Report

Emil Shikhaliyev

15.05.2021

1 Part 1: K-Nearest Neighbor

1.1 K-fold Cross-validation

Below you can find the graph of k-fold cross validation with average accuracies with respect to k values, such as k = 1, 2, 3, ..., 199.

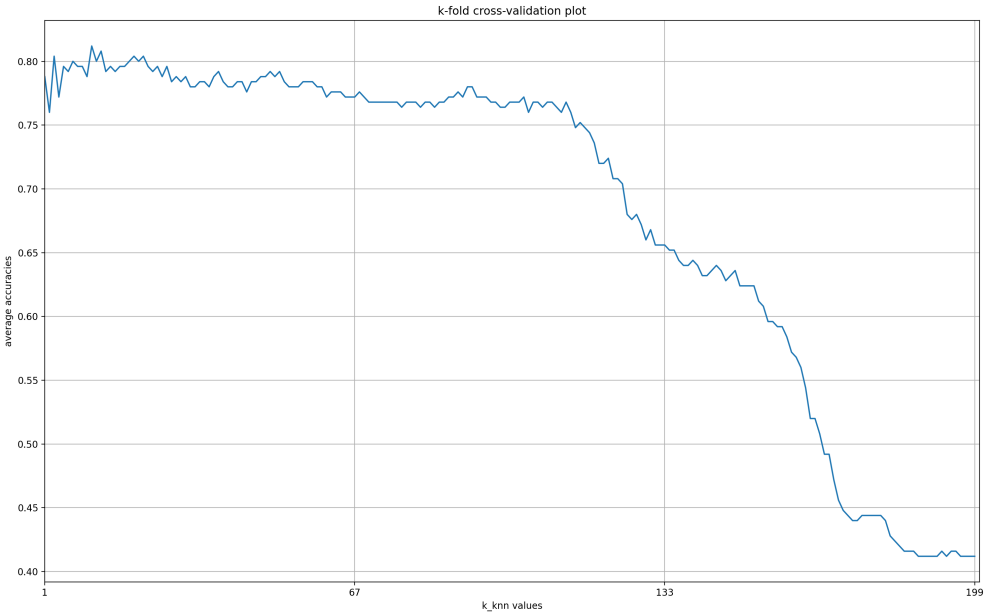


Figure 1: k-fold cross validation plot

1.2 Accuracy drops with very large k values

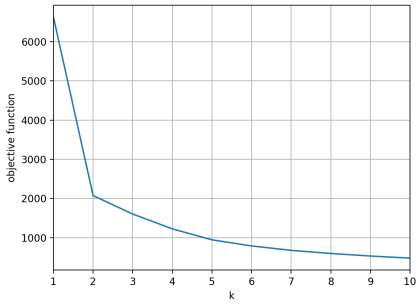
Because, with large k values, overfitting happens.

1.3 Accuracy on test set with the best k

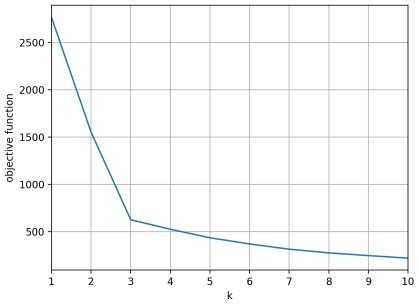
The best k_{KNN} value is **11**. Test set accuracy value is **0.82**, when $k_{KNN} = 11$ was used to calculate it.

2 Part 2: K-means Clustering

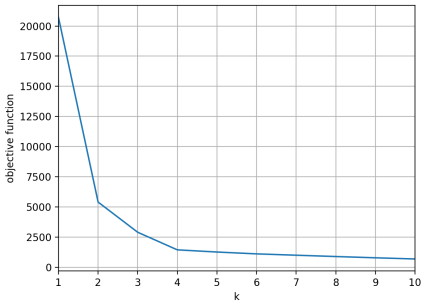
2.1 Elbow method



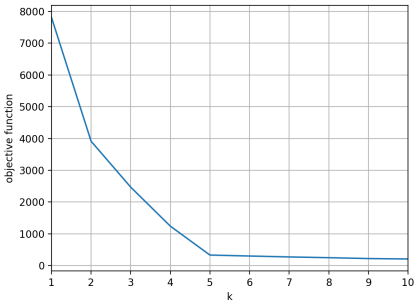
(a) Clustering 1 (suitable k=2)



(b) Clustering 2 (suitable k=3)



(c) Clustering 3 (suitable k=4)

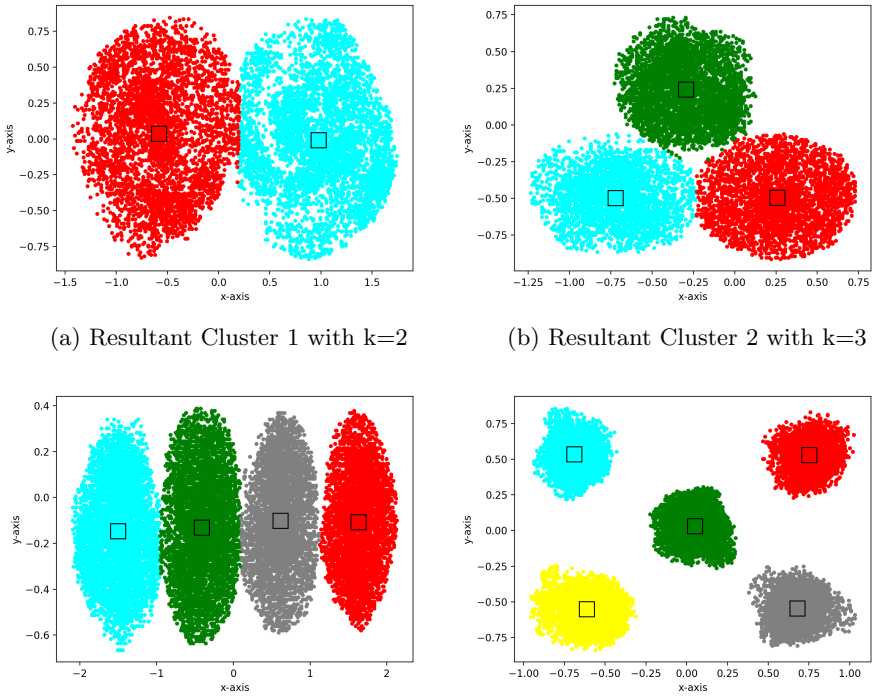


(d) Clustering 4 (suitable k=5)

Figure 2: Plots of Elbow method on Clusters

2.2 Resultant Clusters

Below, you can find plots of resultant clusters with suitable k values.



(a) Resultant Cluster 1 with k=2

(b) Resultant Cluster 2 with k=3

(c) Resultant Cluster 3 with k=4

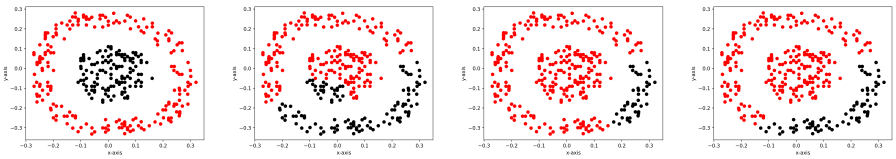
(d) Resultant Cluster 4 with k=5

Figure 3: Resultant Clusters with suitable k values

3 Part 3: Hierarchical Agglomerative Clustering

3.1 data1

Single-Linkage criterion - It's the best criterion for this kind of data, because distances between the points in the inside circle and the points in surrounding circle are far enough.
Complete-Linkage criterion - It doesn't produce suitable output for this kind of data, because some points in one cluster are closer to points in another cluster, than to points in the same cluster.
Average-Linkage criterion - It doesn't produce suitable output also, because of the same reason with Complete-Linkage criterion. We can observe that, the black part is far away from central round than the red parts in surrounding cluster.
Centroid criterion - It doesn't produce suitable output also, because centers of both clusters are too close, that's why we can not get right answer.

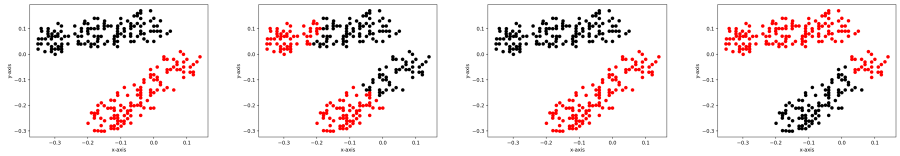


(a) Single-Link. (b) Complete-Link. (c) Average-Link. (d) Centroid

Figure 4: Plots of resultant clusters of data1 using each criterion

3.2 data2

Single-Linkage criterion - It has best result, because distances between points of one cluster are far enough from points of other cluster.
Complete-Linkage criterion - It doesn't have suitable output for this kind of data, because some points, which are in different clusters, are closer each other than points in the same cluster.
Average-Linkage criterion - It has best result, because two clusters are clearly separated.
Centroid criterion - It doesn't have suitable output for this data, because red part in mostly black cluster is a little bit far from the black part. That's why that part is red.

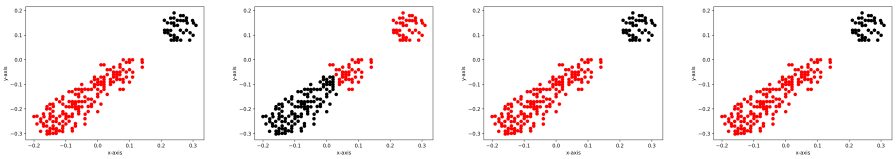


(a) Single-Link. (b) Complete-Link. (c) Average-Link. (d) Centroid

Figure 5: Plots of resultant clusters of data2 using each criterion

3.3 data3

Single-Linkage criterion - It has the best result for this kind of data, because the distances between points in clusters are far away.
Complete-Linkage criterion - It is not suitable for this kind of data, because the some points in long cluster are close to end of smaller cluster than the end of long cluster.
Average-Linkage and Centroid criterion - It has the best result also for this kind of data, because two clusters are separated clearly.

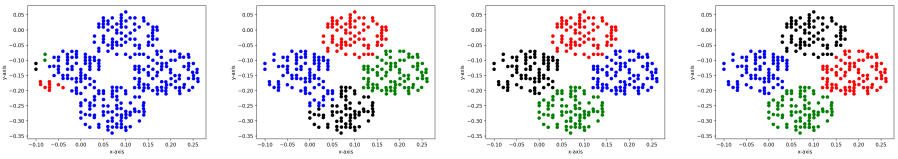


(a) Single-Link. (b) Complete-Link. (c) Average-Link. (d) Centroid

Figure 6: Plots of resultant clusters of data3 using each criterion

3.4 data4

Single-Linkage criterion - It produces the worst output for this kind of data, because the borders of each clusters are near to each other.
Complete-Linkage criterion - It produces better result than Single-Linkage criterion, but it is not perfect, because you can see there're some blue points in black cluster. It's about distribution of points between blue and black clusters.
Average-Linkage criterion - It has best result in this data, because distribution of points and the shapes of clusters are suitable for Average-linkage criterion, which is round.
Centroid criterion - It has the another best result in this data, it is also about the shape of clusters, which is round shape.



(a) Single-Link. (b) Complete-Link. (c) Average-Link. (d) Centroid

Figure 7: Plots of resultant clusters of data4 using each criterion