Report

Ahmet Dara Vefa

20.12.2020

1 Part 1: K-Nearest Neighbor

1.1 K-fold Cross-validation

Check the included images KNN/KnnPlot.png and KNN/KnnPlotZoomed.png for the plots.

1.2 Accuracy drops with very large k values

The data becomes more and more biased towards the group(class) with most members. As we get closer to K=N(data count) the model will only predict the result to be the class with the most members.

1.3 Accuracy on test set with the best k

By looking at the KnnPlotZoomed.png we can clearly see the second K value(K=3) gives the highest average accuracy. After plugging K=3 to our test method we get "accuracy on test set=77.0%" as output

2 Part 2: K-means Clustering

2.1 Elbow method

You can see the plots in the KMeans folder. Using elbow method, the best Kvalues are:
2 for clustering1.
3 for clustering2.
4 for clustering3. I didn't choose K=2 because average objective function value changes too greatly(about 2500) from K=2 to K=3 and K=3 to K=4.
5 for clustering4.

2.2 Resultant Clusters

You can find these plots in the KMeans folder.

3 Part 3: Hierarchical Agglomerative Clustering

You can find these plots in HAC folder

3.1 data1

Average linkage gives a wrong result. The reason the center of the data is included in cluster1 is because the average distance of outer cluster1 vs inner data is just a little bit less than average distance of cluster0 vs inner data.

Centroid linkage gives wrong result. This result includes more data in cluster0 because the center of cluster1 is more towards top left. There is no way this linkage would give the correct result because the center of the outer cluster falls at the center cluster.

Complete linkage gives a better result but it still is wrong. This linkage also can't give the correct result in this type of topology since the distance between two ends of the outer cluster is too large.

Single linkage gives the desired result. This happens because the distance between each data point in outer and inner cluster is larger than the distance between each data point within each cluster.

3.2 data2

Average linkage gives desired result. This happens because the average distance between data points of cluster0 and cluster1 is larger than average distance between each data point within each cluster.

Centroid linkage gives wrong result. This one is interesting because if the top right of the (correctly clustered)bottom cluster was lower this linkage would also give the correct result. However the top right of the (correctly clustered)bottom cluster is further away from the center of the cluster1 than the center of the cluster0.

Complete linkage also gives wrong result in an interesting way. We can see that the furthest distance between each data point within each cluster is almost the same for cluster1 and cluster0, but to be honest I would have expected a bit different clusters because there are data points that are so close to each other but they are in different clusters. I would have expected those points to converge in the earlier stages of the algorithm.
Single linkage also gives the desired result, since the distance between each data point within each cluster is smaller than the minimum distance between each data point of each cluster.

3.3 data3

Only complete linkage gives wrong result, since the top right of the data is closer than the bottom left of the data to the 0.1,0 area.
One important note is that centroid linkage barely manages to give the correct result so I would not use that.

3.4 data4

Average and centroid linkages gives the correct results

Complete linkage gives wrong result because the data is not balanced, there are some points which are spaced out in the outer areas. I am still not sure if removing these outliers would give correct results because there are still some points which are too close to each other even though they are in different (correctly clustered)clusters

Single linkage finally fails, simply because (as stated previously) there are some points which are too close to each other even though they are in different (correctly clustered)clusters