Report

Azat Tologonov

15/12/2020

1 Part 1: K-Nearest Neighbor

1.1 K-fold Cross-validation

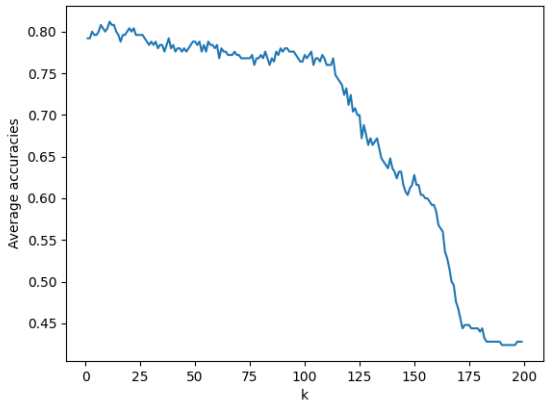

Figure 1: K-fold Cross-validation average accuracies.

1.2 Accuracy drops with very large k values

With very large k values too many points are neighbors, even ones that are far away. And since we use majority vote approach to decide the label, the prediction is no more based on nearest neighbors but on sizes of classes. Major class wins each vote.

1.3 Accuracy on test set with the best k

Best k: 11
Accuracy: 0.83

2 Part 2: K-means Clustering

2.1 Elbow method

Using the elbow method and figures below we can decide that the suitable k for the:

- clustering1 is 2.
- clustering2 is 3.
- clustering3 is 4.
- clustering4 is 5.

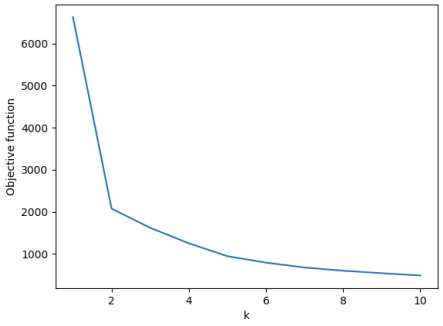

Figure 2: clustering1.

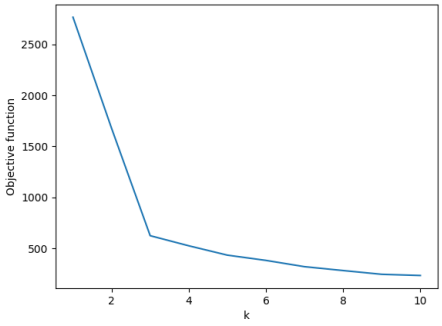

Figure 3: clustering2.

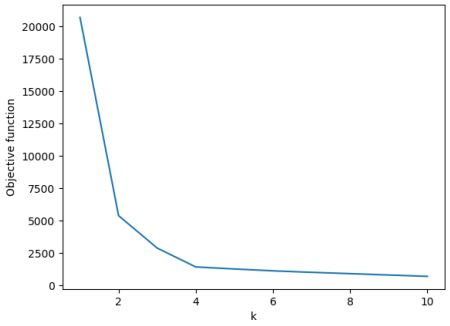

Figure 4: clustering3.

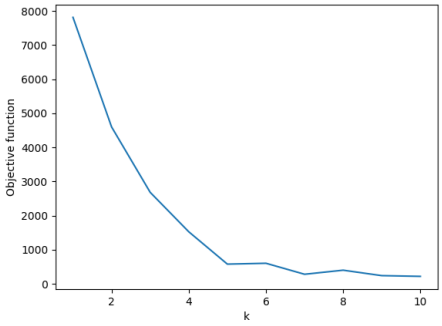

Figure 5: clustering4.

2.2 Resultant Clusters

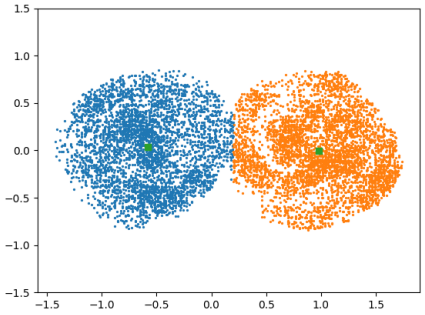

Figure 6: Clustering1.

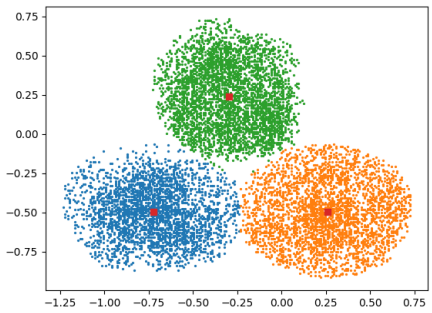

Figure 7: Clustering2.

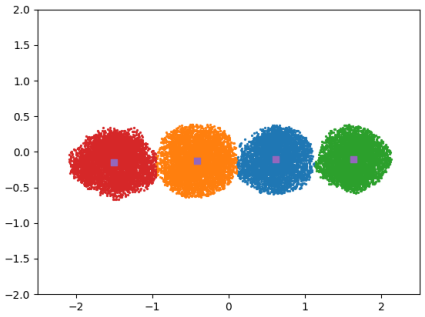

Figure 8: Clustering3.

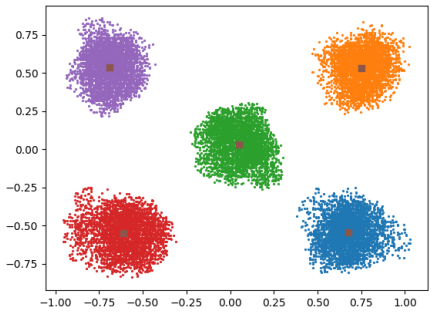

Figure 9: Clustering4.

3 Part 3: Hierarchical Agglomerative Clustering

3.1 data1

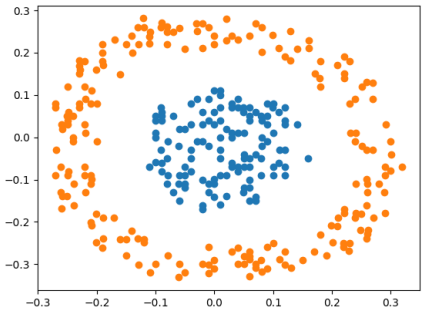

Figure 10: Single-Linkage Criterion

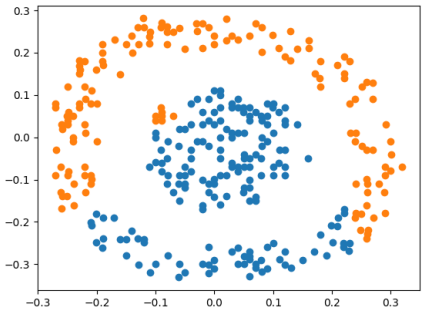

Figure 11: Complete-Linkage Criterion

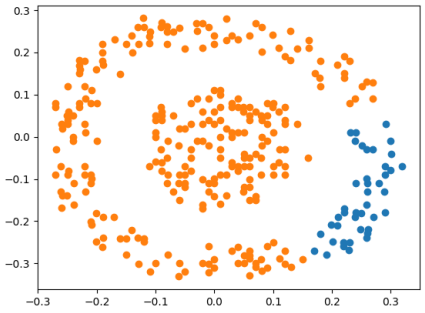

Figure 12: Average-Linkage Criterion

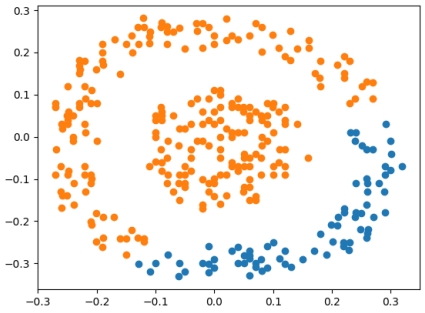

Figure 13: Centroid Criterion

- **Single-Linkage Criterion:** As we can see, data1 is naturally divided into 2 clusters that are separated from each other by the distance between them. There is closely positioned group of data in the center and another closely positioned group around the first one. As Single-Linkage criterion merges groups based on the shortest distance and has chaining effect it was suitable for data1.
- **Complete-Linkage Criterion:** This criterion was less suitable. It tries to make more compact clusters so it was not able to make natural shapes of data1.

- **Average-Linkage Criterion:** This criterion was not suitable at all. It averages all the distances between pairs and such approach does not make any sense for data1 because of the specific shape of the "external" cluster. Average of data items is not sensible.
- **Centroid Criterion:** It uses distance between centroid of groups of data. This criterion was not suitable at all because of the same reason as the Average-Linkage criterion.

3.2 data2

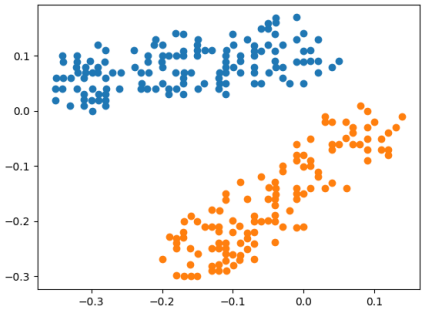

Figure 14: Single-Linkage Criterion

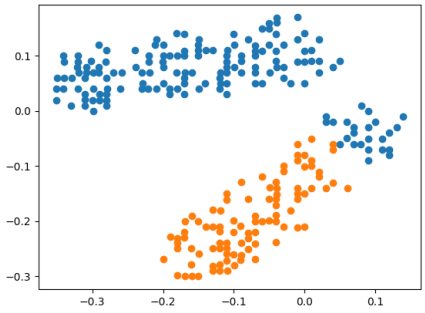

Figure 15: Complete-Linkage Criterion

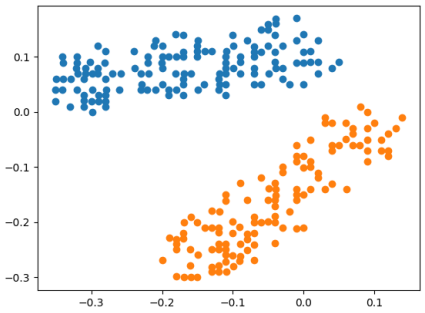

Figure 16: Average-Linkage Criterion

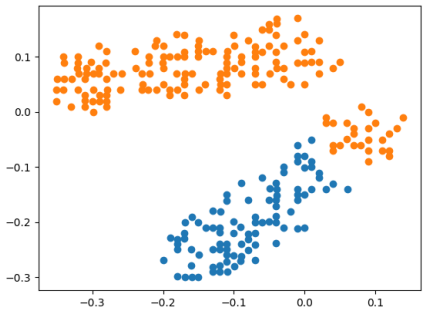

Figure 17: Centroid Criterion

- **Single-Linkage Criterion:** This criterion suited well for that dataset. Data groups have elongated shape and Single-Linkage criterion was able to capture it. Also, two groups have some distance between them and as this criterion uses shortest distances we were able to differentitate them.

- **Complete-Linkage Criterion:** This criterion tried to make more compact clusters and it ended with wrong clustering. It is not able to capture elongated shapes.
- **Average-Linkage Criterion:** This criterion was suitable for the data2. It produces compact clusters but still can have elongated shapes. And it is exactly how data2 looks.
- **Centroid Criterion:** Data2 have elongated shapes and computing the distance between centroids does not male sense. That is why Centroid criterion was not suitable.

3.3 data3

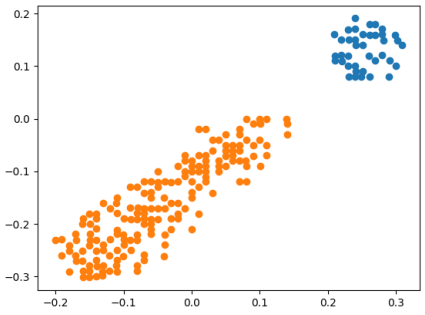

Figure 18: Single-Linkage Criterion

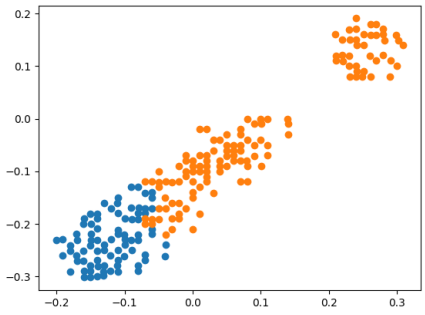

Figure 19: Complete-Linkage Criterion

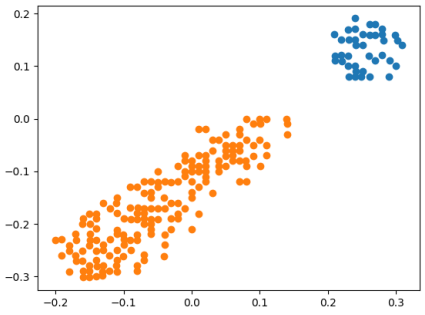

Figure 20: Average-Linkage Criterion

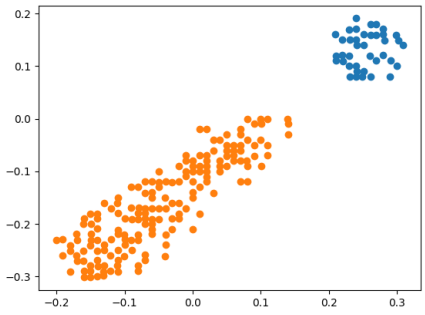

Figure 21: Centroid Criterion

- **Single-Linkage Criterion:** This criterion was suitable for the data3. It uses shortest distances and this dataset has a shape of exclamation mark

with two separated groups, one is elongated and other is circular, with large distance between them.

- **Complete-Linkage Criterion:** This criterion tried to make compact clusters and failed to capture the shape of the dataset. It is not suitable for the elongated shapes.
- **Average-Linkage Criterion:** Because of the large distance between two groups of data and small distances within each group this criterion was suitable. It took average distances between pairs and merged data correctly.
- **Centroid Criterion:** This criterion was suitable for the data3 because of the same reasons as Average-Linkage criterion. Two groupf of data are separated well and there is large distance between their centroids.

3.4 data4

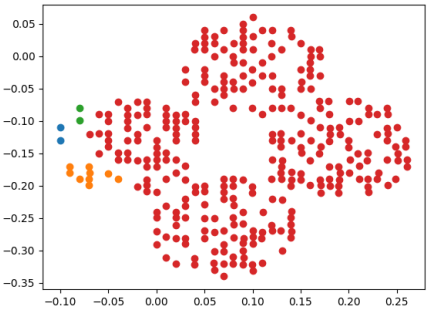

Figure 22: Single-Linkage Criterion

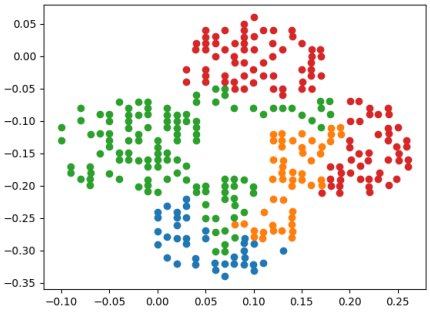

Figure 23: Complete-Linkage Criterion

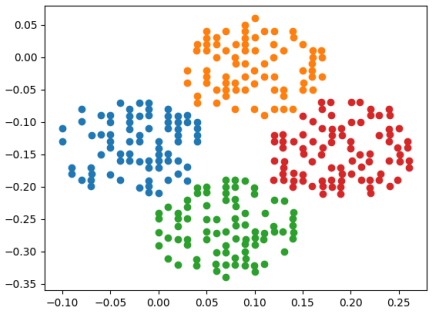

Figure 24: Average-Linkage Criterion

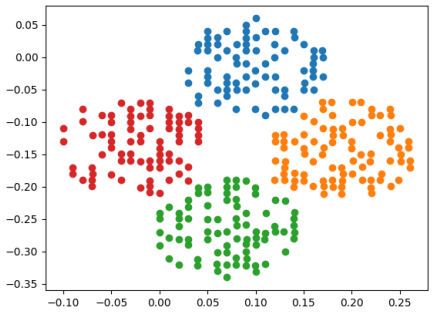

Figure 25: Centroid Criterion

- **Single-Linkage Criterion:** All the data is this dataset is located close to each other and because of that it resulted in the premature merging. This criterion was not suitable at all.
- **Complete-Linkage Criterion:** This criterion tried to make very compact clusters and was not able to capture real shapes of the clusters.
- **Average-Linkage Criterion:** Dataset can be divided into 4 groups with the shape of a circle. This criterion was suitable for the data4. It averaged all possible pairs between the groups and captured shapes correctly.
- **Centroid Criterion:** There are well defined centroids of the 4 groups of data. Because of that Centroid criterion was suitable for this dataset.