Samet Aytac
1819044

ML HW3 Report

I want to start report with brief explanation of my function.

Euclid(x1,x2)

This function take 2 data instances as an input and it returns euclidean distance between them.

$$d = |\mathbf{x} - \mathbf{y}| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \ .$$

k_means(c,ccount,dataset)

This function take initial centers of clusters,desired clusters number and dataset. It calculate euclidean distance between every cluster center and every data instance. It assign data instance to cluster which nearest to it. Then, it update new cluster centers according to changing clusters. After,it checks if is there any changes in clusters or not. If there is no change,halt the process. Otherwise,it continues. It returns final center of clusters and U vector which show which data instance below which cluster.

hierarchicalc(cinit,ccount)

This function take desired clusters number and dataset as input. It thread like every data instance is a cluster. Then, it is try to find two nearest clusters. At that point, I use Ward's minimum distance method to calculate distances between clusters. After it finds two nearest clusters, it merge them and delete this two clusters. This process continues till desired class number is reached. It returns final center of clusters and U vector which show which data instance below which cluster.

hw3.m

In main script, first of all, I take inputs and merge them into two parts : data instances and their natural class. After, I assign initial center of clusters value. I choose those numbers because with this initial center of classes, separation between clusters is more clearly observed than any random c value. But, assign a random c value is OK.(code between commands). After, I call k_means. I choose desired cluster number as 6 but any other value is OK. [ck_means, Uk_means] are c and U values calculating by k means procedure.

After that, I call hierarchical clustering part. At that point, I arrange dataset with 200 elements because, to be fair, I cannot implement vectorization in this procedure very well, so with all dataset, it took a lot of time to return. I choose desired cluster number as 6 again. [c_hier,U_hier] are c and U values calculating by hierarchical clustering procedure.

After, I calculate purity of each cluster of calculating by hierarchical clustering procedure.

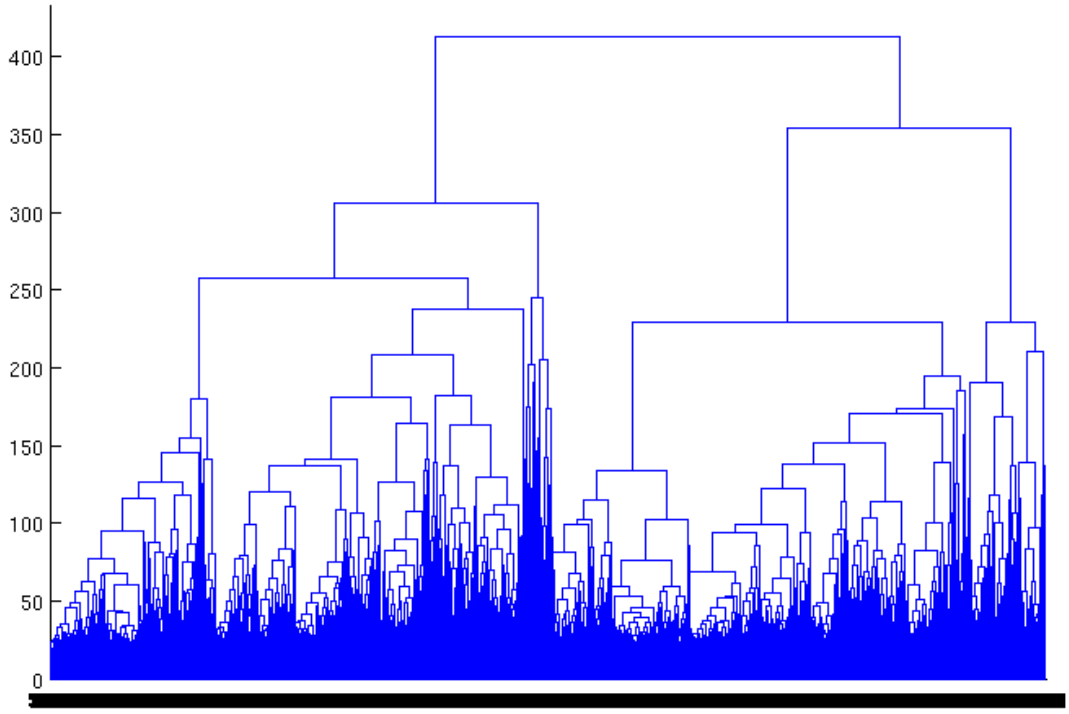
$$Purity(w_i) = \frac{1}{n_i} \max_j (n_{ij}) \quad j \in C$$

When I check purity values, I observe that other than one big cluster, every clusters purity is 1. For example, sixth cluster has 49 elements and all of their natural class is 3. It means, hierarchical clustering procedure works fairly good. After that, I calculate entropy of clusters.

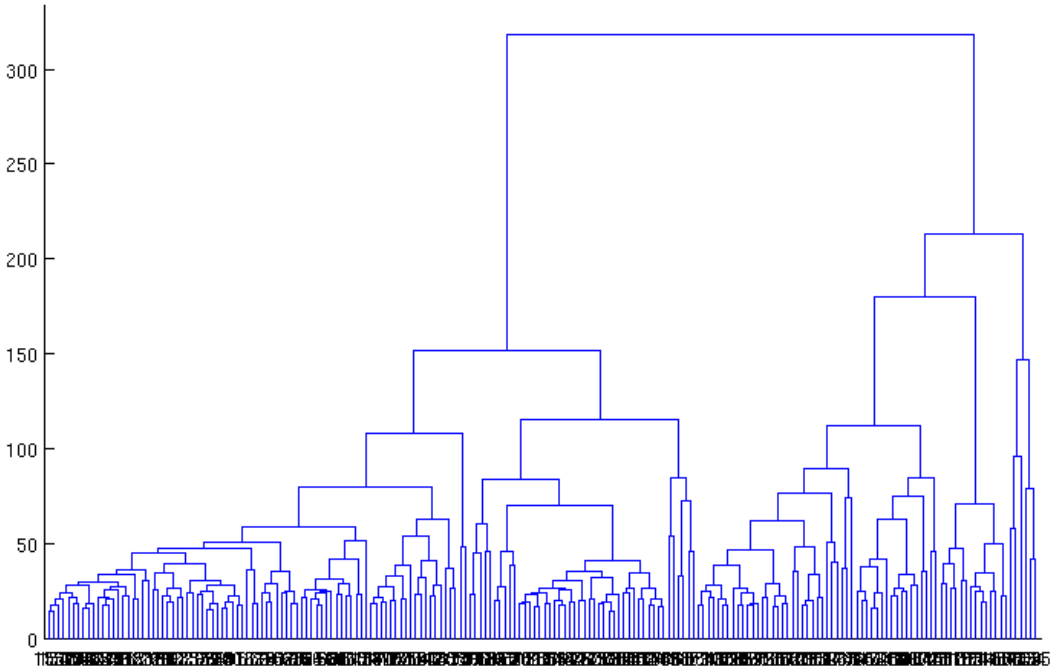
$$Entropy(CC_i) = \sum_{TC_j \in TC} -p(TC_j \mid CC_i) \log p(TC_j \mid CC_i)$$

I find entropy = 0.3150

Then, I draw two dendrograms, one with the all dataset and one with the data which I use in hierarchical clustering part.



Dentrogram with all dataset



Dendrogram with limited data

To sum up, I learn how to implement k_means. I learn that initial cluster choice can cause different clusters. I learn how to implement hierarchical clustering and validate it's results whether it is working good or not.