Report

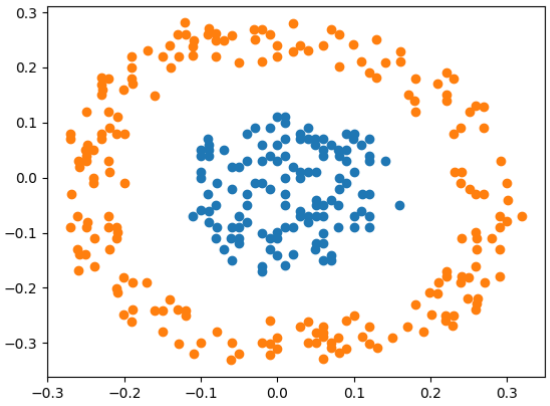
Muhammed Furkan Yağbasan

21.04.2019

1 Part1: Hierarchical Agglomerative Clustering

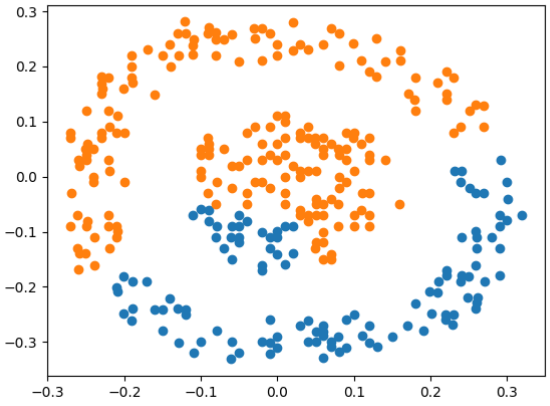
1.1 Data1

1.1.1 single linkage



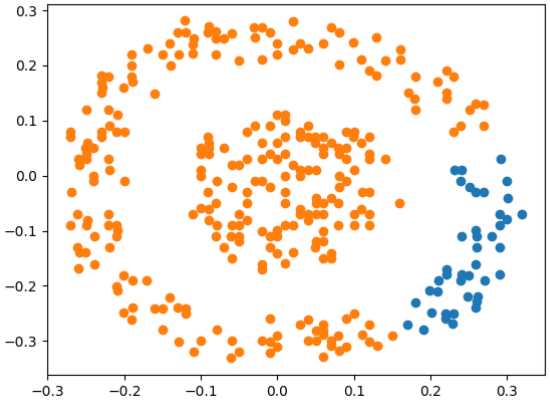
Single linkage is good at clustering curvilinear shapes and shapes that have no connection. Thus, it performed a good job in this dataset. This is the best criterion for this dataset.

1.1.2 complete linkage



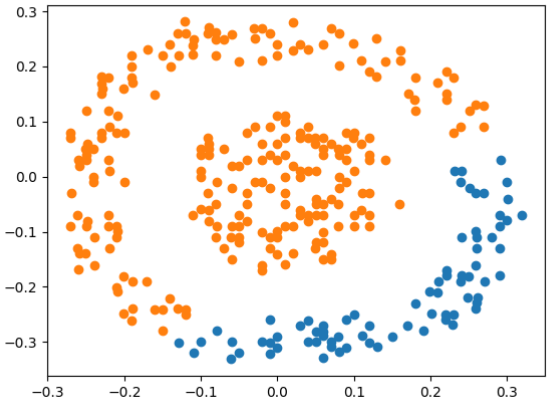
Complete linkage method is not suitable for line shapes since parts of the lines can be clustered differently when there is a near shape to them. This is because the method checks the worse distance between clusters. Therefore, it is an expected result that a part of a line to be clustered with other shapes rather than its farther parts. Apart from these, for nested shapes this method is not suitable at all.

1.1.3 average linkage



This method is not suitable for curvy lines since it needs compact and filled shapes. This method is more suitable for compact amoeba shapes, another words parts of the shape should stick together to each other in circular area.

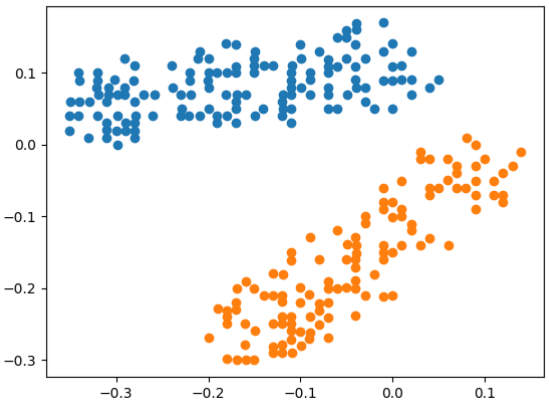
1.1.4 centroid



The shapes are circular in this dataset which is good for centroid method, but the problem is that center points of the shapes move in the same region, since the shapes are on top of each other(or inside). Therefore, centeroid method is not suitable for the set.

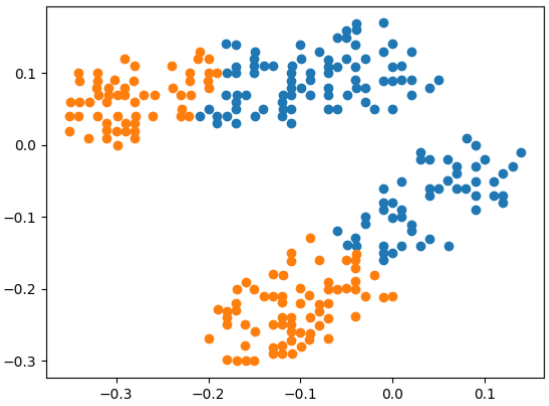
1.2 Data2

1.2.1 single linkage



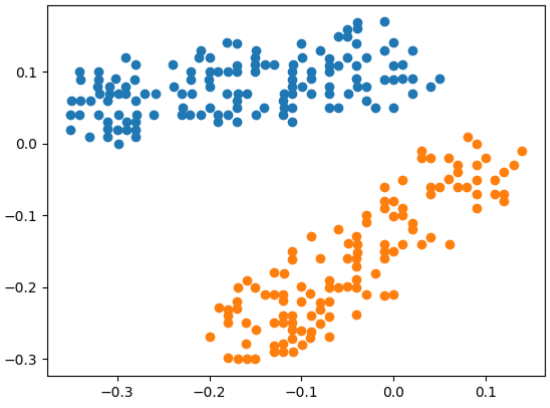
This dataset consist of 2 line shapes apart from each other. Therefore, it is a easy differentiation for single linkage method. (If lines are crossing, it would be a problem)

1.2.2 complete linkage



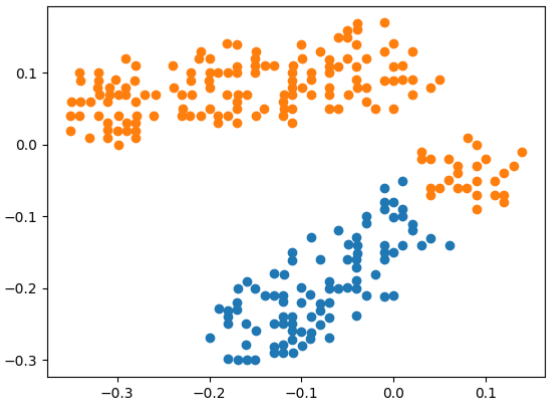
While the lines get longer, parts of the lines can jump to nearset shape to themselves, so this method would not be suitable for sets that consists long lines.

1.2.3 average linkage



Although it clustered correctly, if the lines were longer, near parts of the lines would be clustered together. This methos is not suitable for long shapes.

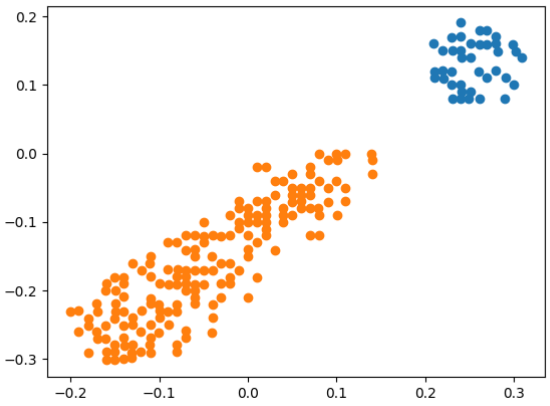
1.2.4 centroid



Centroid method is not suitable for line shapes since the algorithm searches the area around the center point with a circular manner. Therefore, it performed poorly in this dataset.

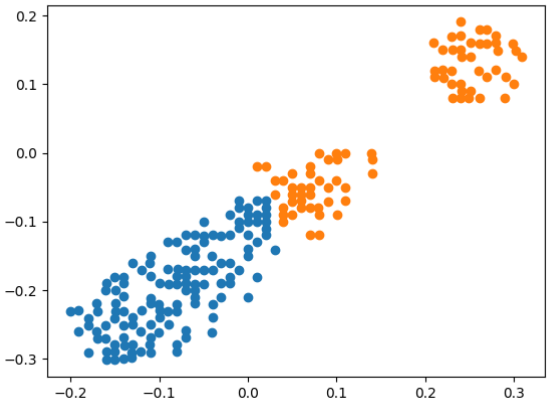
1.3 Data3

1.3.1 single linkage



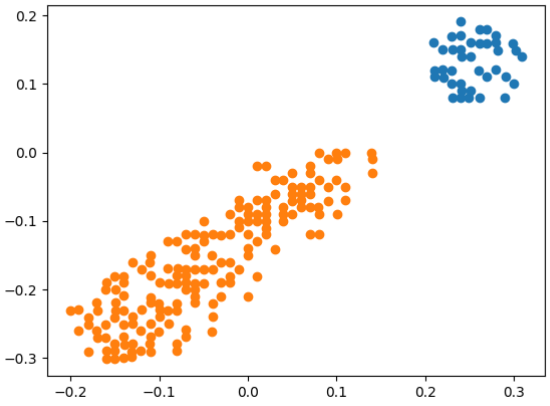
This dataset consist of 2 compact shapes apart from each other. Single linkage is suitable to cluster the shapes that has no connection.

1.3.2 complete linkage



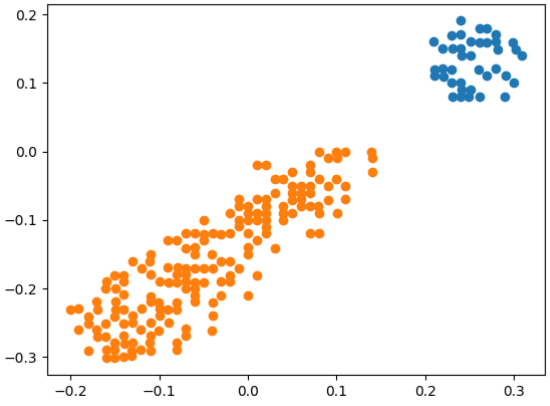
Since the method checks the worse distance between 2 clusters, upper part of the line is merged with the circular shape which is closer to itself rather than bottom part of the line according to the method. While the line gets longer, parts of the line can jump to nearset shape to it, so this method would not be suitable for sets that consists of long lines.

1.3.3 average linkage



Although it clustered correctly, if the line was longer, upper part of the line would be clustered with the circular shape.

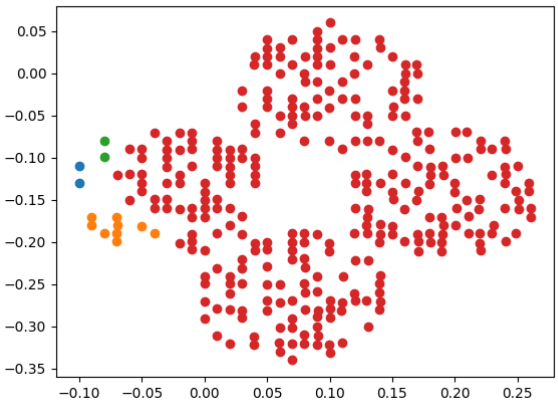
1.3.4 centroid



The centers of the shapes are apart from each other. But, if the line shape was not compact enough and longer, blue cluster could jump to line shape. Although it clustered this dataset correctly, centroid method is not a very good selection for dataset with lines.

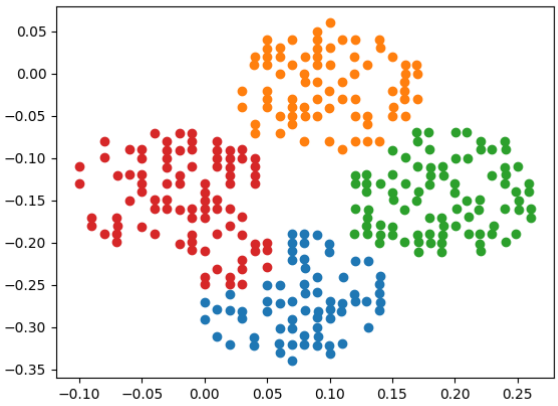
1.4 Data4

1.4.1 single linkage



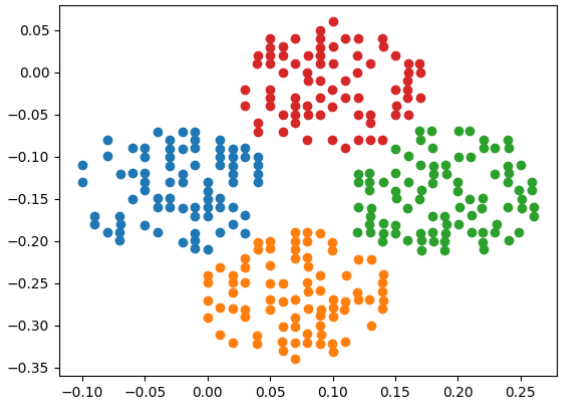
The shapes are touching each other, since single linkage clusters the shapes by looking nearest points of the groups, it jumps easyly to neighbor shape in this dataset.

1.4.2 complete linkage



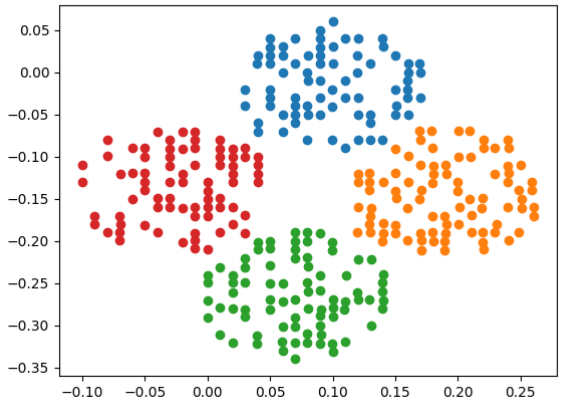
Since the shapes are circular, the method performs better than other datasets in this one. Nonetheless, it is not perfect.

1.4.3 average linkage



Since the shapes are circular and compact, average linkage is a good choice for this dataset. This method is suitable for datasets that consists of compact amoeba shapes

1.4.4 centroid



Since the geometric centers of the compact shapes are apart from each other and the logical cluster shapes are circular, centroid method is a good choice to cluster this dataset.

Strategy	Accuracy on Test Set
Information Gain	0.05
Gain Ratio	0.61
Average Gini Index	0.70
Chi-squared pre-pruning	0.60
Reduced Error post-pruning	0.65

Table 1: Accuracies on test set using different strategies

2 Part2: Decision Tree

2.1 Information Gain

Using information Gain as the strategy gave poor results on given sets. This is because information gain is biased towards choosing attributes with a large number of values. That causes overfitting or fragmentation problems.
(tree representation file: part2Trees/information_gain_tree.txt)

2.2 Gain Ratio

Gain ration is a modification of the information gain that reduces its bias towards multi-valued attributes. Therefore, gain ration gave much better accuracy on test set.
(tree representation file: part2Trees/gain_ratio_tree.txt)

2.3 Average Gini Index

Gini index is an alternative to information gain. It gave better test accuracy than other strategies.
(tree representation file: part2Trees/average_gini_index_tree.txt)

2.4 Gain Ratio with Chi-squared Pre-pruning

Stops growing the tree when there is no statistically significant association between any attribute and the class at a particular node. Pruned the tree but accuracy did not change much.
(tree representation file: part2Trees/preprining_tree.txt)

2.5 Gain Ratio with Reduced error post-pruning

Since the validation set is chosen randomly from the training set, pruning result changes one run to another. By using randomly chosen 5% of the training set as validation set, test accuracy was 65% in average while the training accuracy was 61% in avarege. Over 70% test accuracy was observered in some runs. Moreover, Resulting trees were quite simple wrt. the other results.
(tree representation file: part2Trees/postpruning_tree.txt)