Report HW2

Sinan Talha KOŞAR - 2099190

Task 1

(a) What are the shortcomings of the Euclidean distance?

When we use Euclidean distance, the meaning of features go away, Euclidean distance doesn't make sense when our data is discrete, because it is the line of sight distance between the points. The problem is at the level of measurement scale. An example of problematic dataset for Euclidean distance, if we have mixed data having height, length etc as numbers but also color feature which is not number. If the data have a correlation structure Euclidean distance is not the appropriate metric.

(b) Why does the dataset trigger the shortcomings of the Euclidean distance?

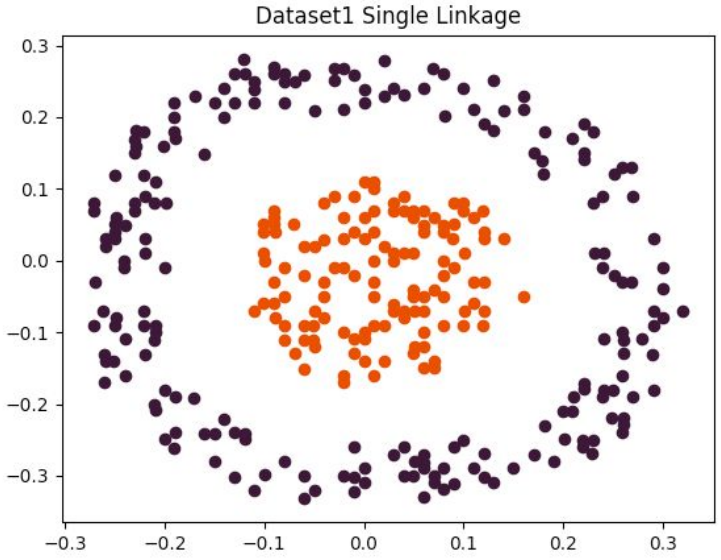
Features have no commons and difference of distances are small so it triggers the disadvantages of Euclidean distances.

(c) How can the dataset be preprocessed so that the test set accuracy improves?

Increasing the number of training example does not guarantee gaining better results. To enhance the accuracy of your model, we may need to do feature selection. We might need to scale your features before to use them with the selected classifier. Leave-One-Out method could overcome the problem, if we have limited data

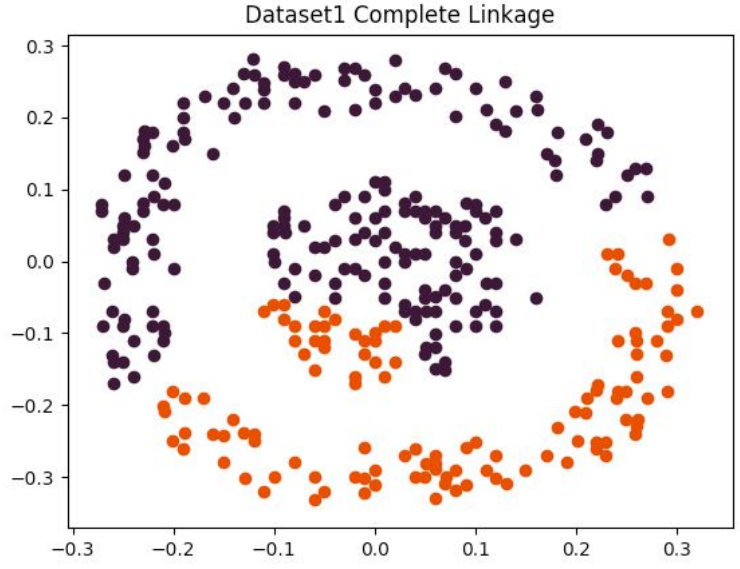
Task 2

Dataset 1

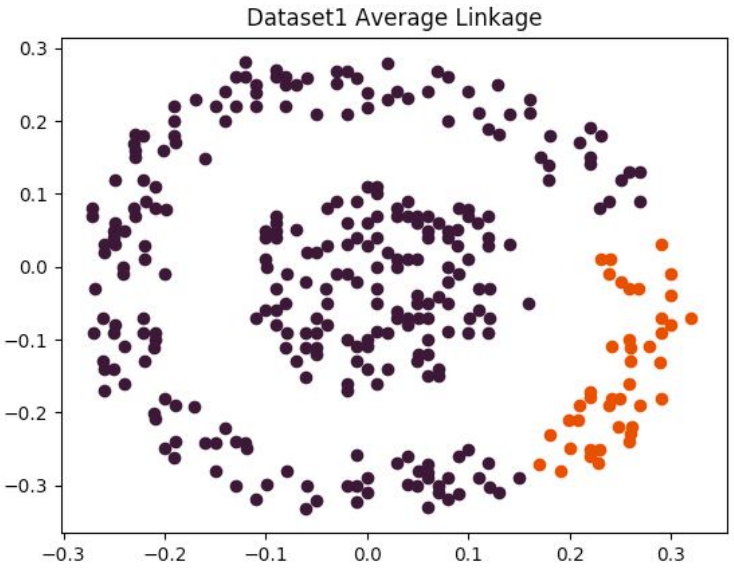


The data distributed like elongated shape of the pinwheel lobes, therefore it is suitable for this dataset

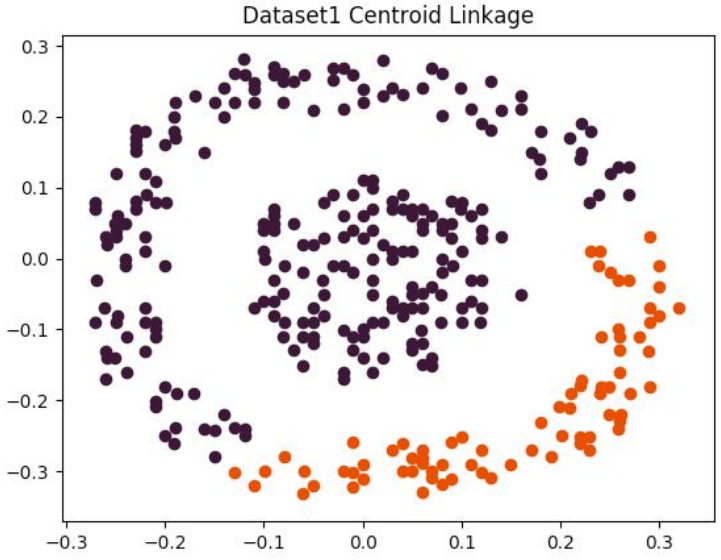
---



As it requires *all* of the distances to be small, it is not suitable for this dataset.

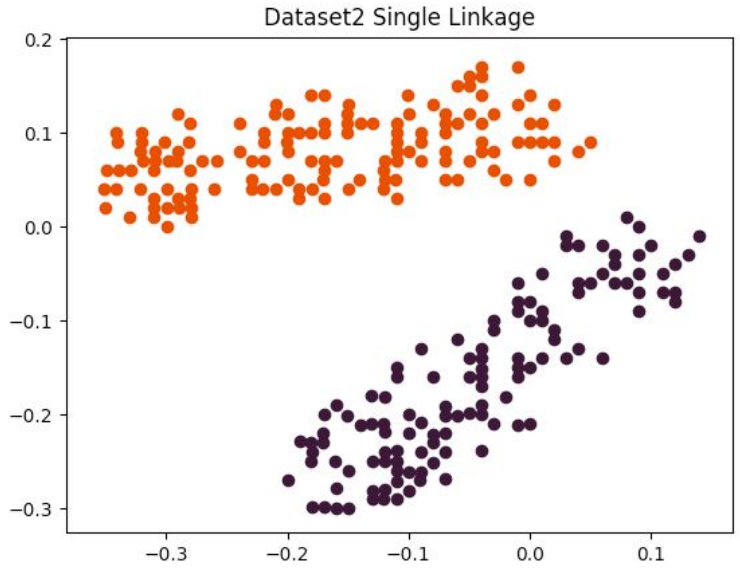


Although it is good for elongated shapes, the clusters are not compact, so it is not suitable for this data set.

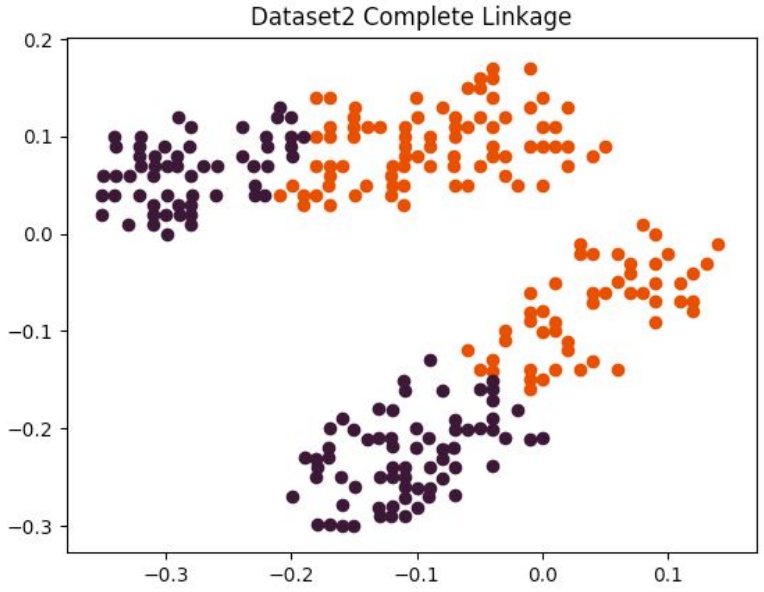


It is not suitable since centroid linkage only makes sense if an average linkage of data items is sensible.

Dataset 2

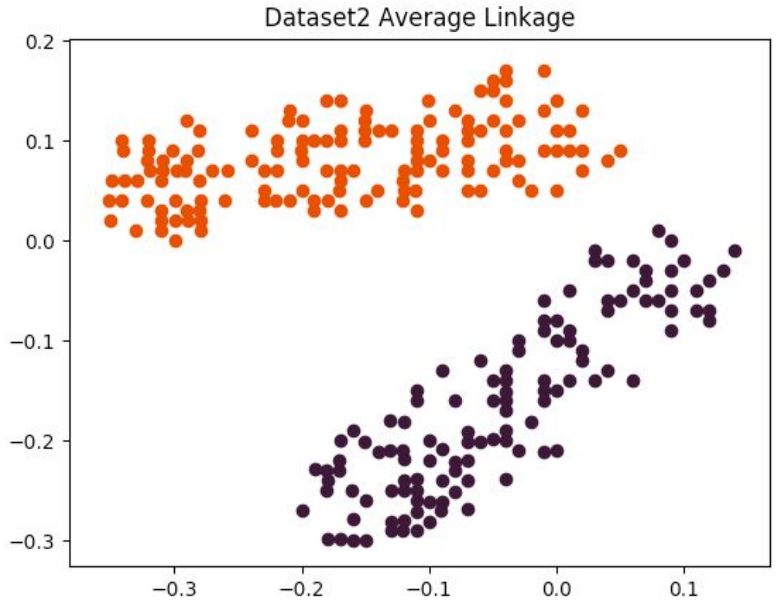


Single linkage decrease minimum distance much, it is suitable since the data are not spread closer to each other.



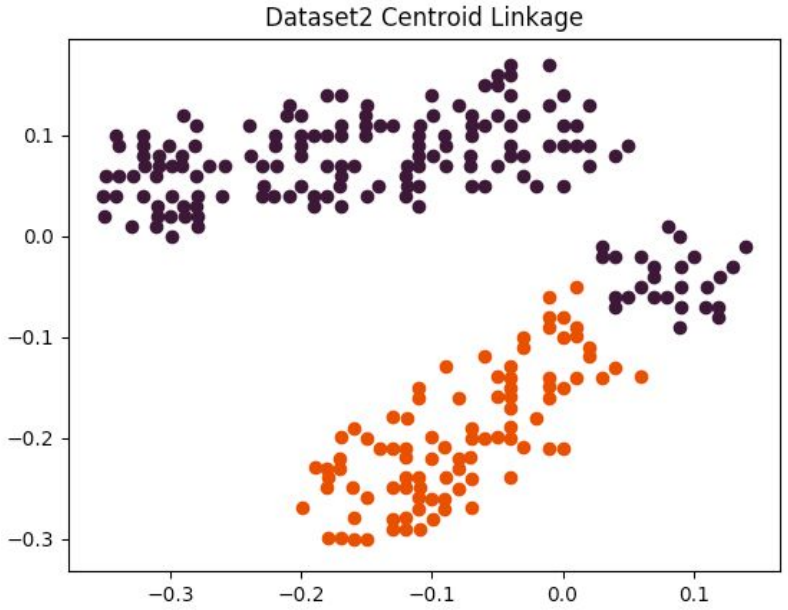
Since complete linkage increases the maximum distance much, it is not suitable for this dataset

---



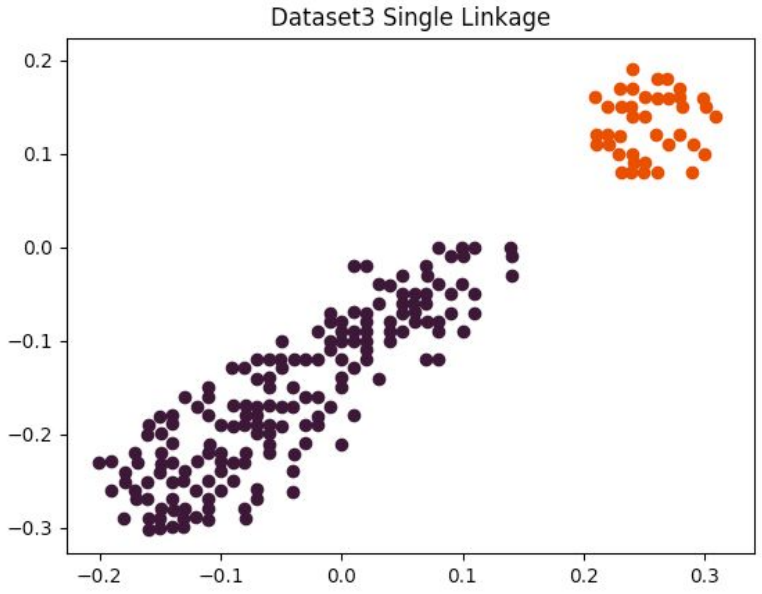
The dataset is compact elongated shaped, so it is suitable for this data

---

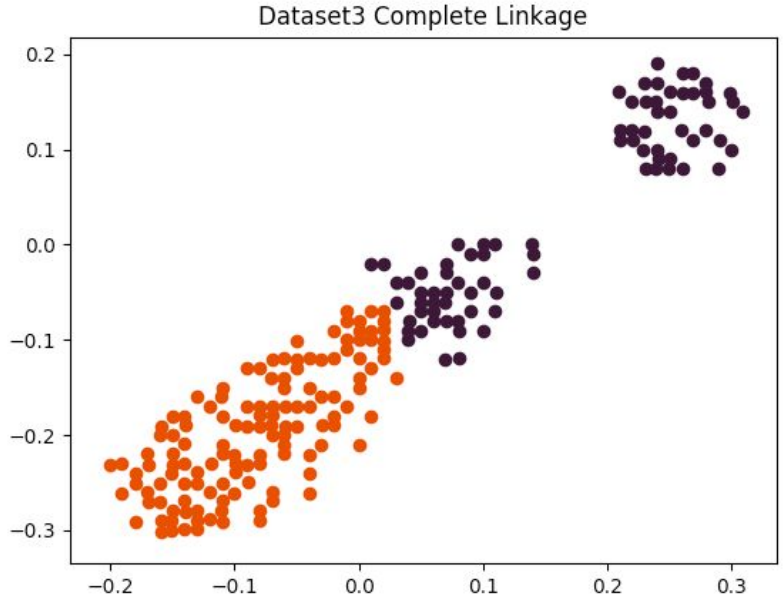


The cluster's geometric centroids are not obvious, so it is not suitable for this dataset.

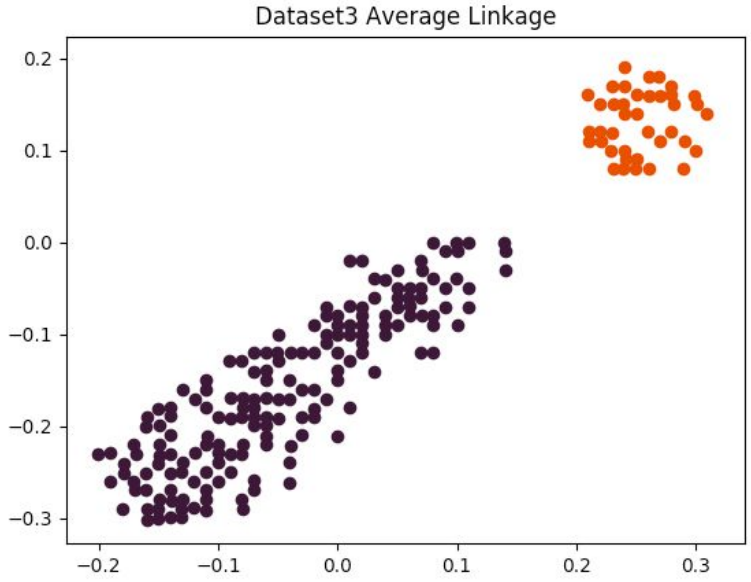
Dataset 3



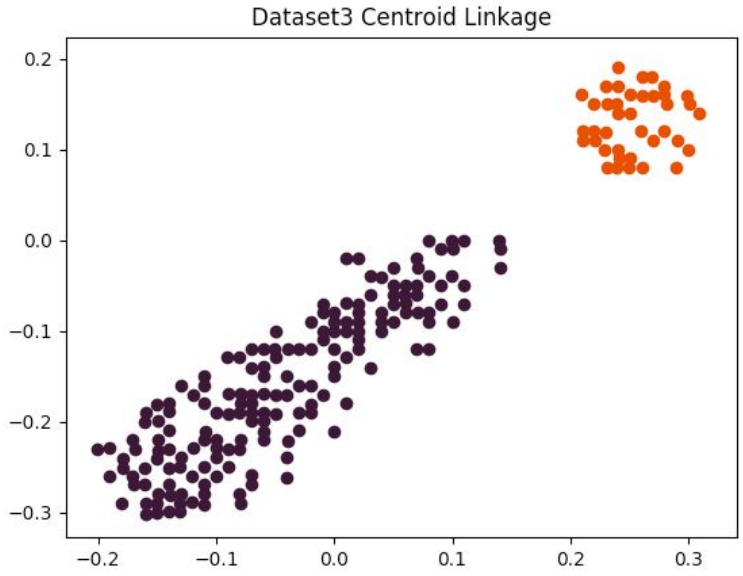
One cluster is like dataset1 and other like dataset2, so both advantages are met for single linkage, therefore it is suitable for this dataset.



Two most distant from each other members are much more dissimilar than other (as there exists circle), so it is not suitable for this dataset.

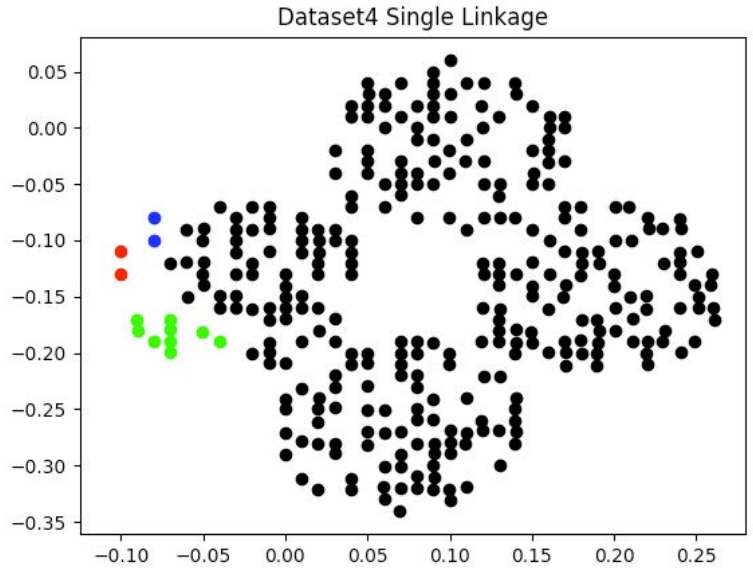


Dataset is compact elongated shape, so it is suitable for this dataset.

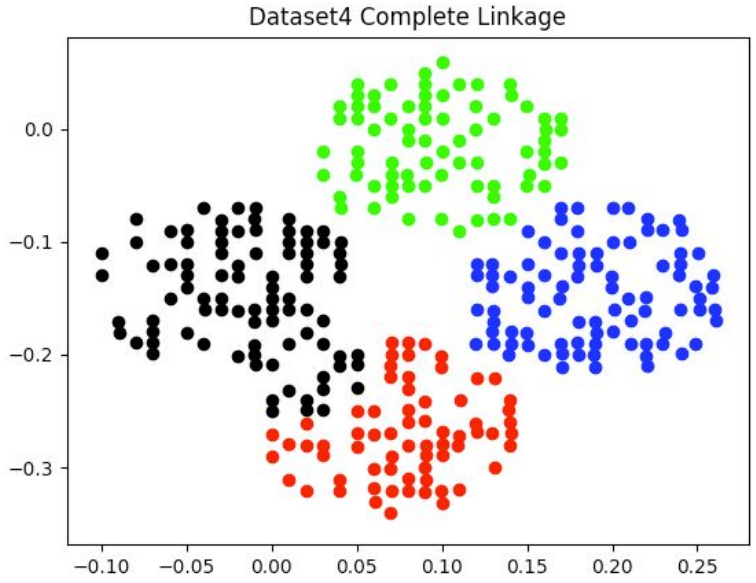


Each geometric centroids are obvious and average linkage is suitable so it is suitable for this dataset.

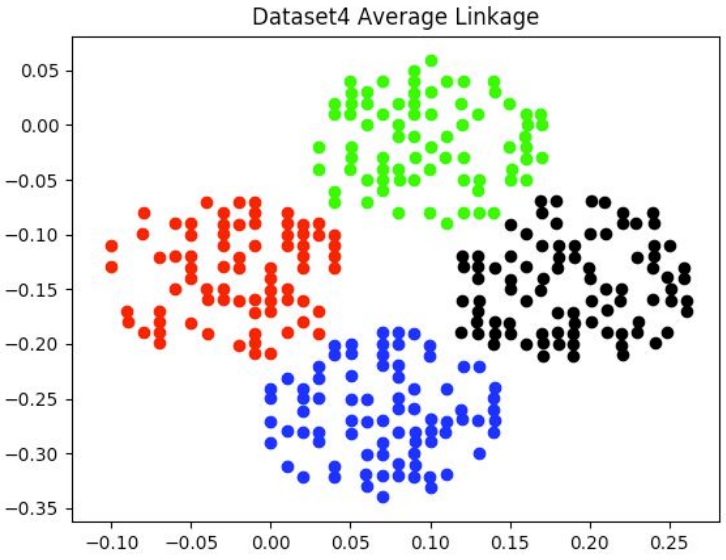
Dataset 4



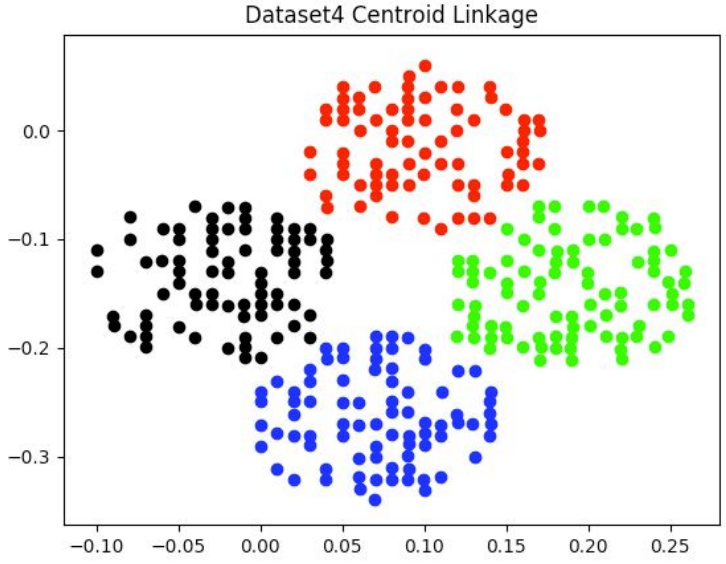
Although it is goog for elongated shape of pinwheel lobes, this resulted in premature merging of clusters in the tree, so it is not suitable for this dataset.



The borders of circles are more similar since complete linkage maximize the distance, it is not suitable for this dataset, but more suitabşe than single linkage in this dataset.



Since it is compact clusters that have some elongated shapes, it is suitable for this dataset.



Since average linkage is suitable and geometric centroids are obvious for circle shaped clusters, it is suitable for this dataset