Name: Prathamesh Nitin Lonkar

B-Number:B-00811727

Email:plonkar1@binghamton.edu

Assignment 2

Q1.)Clean up the data set. This includes filling up the missing values and normalizing all the data items. Please state clearly the methods you use for filling up the missing values and normalizing the values in English to answer this question?

Ans.) To clean up the dataset by replacing the missing values i have used the cummulative mean of all the values in each row and replaced them with the mean in each row. For this i have created a function transform in which i am passing each row the original dataframe and the index of the row as input, then after calculating the mean i am replacing every ’?’ with the mean of that row using the inbuilt function .replace.

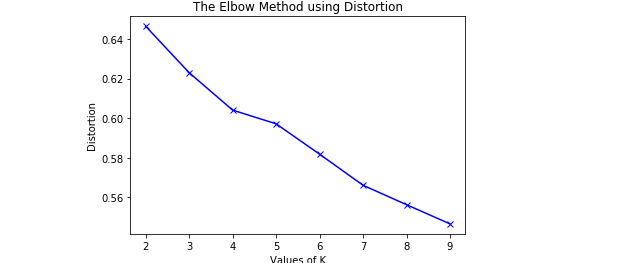
For normalizing the dataset i am using the method specified by Prof i.e. the min max normalization method by using the formulae df3=(df4-df4.min())/(df4.max()-df4.min()) where df4 is the datasets to be normalized, before this i am dropping the first attribute of the table as it is not important for us.

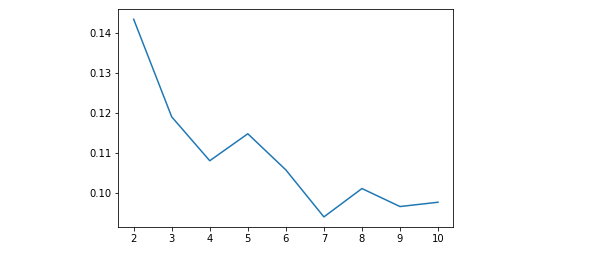
Before doing the normalization i had to convert every value in the dataset to float or int as some of the values were been taken as string.

Q2.) It is well-known that the k-means algorithm requires that the number of clusters, k, be given in advance. In this problem, we do not know the k value in advance. Propose a specific termination condition for the modified k-means when searching the true k value. State clearly your proposed condition or method in English.

Ans.) To calculate the number of k in k-means there are various methods by which we can find the k. Here I have used two methods Elbow method and silihoutte method, by cross referencing both methods I came to the conclusion that the true value of k is 3. The termination condition for both of these conditions were to find the score for maximum of 10 clusters.

After calculating the scores by both methods we can plot these scores on a graph and check the score for each cluster and decide on what should be the true k value. For eac method there are different consideration, for the elbow method we should take the middle value after which the score changes as shown in the figure below, for silihoutte method we will take the cluster with the maximum silihooutte score.



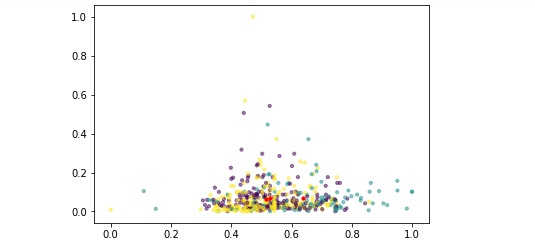


Silihoutte Score

Q3.) Implement the modified k-means algorithm with your proposed termination condition and run the algorithm using the water-treatment dataset. Please note that you must use the output format given in the description file. Report your output.

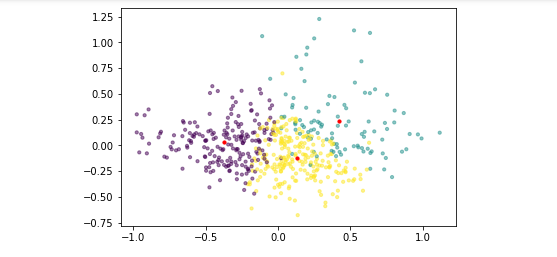
Ans.)After implementing the modified k means we will get the output as descripted in the output file provided in the folder.

In this output file each index will represent each row of the data and its value will represent the cluster number to which that datapoint belongs to.



Q4.)Apply the PCA method you implemented in the first assignment to this dataset. Then apply the implemented modified k-means method above to this reduced data set to report the output. Please follow the same protocol of the output format specified in the description file.

Ans.)After implimenting the pca on a data and then again doing k-means on the new data the result will be same as in the above question as provided in the output file.



Q5.)Compare the two clustering results and analyze any differences that you have observed and state why there is such difference if there is or why there is no difference if there is no.

Ans.) Before the Pca was applied to the dataset the cluster which were formed were compact and been overlaped over each other, and the cluster datapoints were not distinct. As shown in the figure below the datapoints are not distinct as in which cluster they belomg to, and after the pca was applied, the datapoints for each cluster were were very distict as they seemed to be classified correctly.

Q6.) (50 pts.) Implement an autoencoder (either shallow or deep) for dimensionality reduction and apply the implemented autoencoder to the given dataset. Report the dimensionality reduction result using the autoencoder and discuss the difference between PCA and autoencoder for dimensionality reduction with this dataset.

Ans.) After the encoding and decoding layers of the auto encoder are been set with the proper weights and values we will provide the dataset as inout to this autoencoder, what will happen is that it will compute each of this dataset’s value with each encode and pass it on to each decode layer the difference between the pca method and auto encoder method is that the values computed after the dataset was provided will keep on changing after each itteration, i.e every time we give our data to the autoencoder it will give us different results hence we can see different result if we plot a fixed number of columns or rown over a plot.