Final Project Report

Project problem:
The project is Hierarchical Clustering for Seed Categorization. In this project, we Implement Hierarchical Clustering on the UCI seed dataset to divide it into cluster groups and use the cluster IDs as labels for a subsequent K nearest neighbor classifier to identify the species. We need to use multiple clusterings here and determine what a good number of clusters would be and how to choose the similarity between clusters and a data point.

Methodology:
First, read the UCI seed dataset. We need to take the features part of the dataset for the clustering. For clustering, we create a matrix of the size of the dataset. Calculate the euclidian distance for the data points and store them in the lower triangular matrix. Find the minimum value of the matrix those matrix row and column values will form a cluster. Update the matrix based on the linkage criteria. Repeat the clustering until the required number of clusters has been found.

After forming the cluster we need to find a good cluster for this we find the accuracy of the cluster by taking the max vote for each cluster and finding the accuracy. I have set the accuracy limit to be greater than 0.9. So when we get an accuracy greater than 0.9 then that cluster is considered a good cluster. The we find the cluster for that number.

Then we take an average point for each cluster and find the euclidian distance for that point to all other points and these distances form a new feature for the dataset.so if there is 5 cluster then we need to add 5 new features of euclidian distance to the dataset because we are comparing the distance from that datapoint to all other clusters.

Now we train a KNN classifier with the X and Y. For KNN we need to calculate the euclidian distance of the data point that we need to predict and to all the other data points and pick the K nearest data points and return the mode of the K nearest data points. So now these predicted values form the predicted Y value. Finally, with the actual Y and the predicted Y, we will find the accuracy of the classifier.

We need to run this code for different numbers of K values and find its accuracy. The higher the accuracy the better the model is. so based on accuracy, we can find the best clusters.

Experiments done:

We need to find a good number of clusters to do classification. So I have implemented clustering in Average (mean) linkage, Single (minimum) linkage, and Complete (maximum) linkage. For K values 3,5 and 7 we have found the accuracy

For the first experiment, I trained the classifier using all the data points and predicted labels for all the data points.

For the second experiment, I trained the classifier using the first 180 data points and predicted labels for the last 30 data points.

Results:

The accuracy result for the first experiment are

Without using any of the clusters the KNN is:

K 3:0.9

K 5:0.8666666666666667

K 7:0.8666666666666667

For Average (mean) linkage

```
a good cluster is
18
```

K 3:0.9380952380952381

K 5:0.9238095238095239

K 7:0.8904761904761904

For Single (minimum) linkage

```
a good cluster is
46
```

K 3:0.9333333333333333

K 5:0.9142857142857143

K 7:0.9142857142857143

For Complete (maximum) linkage

```
a good cluster is
13
```

K 3:0.9380952380952381

K 5:0.9238095238095239

K 7:0.9

The accuracy result for the second experiment are

For Average (mean) linkage

```
a good cluster is
18
```

K 3:0.9

K 5:0.8666666666666667

K 7:0.8666666666666667

For Single (minimum) linkage

```
a good cluster is
46
```

K 3:0.9

K 5:0.9

K 7:0.8666666666666667

For Complete (maximum) linkage

```
a good cluster is
13
```

K 3:0.9

K 5:0.8666666666666667

K 7:0.8666666666666667

Conclusion:

From the results, we can see there is an increase in accuracy for adding cluster features to the dataset. For the similarity between clusters and a data point, we have used Euclidian distance to calculate the similarity for the clusters.