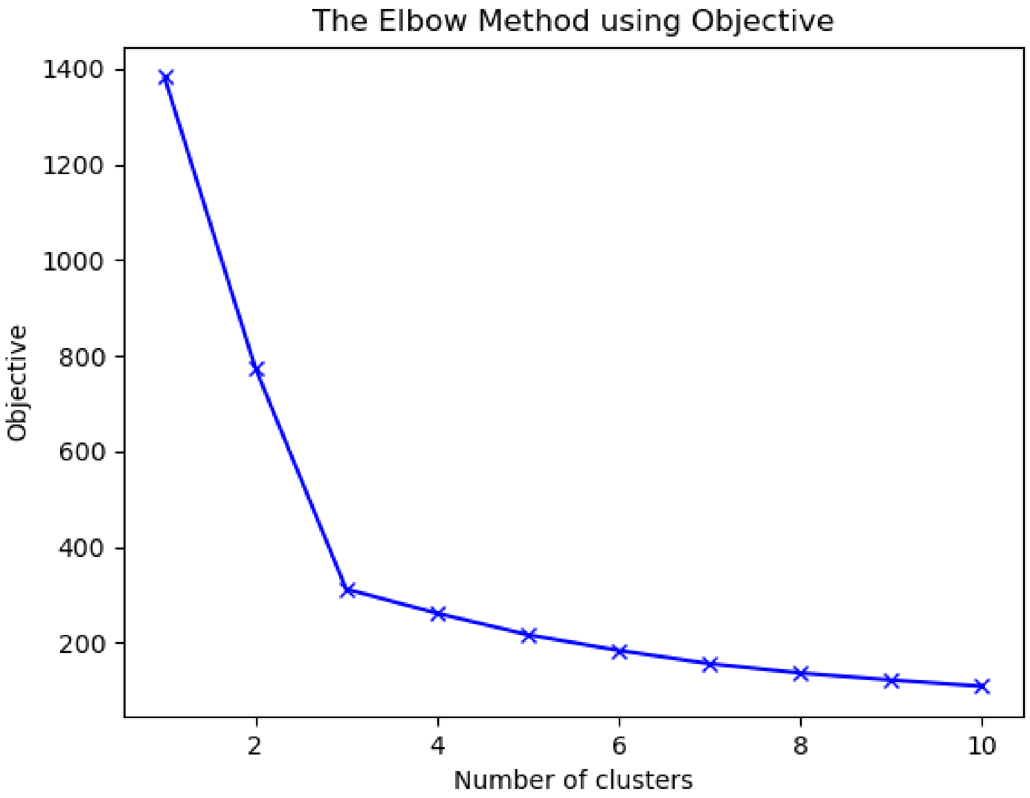


K = 2 chosen.

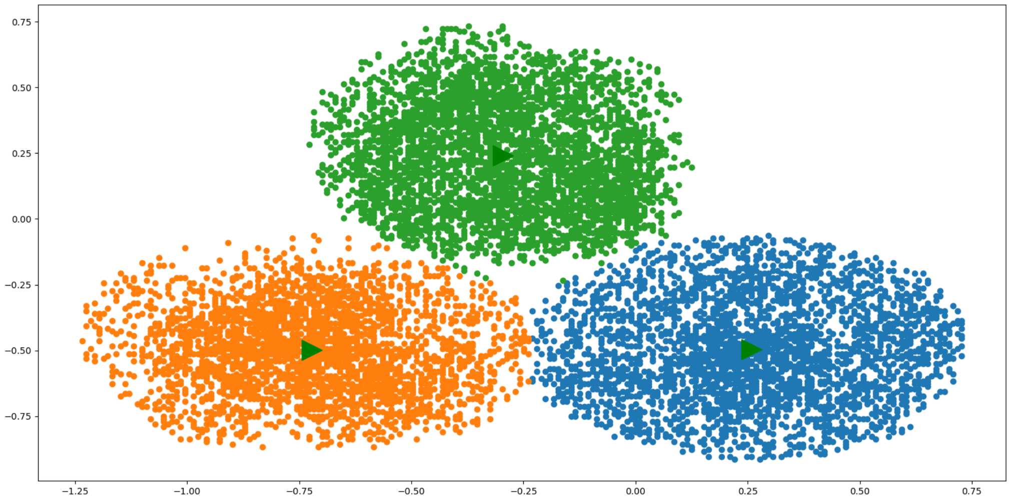

0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

,

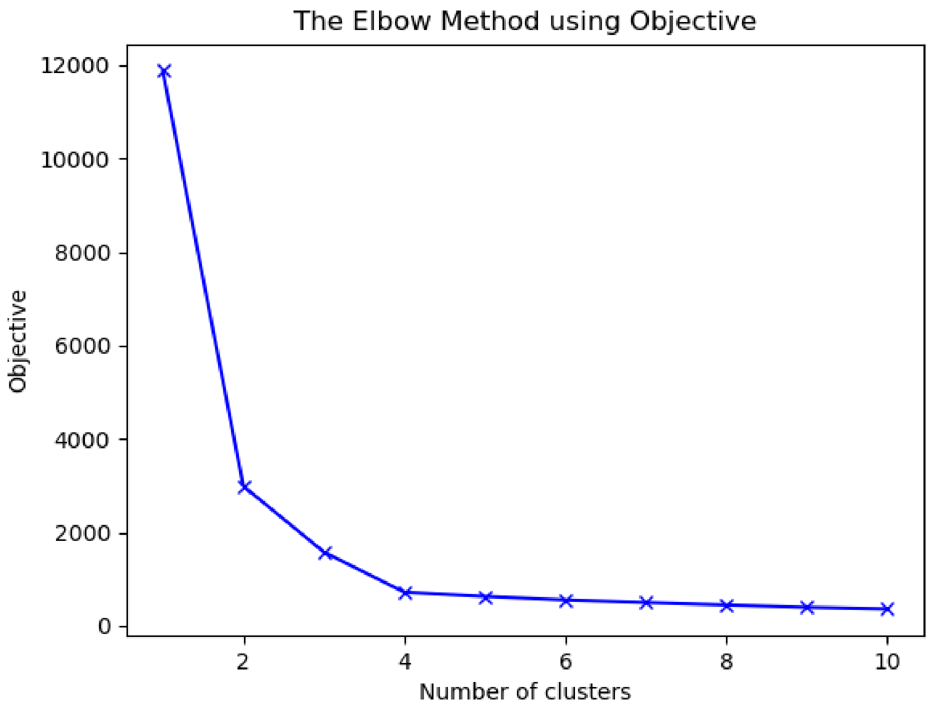
Dataset 2



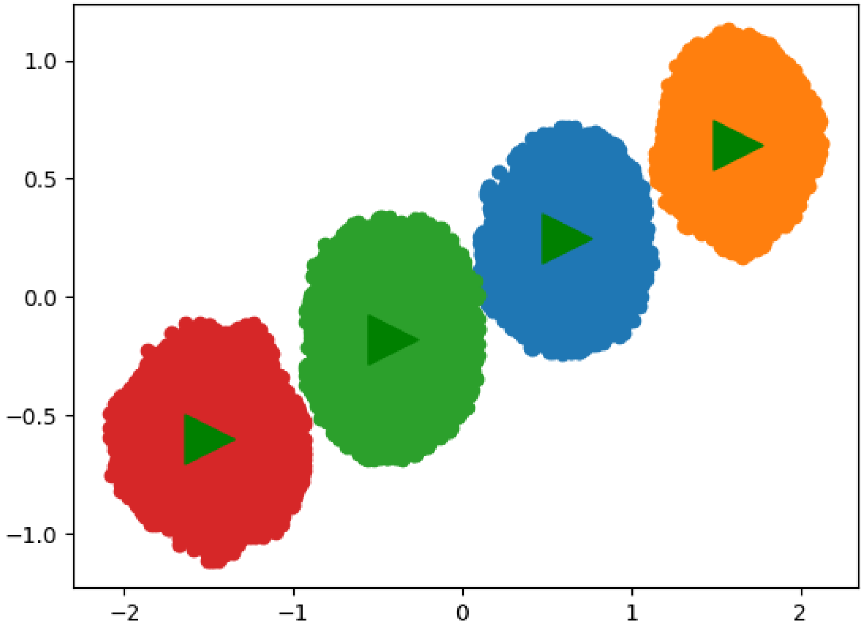
K = 3 chosen



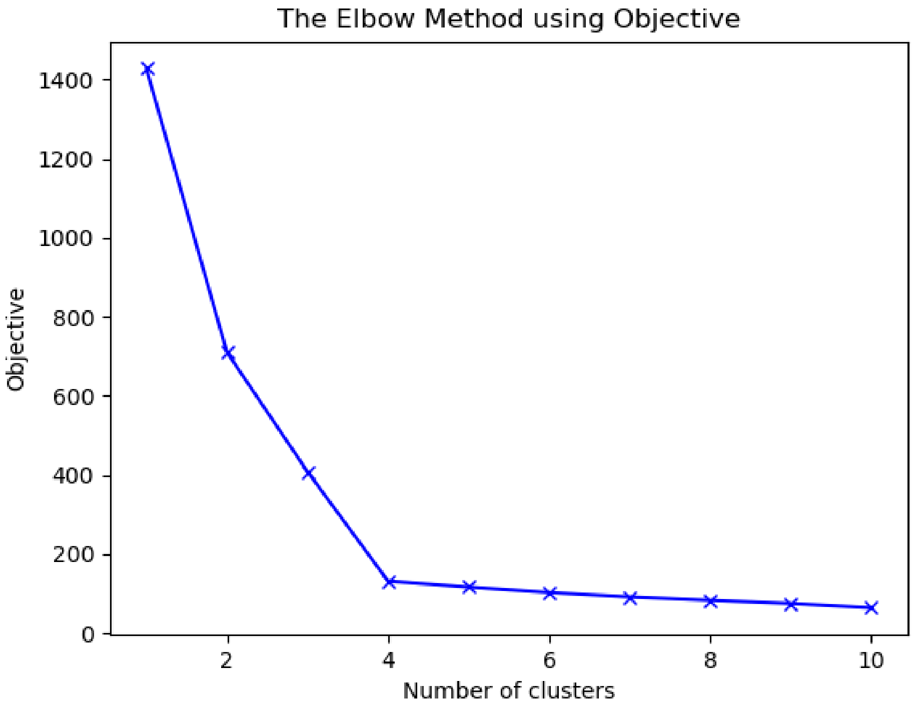
Dataset 3



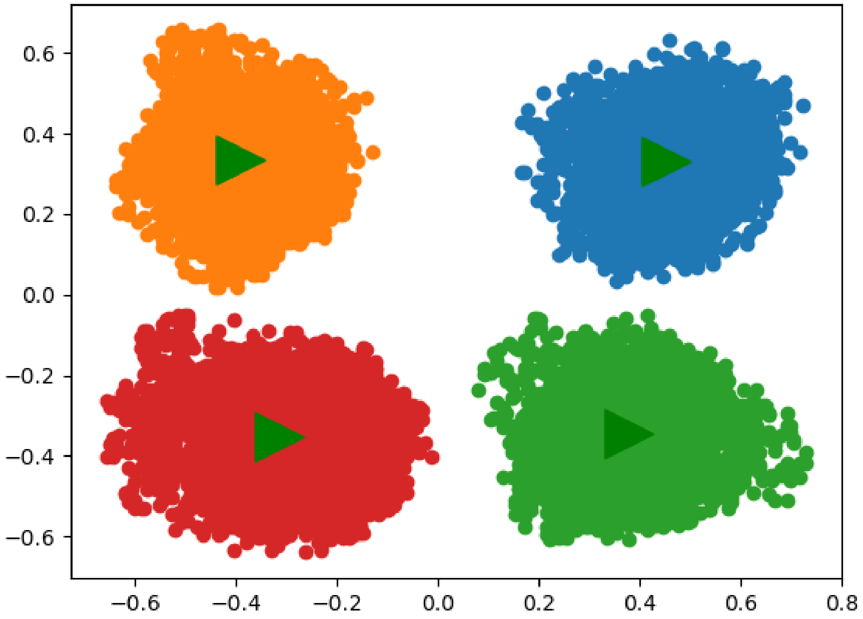
K = 4 chosen



Dataset 4



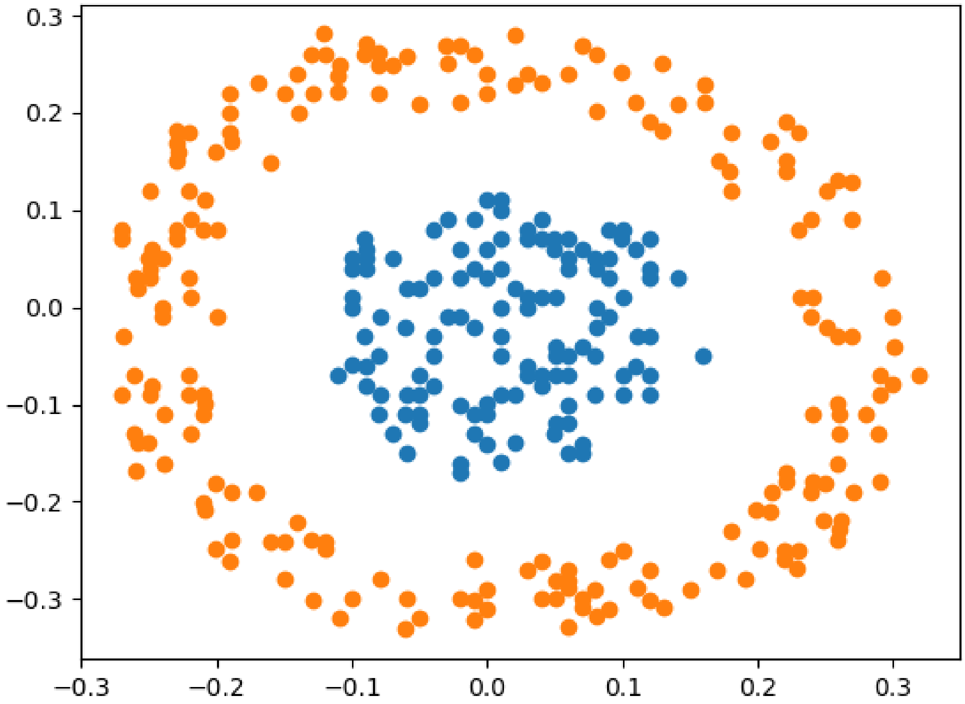
K = 4 chosen



Hierarchical Agglomerative Clustering (HAC)

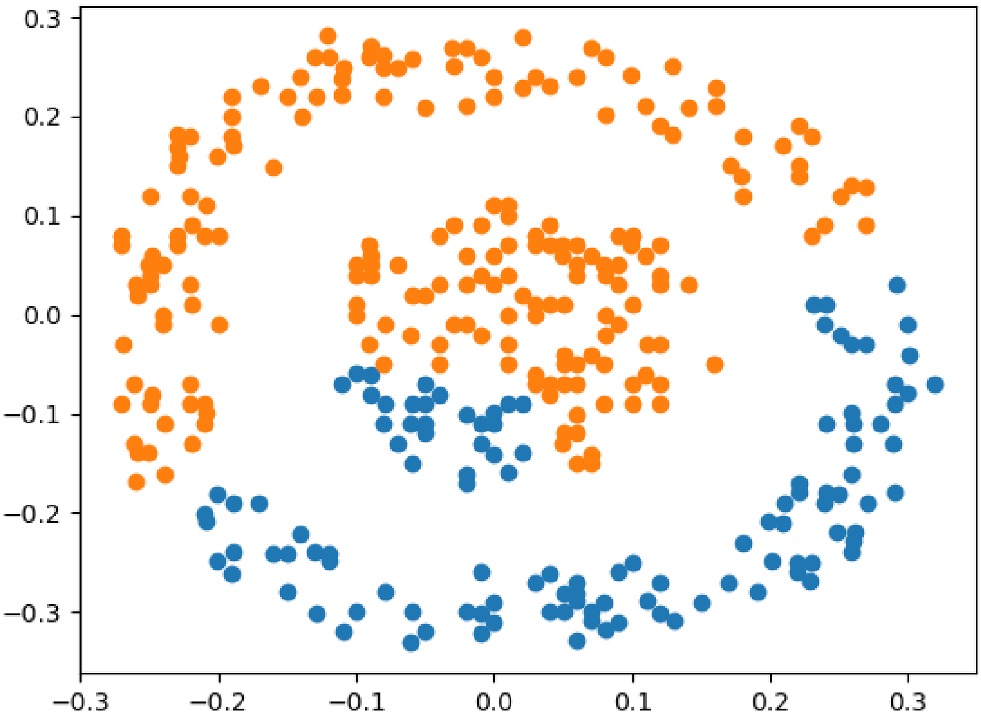
Dataset 1

Single-linkage



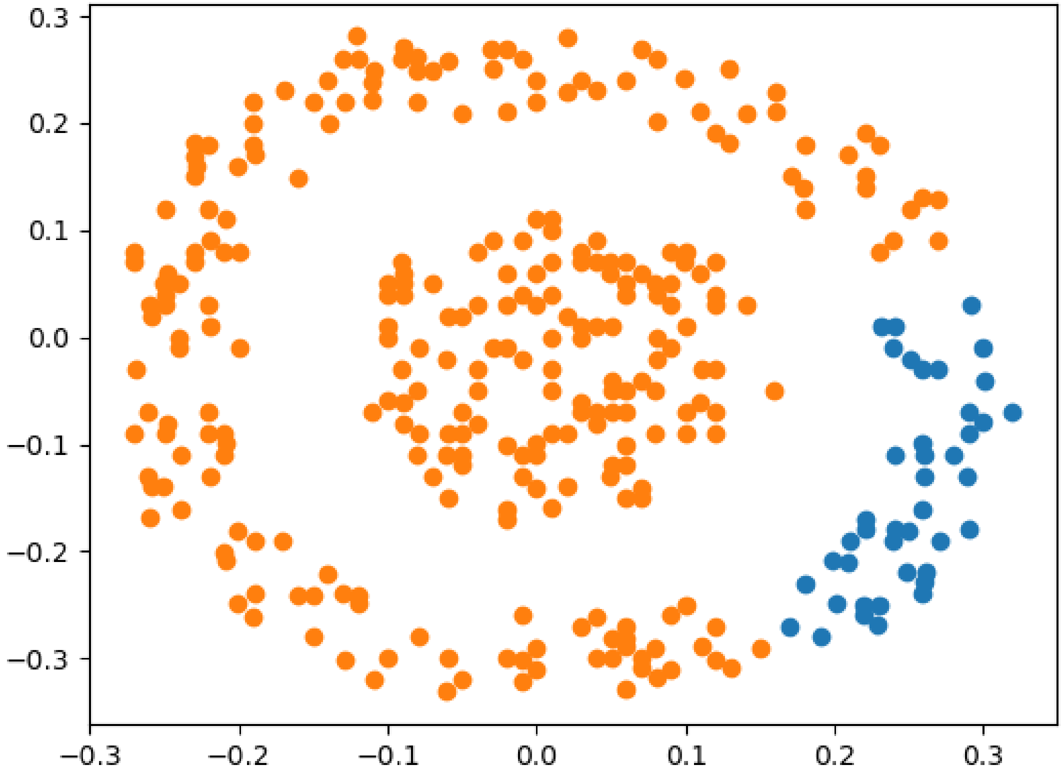
Single-linkage seems to perform well with dataset 1. It is known to be susceptible to noise and outliers, however, the dataset does not seem to contain them. Also, the minimum intra-cluster distances are shorter than the minimum inter-cluster distances for each point considering the initially formed clusters. That is, for all close neighbors of a point in the other cluster, there is a closer neighbor which is in the same cluster. That‘s why no point is assigned to a different cluster.

Complete-linkage



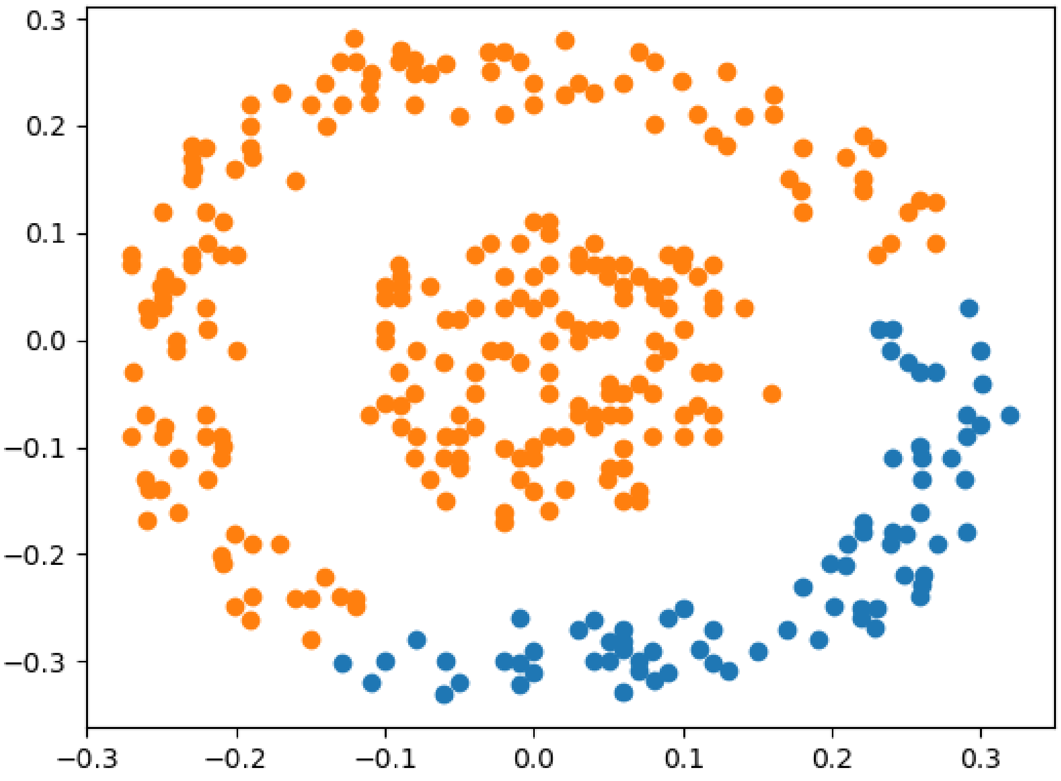
For this dataset, complete linkage did well although there are mis-classified points. Considering the initially-formed clusters, ie, most of the closest points, their neighbors which need to be in the same cluster with them, are less distant to the other clusters. That is, the real inter-cluster distance is more than the intra-cluster distances for almost every point. Consider the blue ones on the bottom left of the central cluster, the maximum distance from them to another point in the middle cluster is more than the minimum distance from them to the points in the other cluster. This cause of mis-clustering of these points is the initially formed clusters and the explained intra and inter-cluster distances.

Average-linkage



It seems to have failed to form the clusters. It is biased towards the more globular cluster whose points are closer to each other.

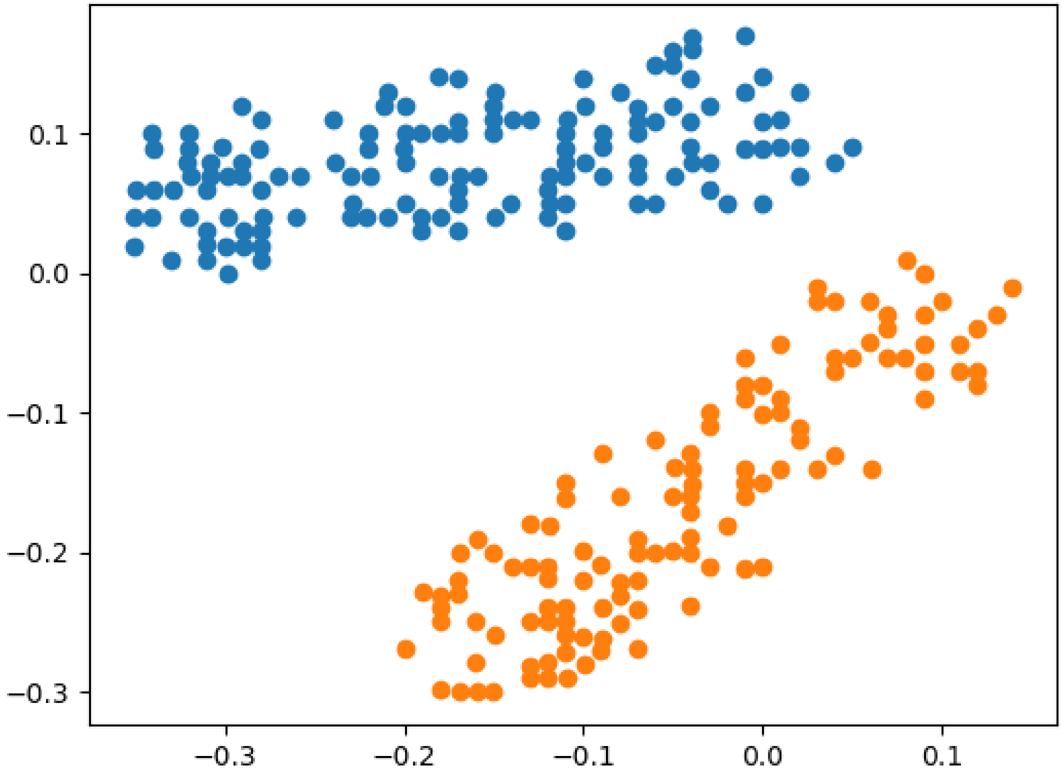
Centroid-linkage



Performed poorly due to similar reasons stated for the average linkage method.

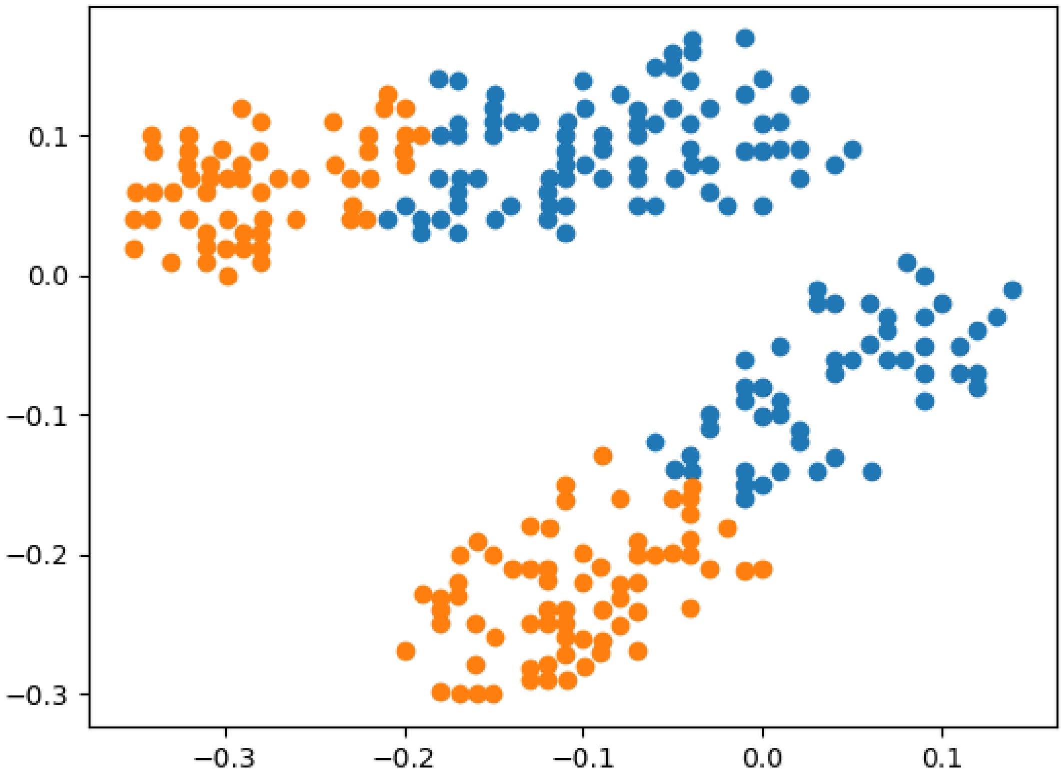
Dataset 2

Single-Linkage



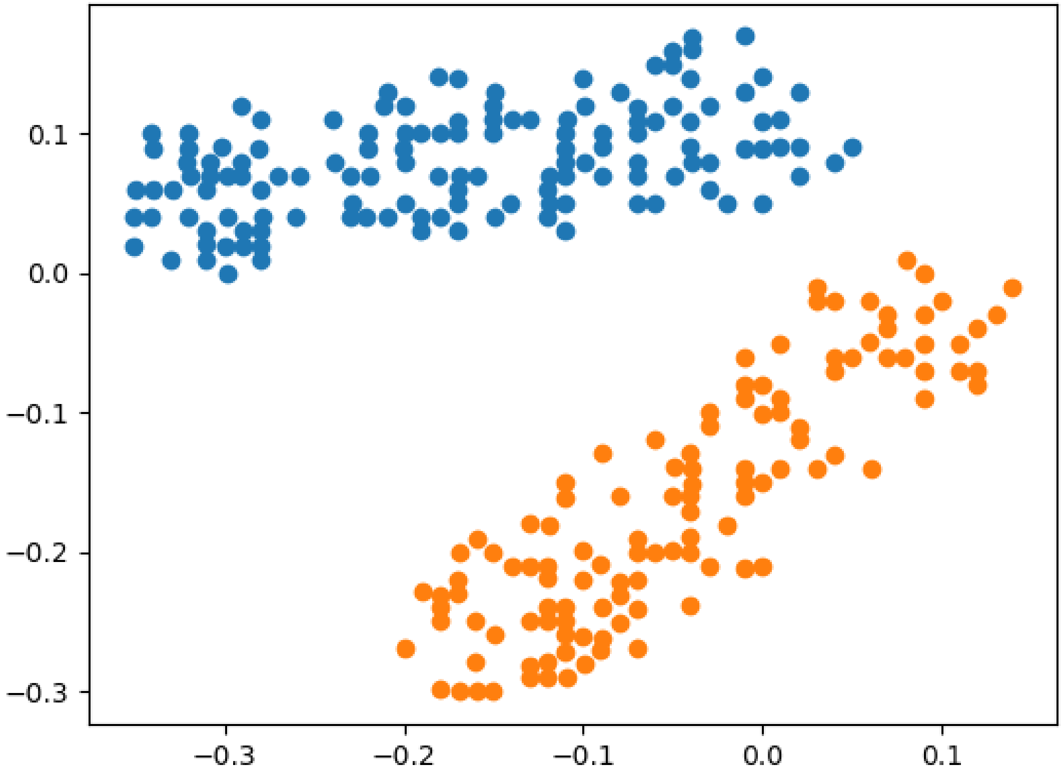
Although the formed clusters do not seem to be hierarchical, it is a feasible clustering, the reason for which is as stated in the single-linkage for dataset 1.

Complete-linkage



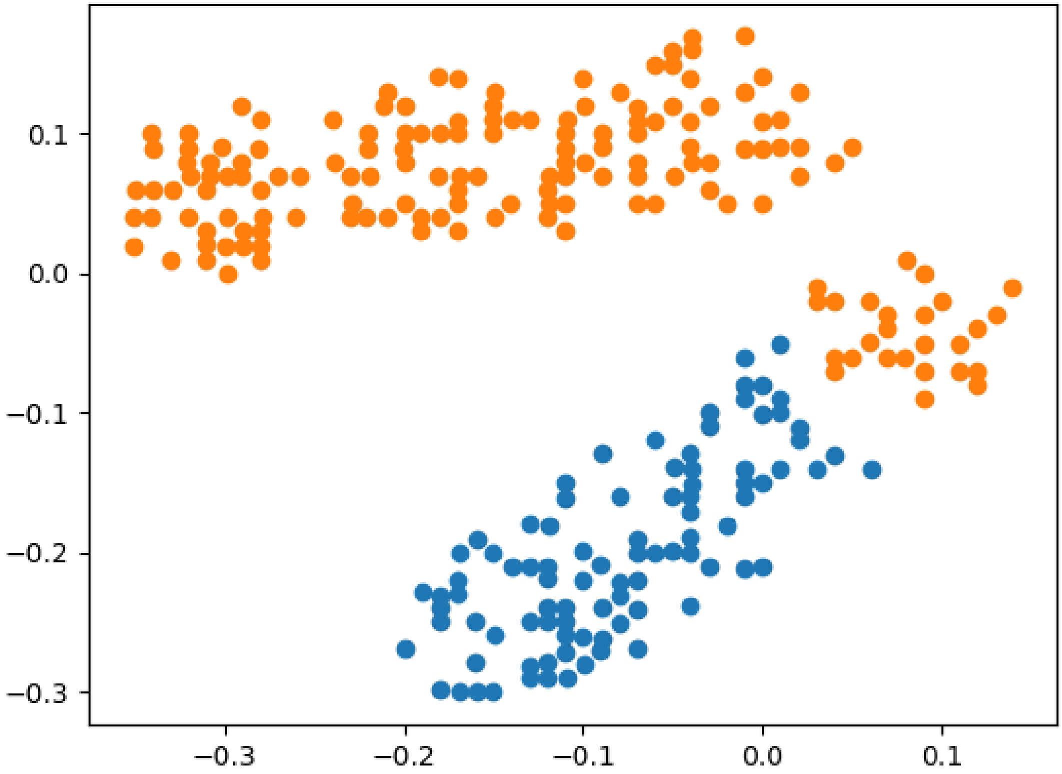
These clusters formed do not seem feasible. The complete linkage method is known to break large clusters, as the intra-cluster distance is also large for a large cluster. And here, the blue ones are clustered together because they are closer to each other than those orange dots that are in the same cluster.

Average-linkage



Clustered well. The average distance from each cluster formed which are actually in the same cluster is less than that for the ones in other clusters.

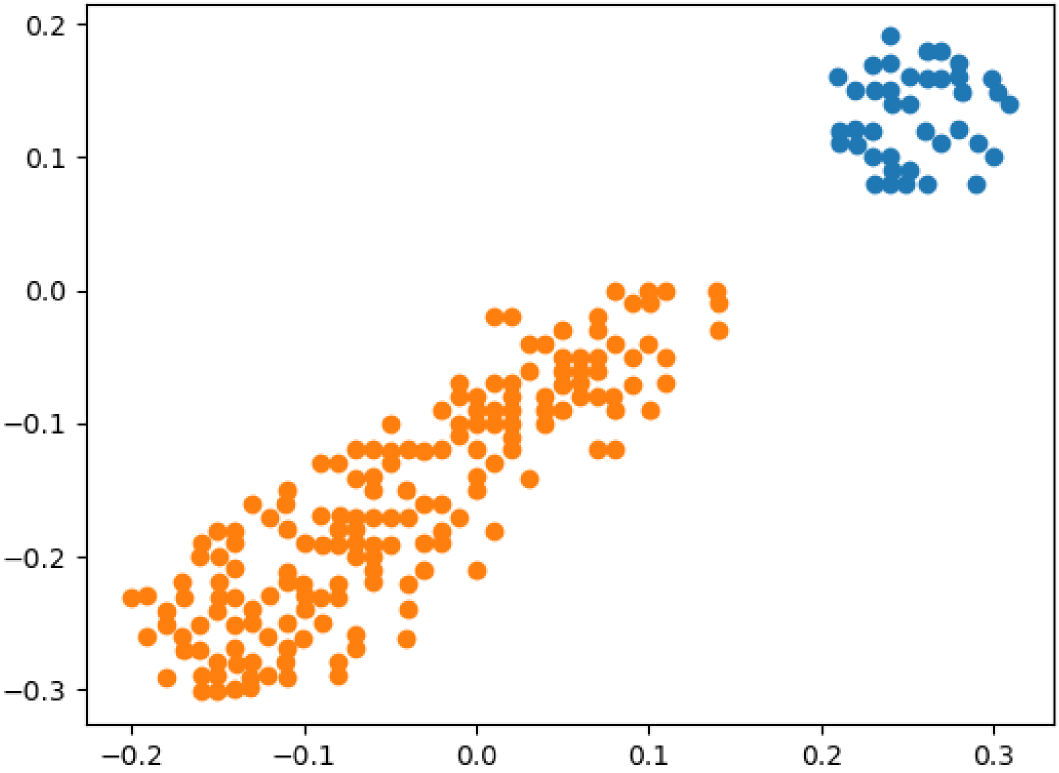
Centroid-linkage



As bigger clusters are formed, the cluster centers for the blue and orange parts in the bottom cluster got away from each other, as the points on the bottom left are closer to each other.

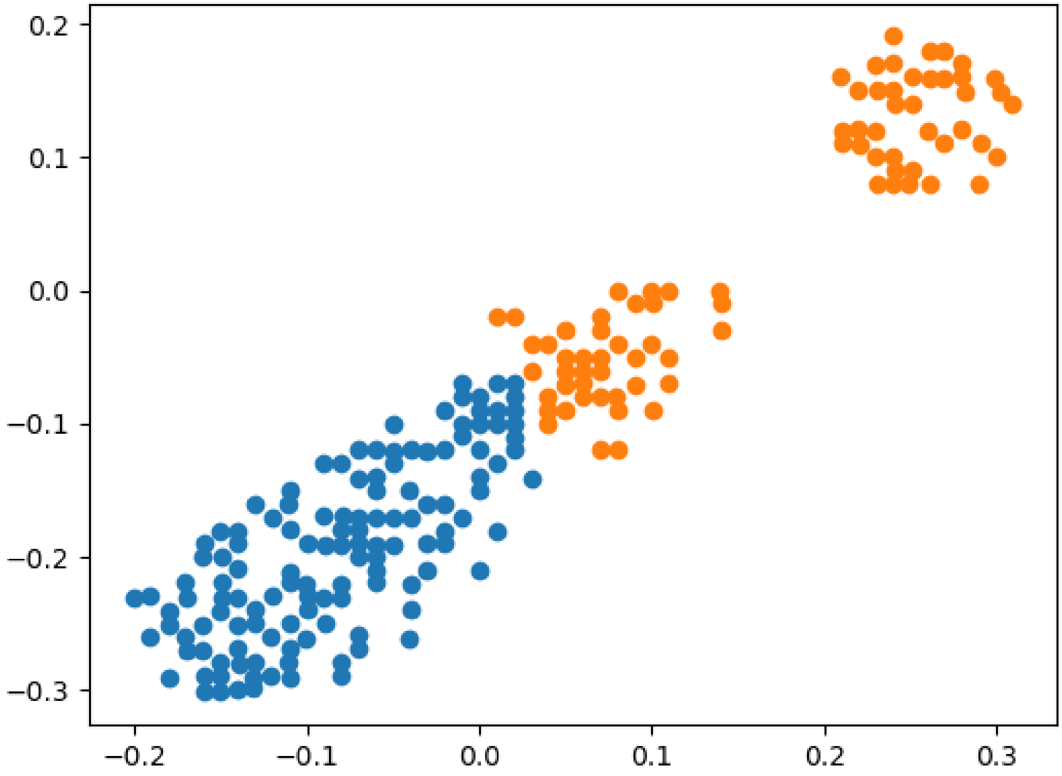
Dataset 3

Single-Linkage



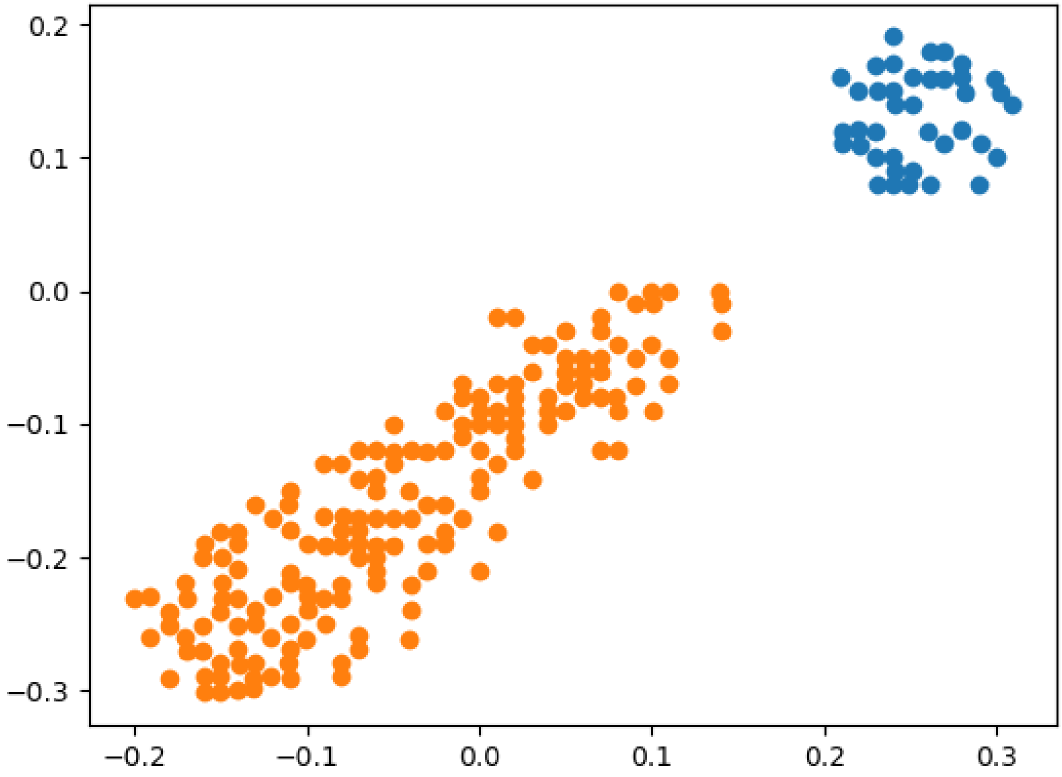
Clustered well for the same reasons as in the single-linkage method with dataset 2.

Complete-Linkage



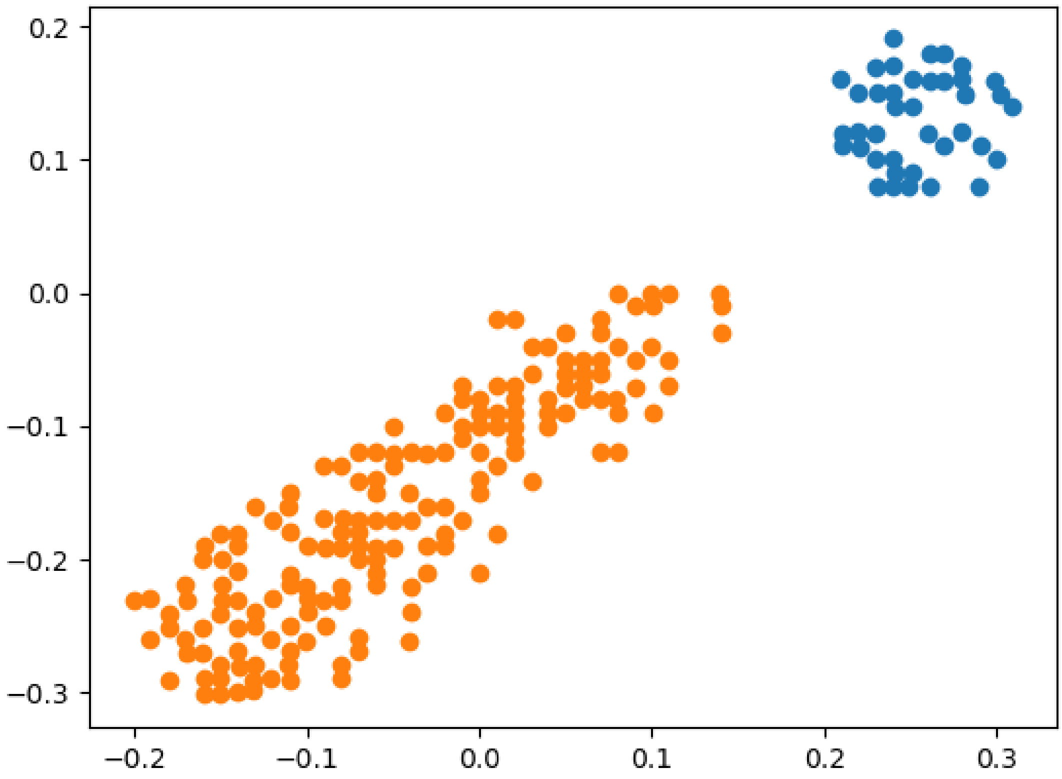
Poor clustering for the same reasons as in the complete-linkage method with dataset 2.

Average-Linkage



Clustered well for the same reasons as in the complete-linkage method with dataset 2.

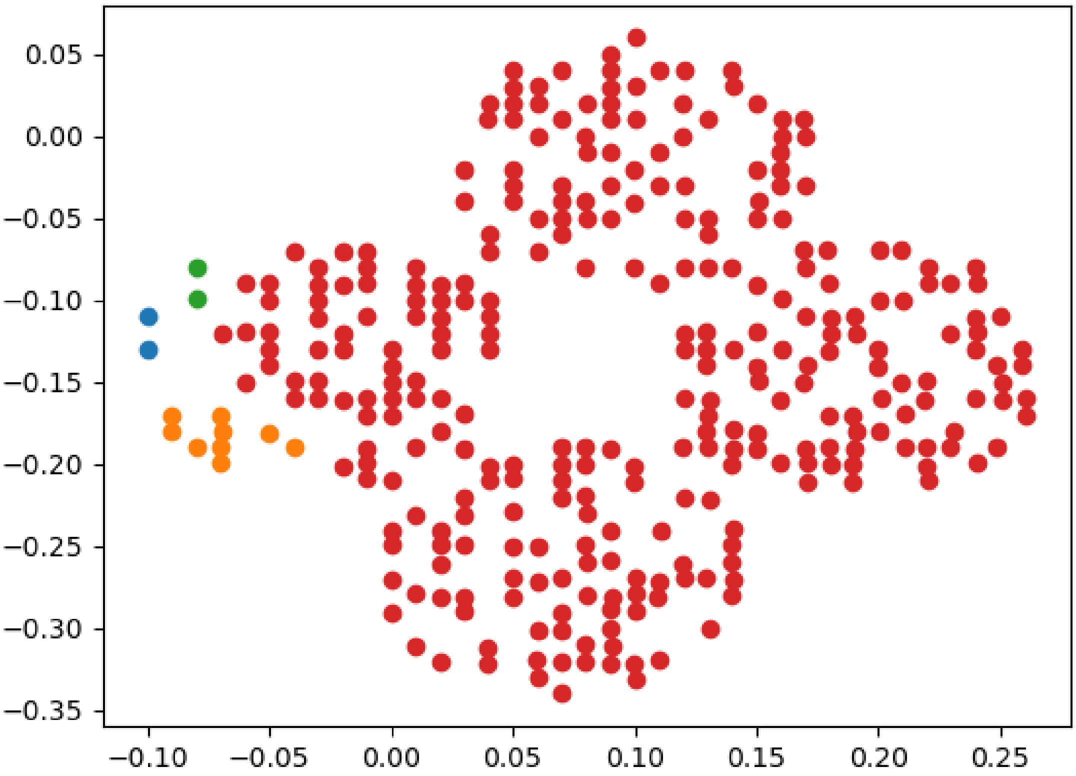
Centroid-Linkage



The formed cluster centers are coherent with the real clusters, hence the resultant clustering is feasible.

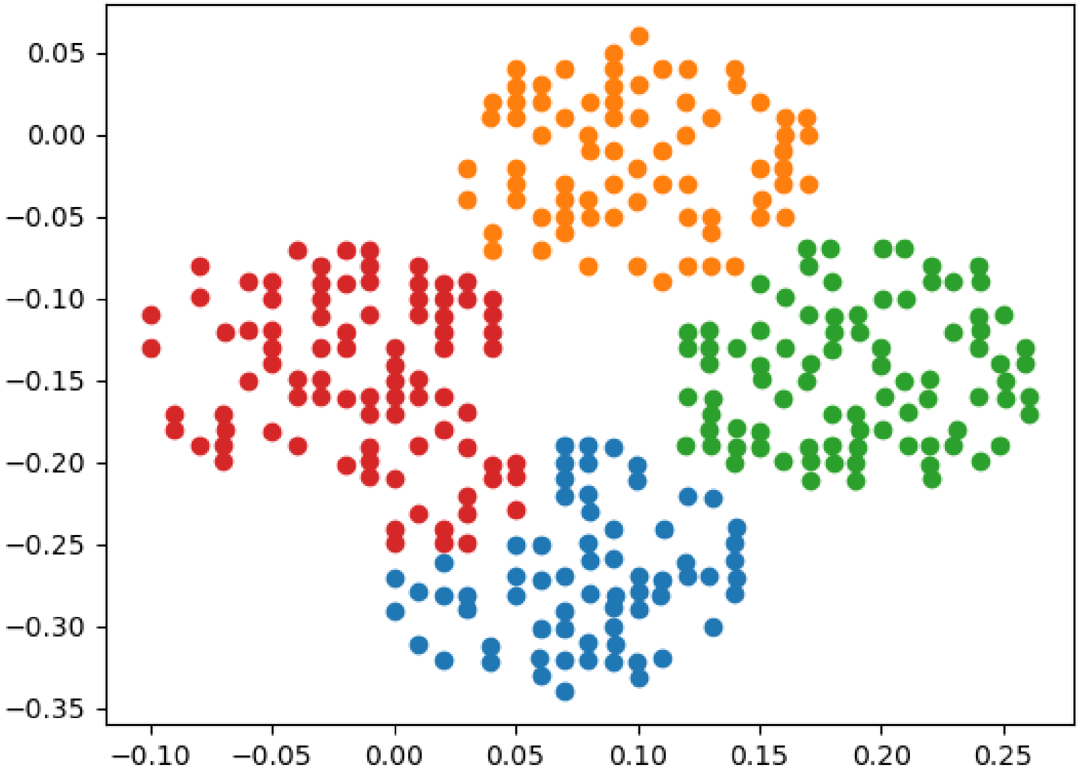
Dataset 4

Single-Linkage



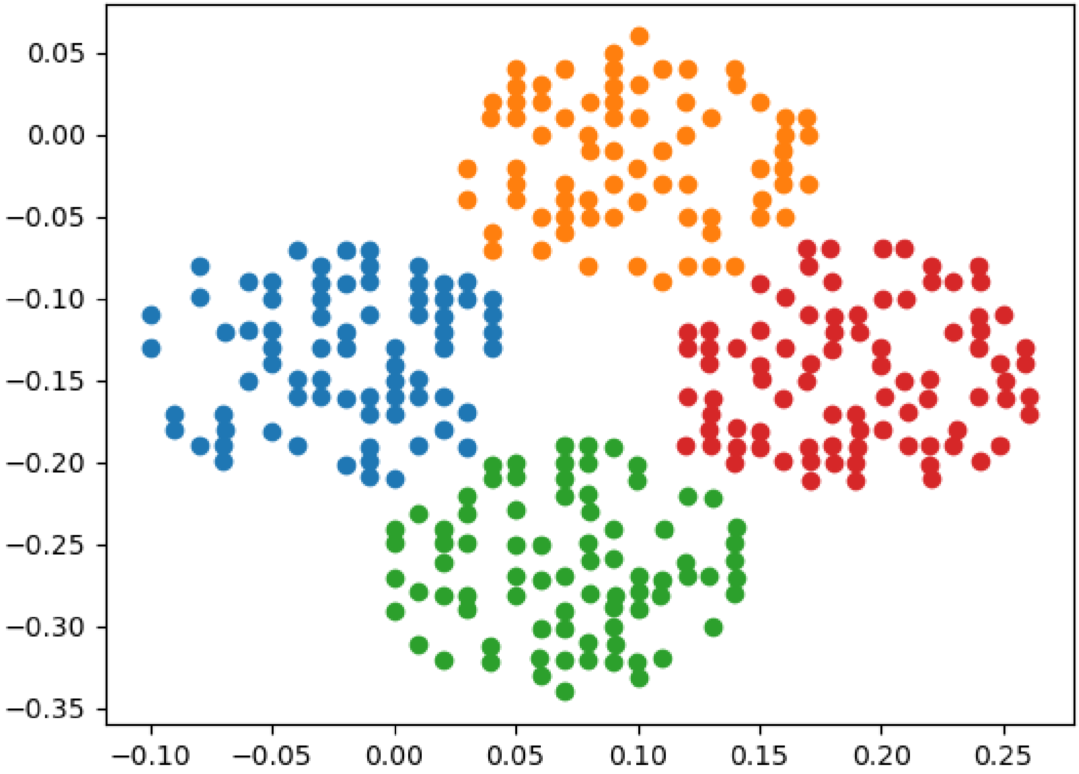
The single-linkage method performed poorly, as there are many data points on the edges of the real clusters. That is the points on the edges of the clusters are both close to the points in the same cluster as well as to their closest neighbors in different clusters. That's why the single linkage method fails to distinguish between the neighbors within the same cluster and the others.

Complete-Linkage



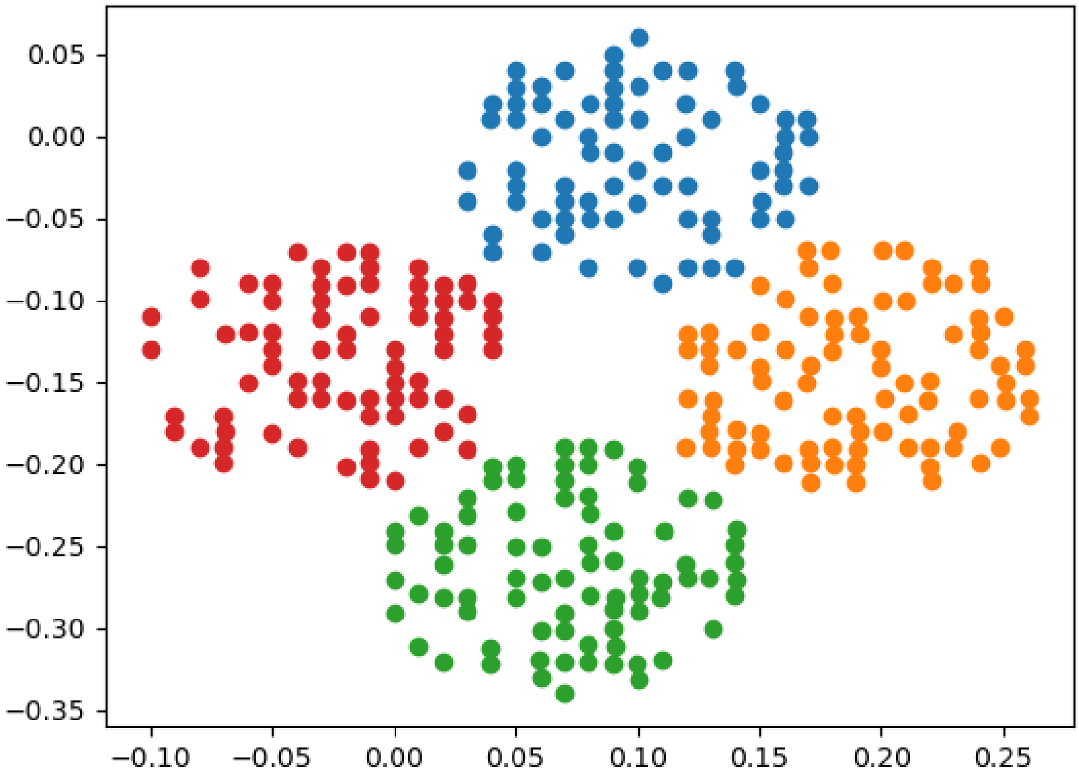
Here the complete linkage performed really well for almost the same reasons stated for the single-linkage. That is, for these clusters that are not separate enough, the maximum distance is a better measure to determine the cluster.

Average-Linkage



The average distance is a good measure for inter-cluster distances for such not obviously separate clusters, as can be seen from the resulting clustering.

Centroid-Linkage



The average distance is a good measure for inter-cluster distances for such not obviously separate clusters, as can be seen from the resulting clustering.