CENG499 HW2 Report

Alperen Oğuz Çakmak

22 December 2020

1 Part 1: K-Nearest Neighbor

1.1 K-fold Cross-validation

The best result for k is 11 according to the plot. Since all of the numerical values for the x axis did not fit, I removed some of them from the axis but all the points are in the graph from 1 to 199.

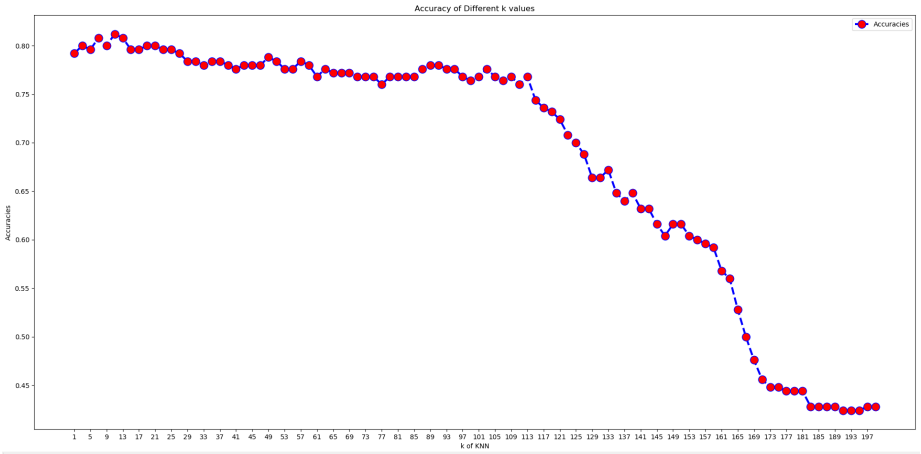

Figure 1: Accuracy for different k values

1.2 Accuracy drops with very large k values

The problem with very large k value is that we take more and more samples into consideration while making a prediction which might lead to taking all of the samples of another class. For example, in that case maybe our point is very close to the class A's neighbourhood but if class B has more samples and if we choose our K so that all of the B's data is included in our neighbours than the prediction will be wrong.

1.3 Accuracy on test set with the best k

As stated in 1.1, best result is obtained with k=11. With k value of 11 the test accuracy is %85.

2 Part 2: K-means Clustering

2.1 Elbow method

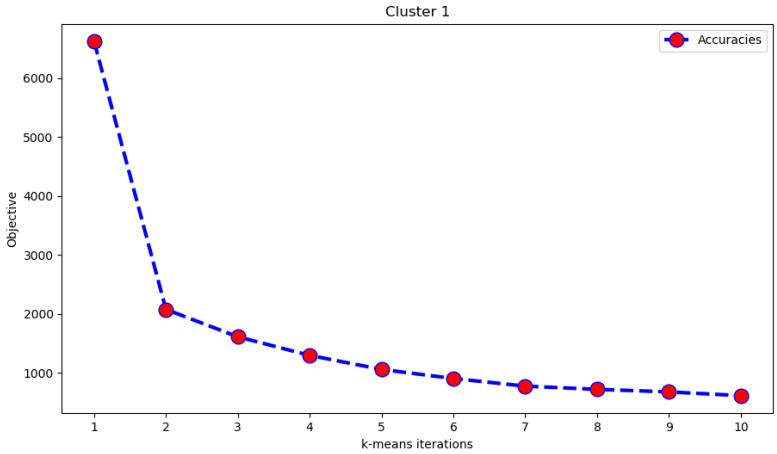

Figure 2: Graph of objective function for cluster 1. Best k is 2 for Cluster 1.

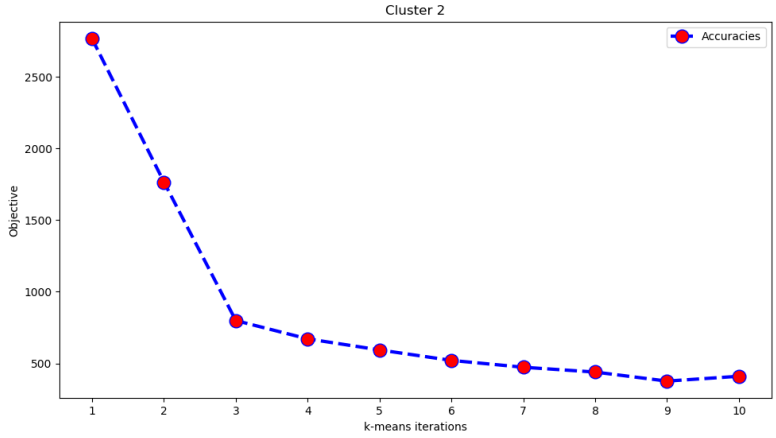

Figure 3: Graph of objective function for Cluster 2. Best k is 3 for Cluster 2.

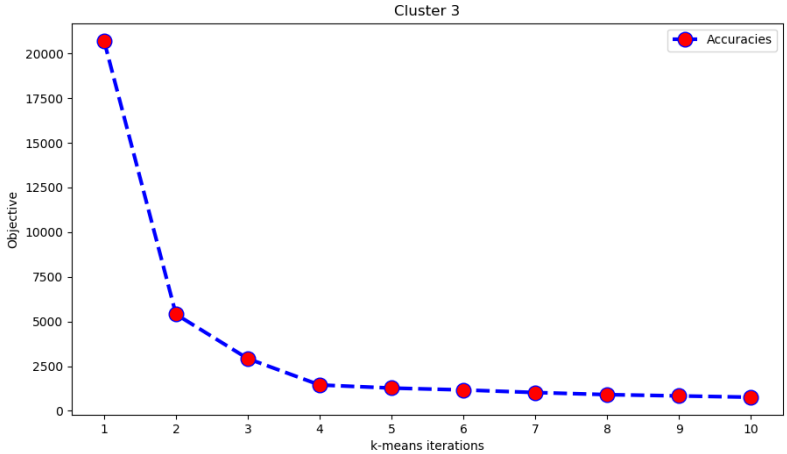

Figure 4: Graph of objective function for Cluster 3. Best k is 4 for Cluster 3.

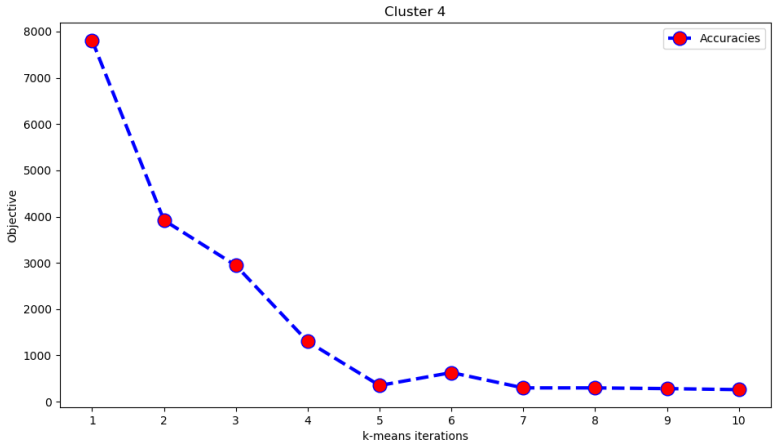

Figure 5: Graph of objective function for Cluster 4. Best k is 5 for Cluster 4.

2.2 Resultant Clusters

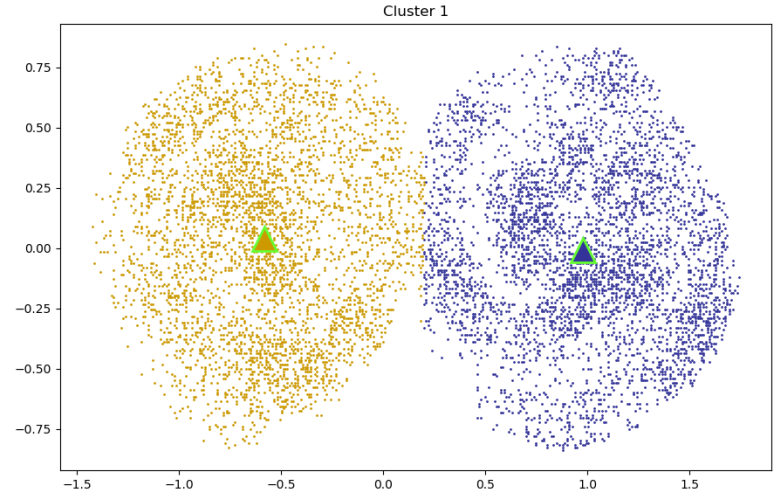

Figure 6: Cluster 1

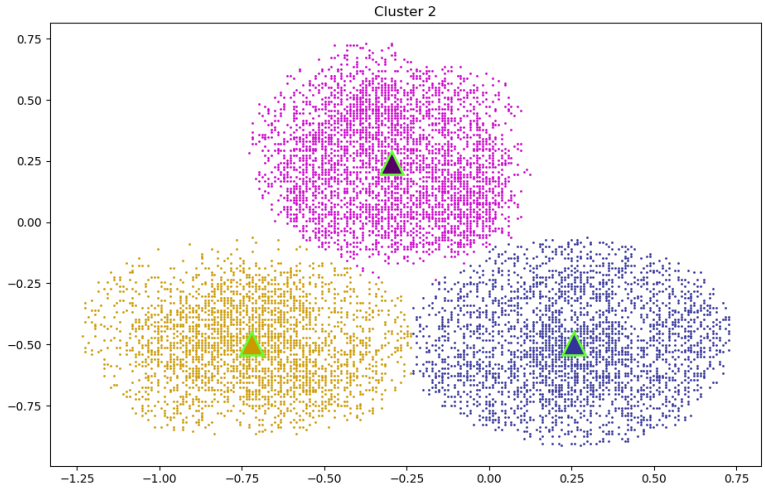

Figure 7: Cluster 2

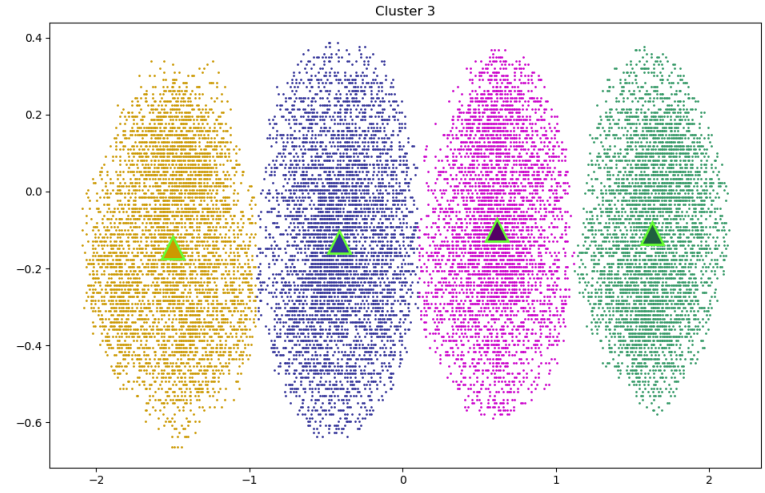

Figure 8: Cluster 3

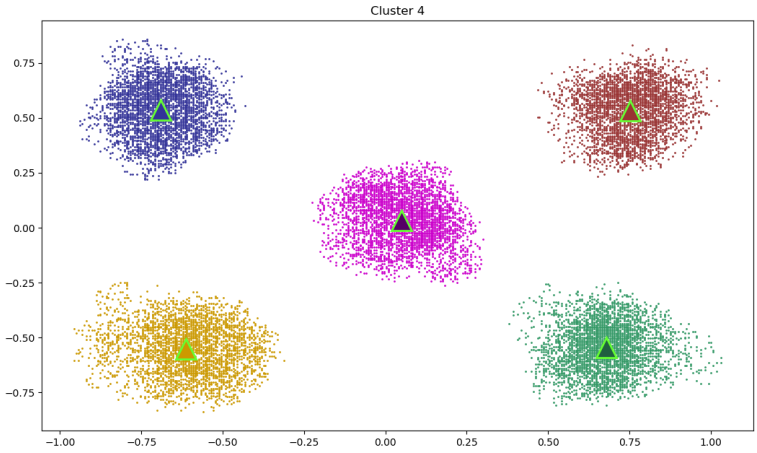

Figure 9: Cluster 4

3 Part 3: Hierarchical Agglomerative Clustering

3.1 data1

For data 1 most accurate criterion is single-linkage because in this dataset, the points form 2 groups but one group is encapsulated by the other as it can be seen from the figures. For example, when complete-linkage is applied, points lying in the outer ring are clustered with a point from the inner group of points because with complete-linkage, distance of two clusters are obtained from the points which have maximum distance in between.

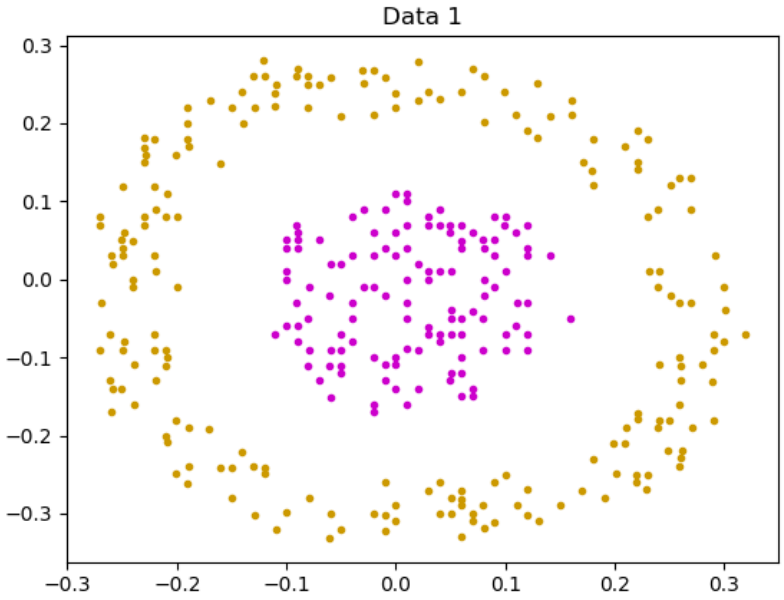

Figure 10: Data 1 with single-linkage criterion.

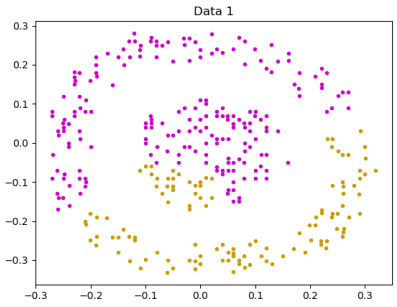

Figure 11: Data 1 with complete-linkage criterion.

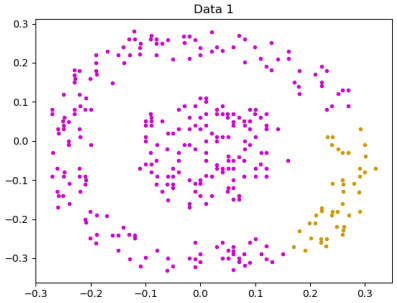

Figure 12: Data 1 with average-linkage criterion.

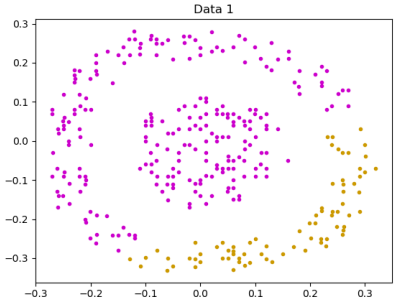

Figure 13: Data 1 with the centroid criterion.

3.2 data2

In data 2, both single-linkage and average-linkage criterion seems to be good. If a person were to look at non colorized versions of these figures, everyone would say there are 2 groups so it is not a surprise that single linkage again is very accurate. Average-linkage uses average of all of the points' distances across each cluster. Since these data points form 2 groups which are very distinct from each other it gives accurate result too. Lastly, If I were to comment on the centroid criterion, I would say it did not do a great job. The order which we iterate through the points is important with this criterion, and in this data the center points moved in a way that the cluster of upper group of points also included some of the other group's points in the brown cluster

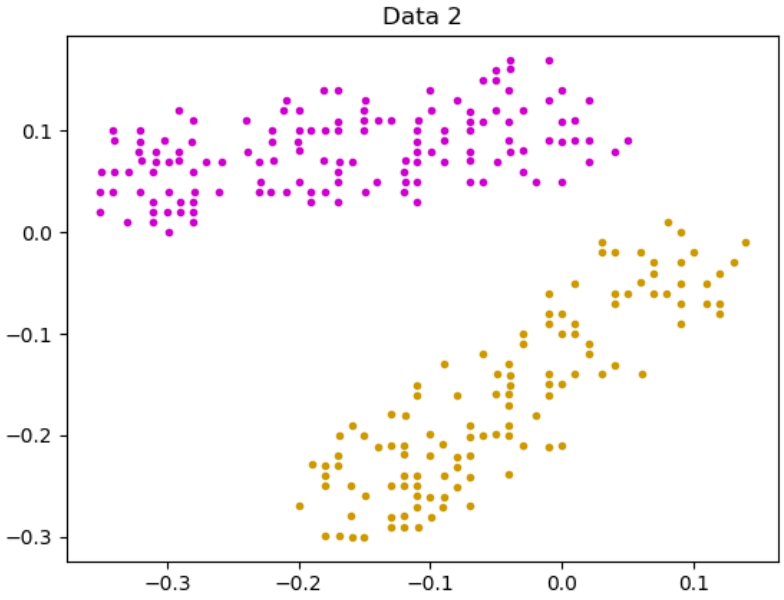

Figure 14: Data 2 with single-linkage criterion.

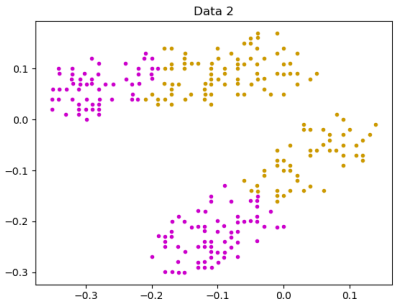

Figure 15: Data 2 with complete-linkage criterion.

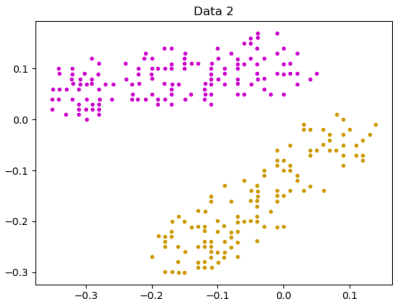

Figure 16: Data 2 with average-linkage criterion.

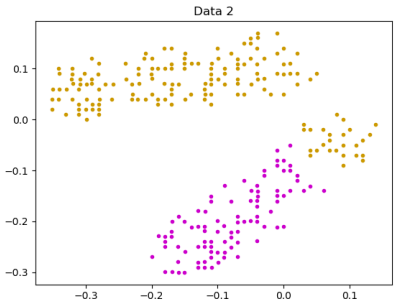

Figure 17: Data 2 with the centroid criterion.

3.3 data3

In data 3, all of the criterions except complete linkage did great job since 2 groups of data are very distinct and far away from each other. But the problem with complete-linkage is, lets think that we had 3 clusters which are purple part, brown part of the lower left group and brown group at the upper right corner. The middle cluster is closer to the brown cluster in terms of complete-linkage criterion because it uses max distance of points, because of that middle cluster is merged with the brown cluster not the purpler one which lead to the figure 19.

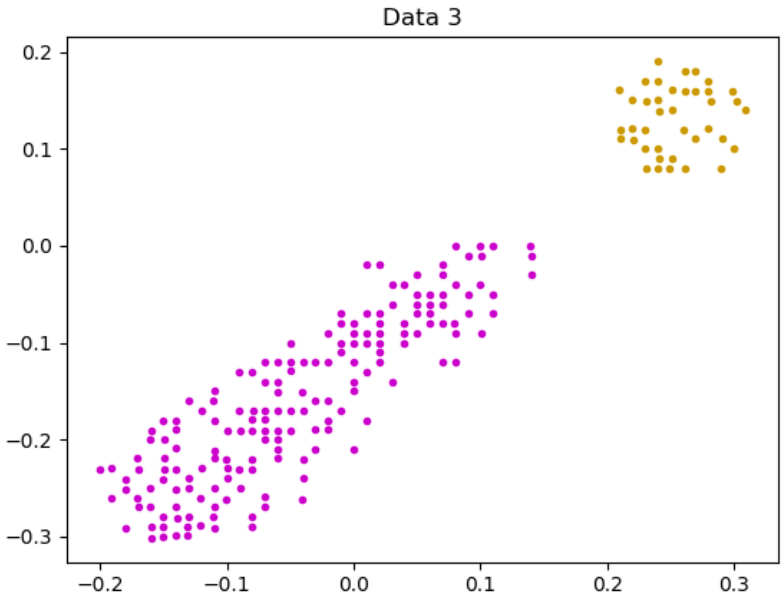

Figure 18: Data 3 with single-linkage criterion.

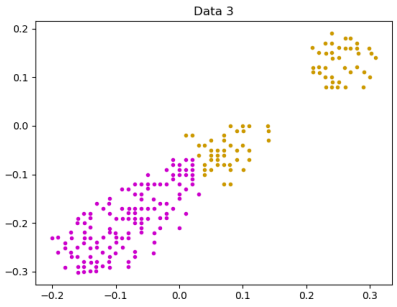

Figure 19: Data 3 with complete-linkage criterion.

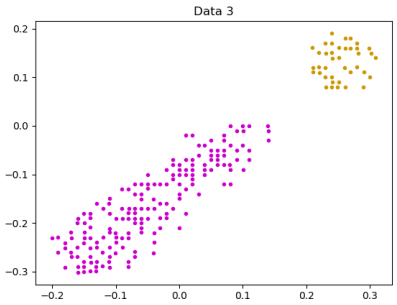

Figure 20: Data 3 with average-linkage criterion.

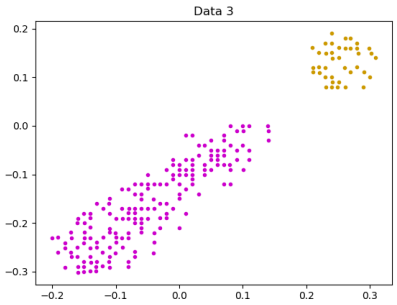

Figure 21: Data 3 with the centroid criterion.

3.4 data4

For data 4, centroid and average-linkage are accurate, complete-linkage okay and single-linkage is very bad. Since this data looks like 4 groups which are very close to each other most of the criterions are okay except the single-linkage because these groups are so close to each other. Order of traversal lead to brown cluster taking almost all of the points so that there is not much left for other 3 clusters.

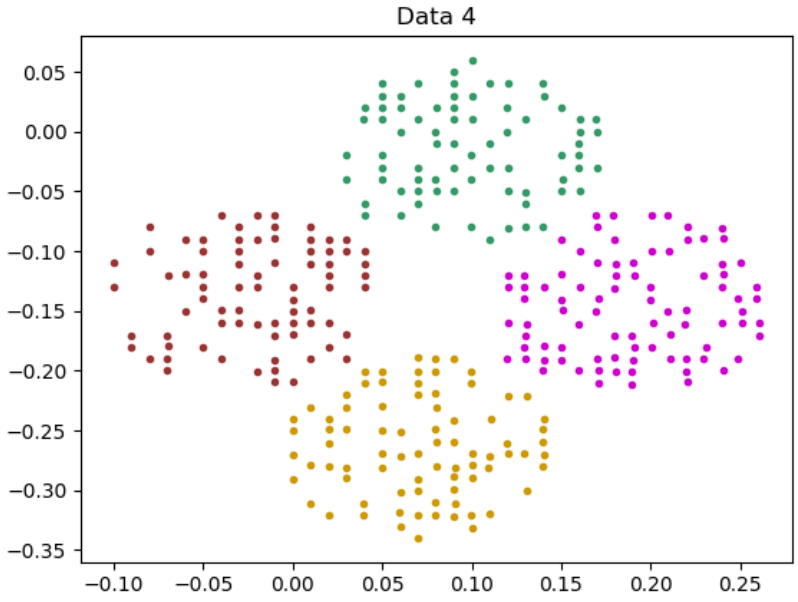

Figure 22: Data 4 with the centroid criterion.

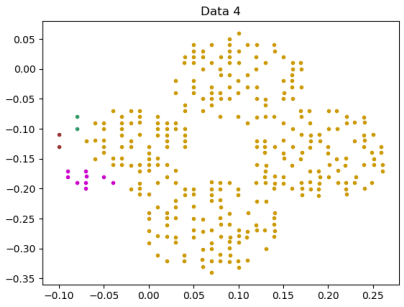

Figure 23: Data 4 with single-linkage criterion.

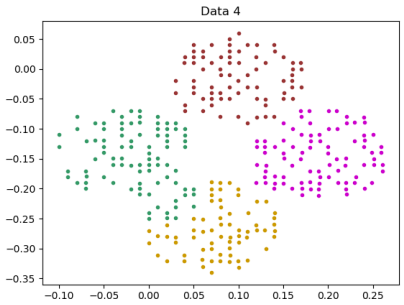

Figure 24: Data 4 with complete-linkage criterion.

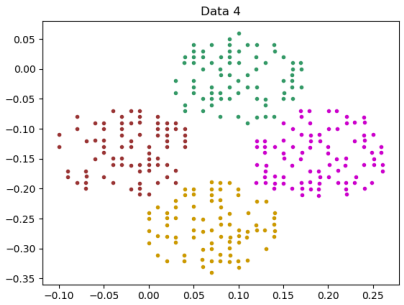

Figure 25: Data 4 with average-linkage criterion.