CS498 Applied Machine Learning

Homework 4

Group Work

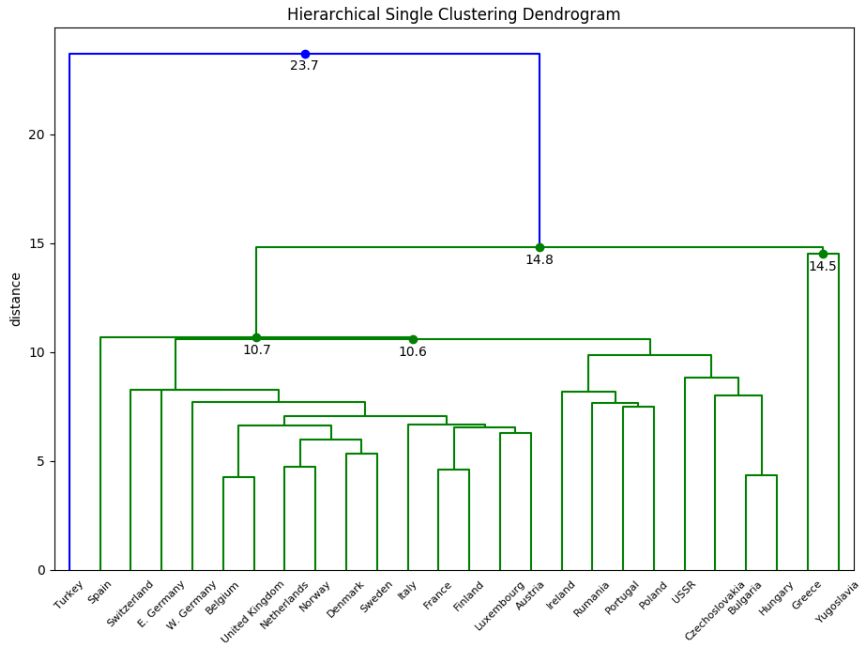
Team Leader
Name: Richard Wheeler
NetID: rw6

Team Member
Name: Jason Neal
NetID: jasoncn2

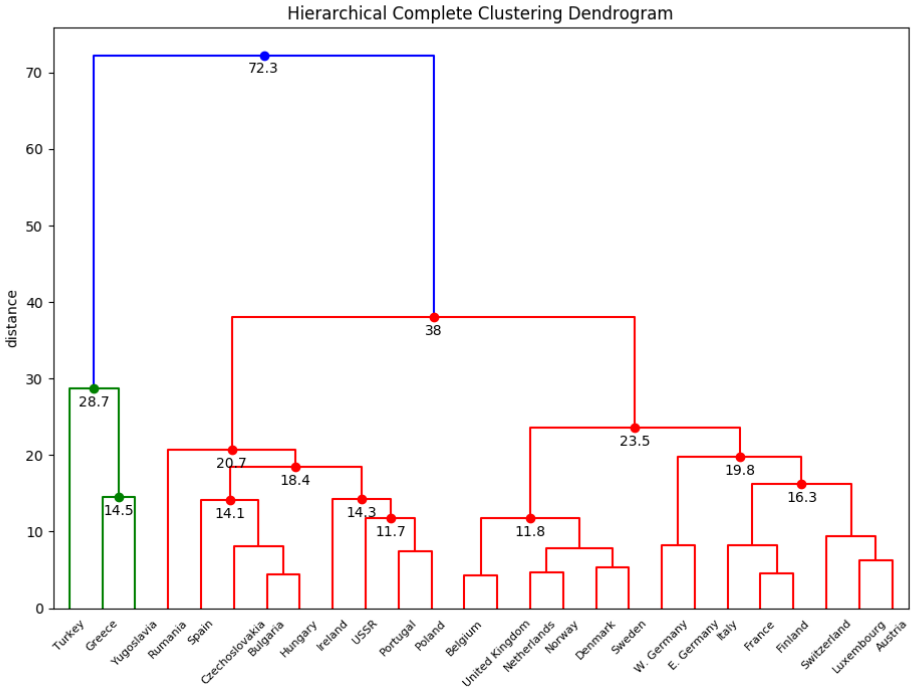
Name: Ling-Hsi Liu
NetID: lhliu2

Part1.1

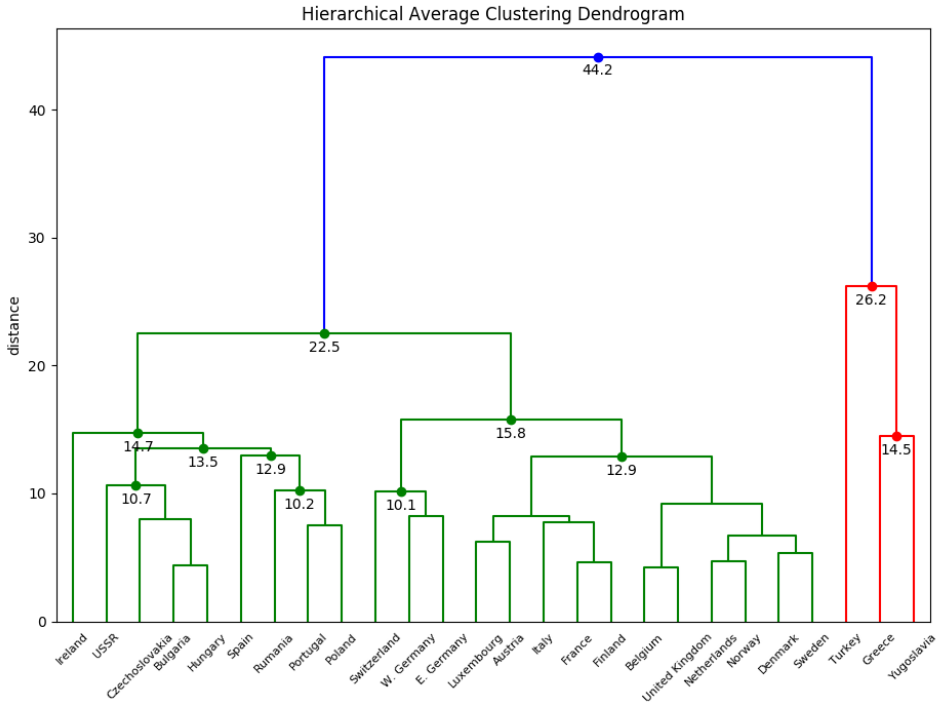
Hierarchical Single Clustering Dendrogram



Hierarchical Complete Clustering Dendrogram



Hierarchical Average Clustering Dendrogram

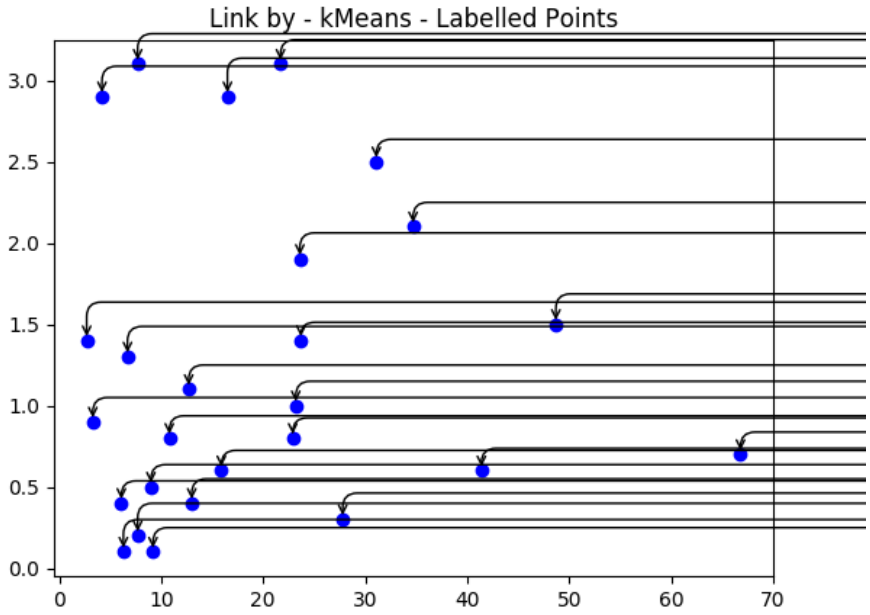


What structure in the data does each method expose?

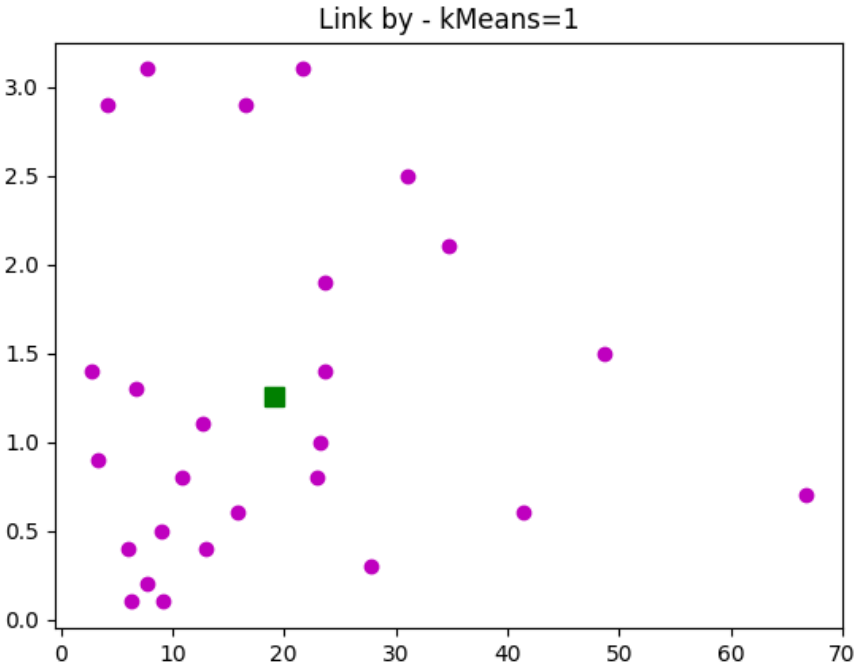
Exposes the average employment by industry which reflect the political geographic alignments of that time period in 1979. The complete clustering dendrogram shows this alignment most clearly. The average clustering dendrogram is 2nd best and single clustering the alignments are less clear but still distinguishable.

Part1.2

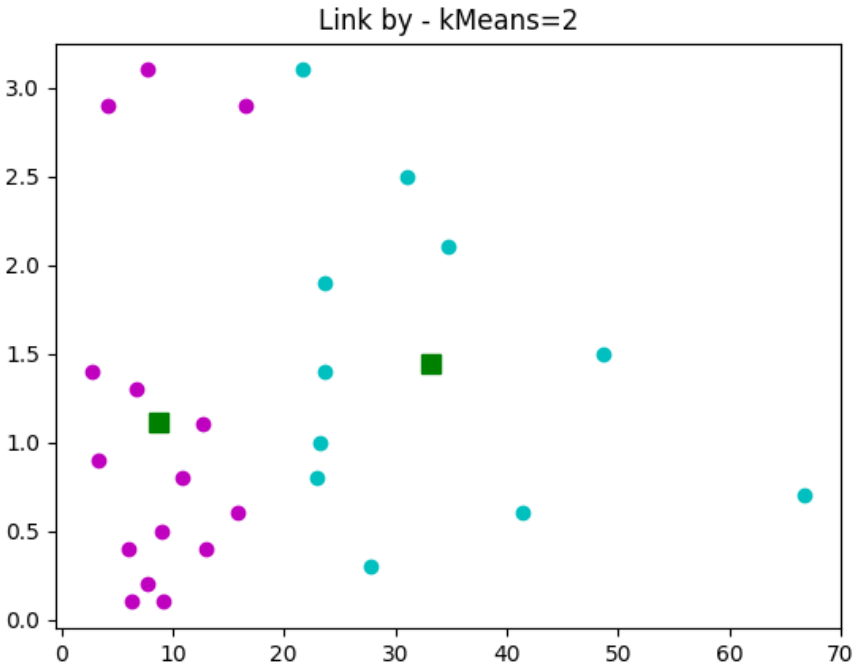
Link by – kMeans – Labelled Points



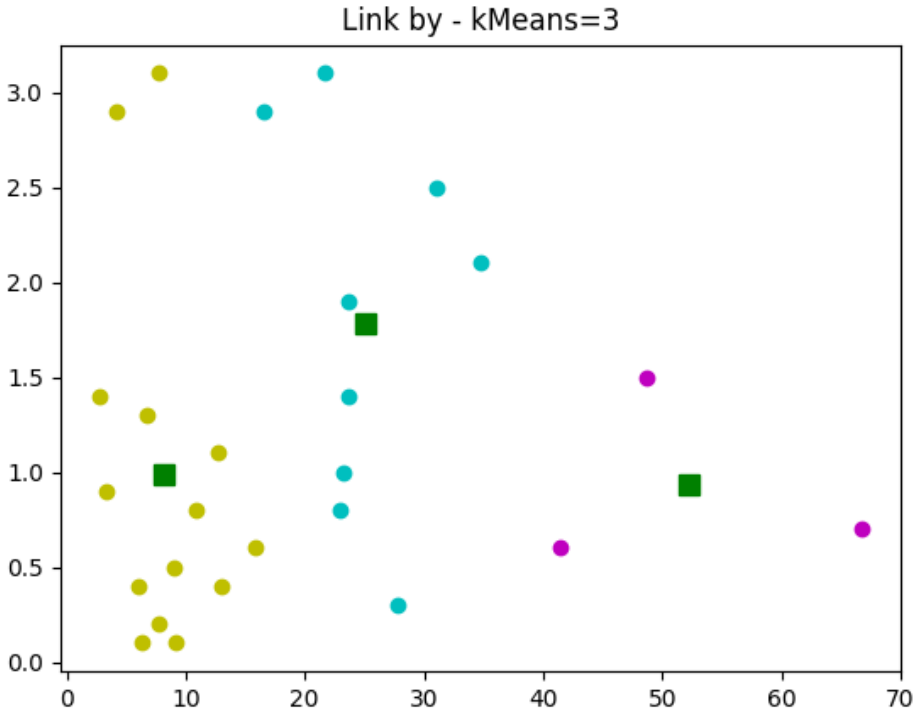
kMeans = 1



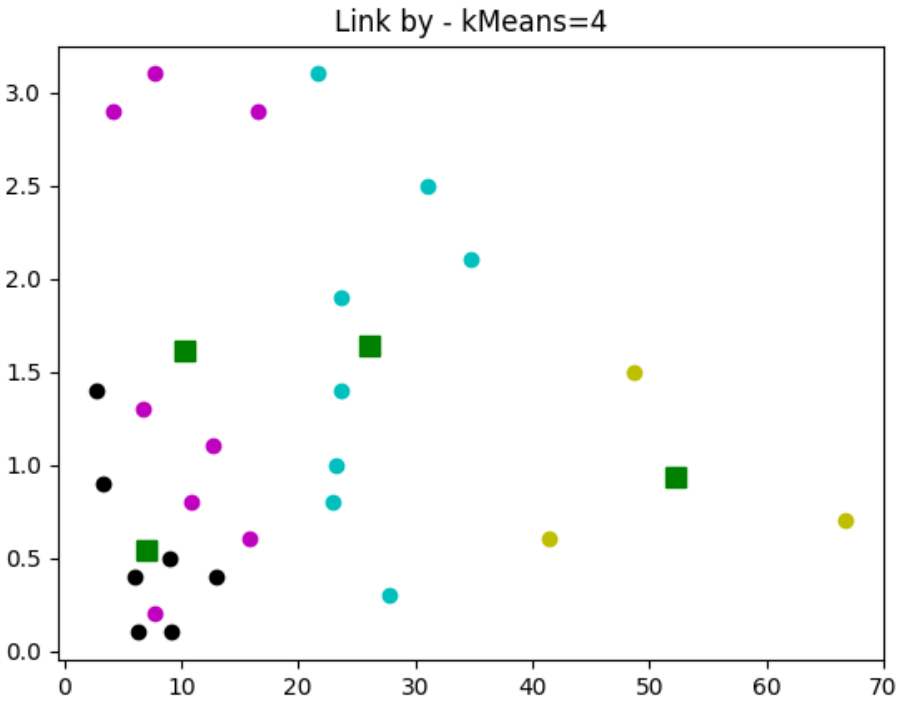
kMeans = 2



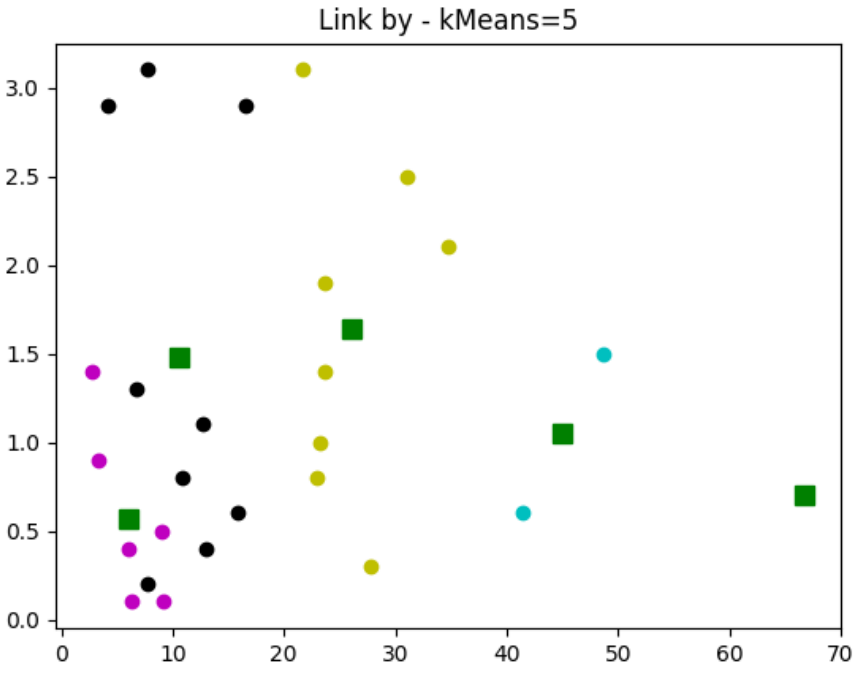
KMeans = 3



kMeans = 4


Link by - kMeans=4
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
60
70

kMeans = 5


Link by - kMeans=5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
60
70

What is a good choice of k for this data and why?

K=5 is a good choice.

When the KMeans is 5, we could clearly know each cluster. And, every data point is more closed to centroids. There are three points on the right side which are far from the left cluster. When the k-mean is bigger, the data points have short distant to centroids.

Part2

6.2. (a). (a) Total Error rate: 0.9624217118997912

(b) The class confusion matrix of your classifier is below:

cnf_mat

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0	0	0	3	0	0	1	0	4	0	0	2	0	0
1	0	0	0	0	0	0	0	0	65	0	0	0	0	0
2	0	0	0	0	0	0	0	0	25	0	0	0	0	0
3	0	0	0	0	0	0	0	0	25	0	0	0	0	0
4	0	0	0	0	0	0	0	0	59	0	0	0	0	0
5	0	0	0	0	3	0	1	0	0	0	0	0	0	0
6	0	0	0	1	0	0	0	0	0	0	0	1	0	0
7	0	0	0	0	0	0	0	0	73	0	0	0	0	0
8	0	0	0	0	0	0	0	0	18	0	0	0	0	0
9	0	0	0	0	0	0	0	0	76	0	0	0	0	0
10	0	0	0	0	0	0	0	0	18	0	0	0	0	0
11	0	0	0	0	0	0	0	0	16	0	0	0	0	0
12	1	0	0	0	0	0	1	0	8	0	0	0	0	0
13	0	0	0	1	0	0	3	0	72	0	1	1	0	0

Below is the approach we took for Problem 2.

1. Segmented data into 80/20 Train/Test sets
2. In the training data, segmented the signals into 'chunks' (first try was 32 to represent 1 second of activity)
3. Flattened the 32x3 arrays into 1x96 arrays to represent each segments x/y/z over the segment
4. Assigned each segment to nearest centroid
5. Built a histogram of centroids to define our classifier feature set for all 14 ADL's
6. Took the test data and transformed it similarly to training data by segmenting, flattening, and clustering to build out similar histogram features for each observation
7. Then built both a random forest and naive bayes classifier against our untouched test data using our trained feature set.
8. Once the classification/prediction was completed scored it and produced a confusion matrix measuring actuals vs predicted.

Additional efforts that we attempted to improve accuracies:

A. One was using both raw histogram counts as well as normalizing the histogram counts by dividing each element by the sum of the entire vector - this effort showed no improvement in our accuracy measurements.
B. Second effort was using random forest classifier by using the training histogram counts as predictor values against the test data centroids. Breaking out the centroids cause library function issues that could not be resolved by the homework deadline.

6.2(b)(a) modifying the number of cluster centers in your hierarchical k-means (b)modifying the size of fixed length samples that you use.

The additional efforts outlined above were attempts to complete this section. However the issues with the data structures and library calls prevented gathering any results.