Assignment 2

Machine Learning (**CS564**)

Department of CSE, IIT Patna

Date:- 26-Aug-2019 **Deadline:-** 02/09/2019

Instructions:

1. Markings will be based on the correctness and soundness of the outputs. Marks will be deducted in case of plagiarism.
2. Proper indentation and appropriate comments (if necessary) are mandatory.
3. You should zip all the required files and name the zip file as ***roll_no_of_all_group_members*.zip**, eg. **1501cs11_1201cs03_1621cs05.zip.**
4. Upload your assignment (**the zip file**) in the following link: https://www.dropbox.com/request/UOritRo8BLNMb0xoqzXk

For any queries regarding this assignment contact:

Pratik Dutta(9007216781/9431432366)

Questions:

1. In the computational biology field, identifying the correct set of clusters helps to analyze the characteristics of the genes that helps to understand the unknown gene characteristics. Also, finding the correct cluster centers is one of the important steps for getting good results. In this assignment, you need to find the correct clusters for the gene expression dataset.

   Dataset:- The dataset is related to B-chronic lymphocytic leukemia(BCLL) dataset which you can download from preprocessed_BCLL.txt . The general structure of the dataset is as follows

<Gene_ID> <Gene_name> <feature_1><feature_2> ... <feature_21>

a. Apply ***k-means*** algorithm on the dataset. Run the k-means algorithm 10 times where each time the number of clusters is randomly chosen between 2 to sqrt(N), where N represents the number of datapoints. For each iteration, also report the ***Silhouette score*** to understand which is the best number of clusters. Show the answer in the following format

   ##

   Iteration 1:- No of clusters = ##, Silhouette score= #####

   . . .

   . . .

   . . .

   Iteration 10:- No of clusters = ##, Silhouette score= #####

Please note, try to avoid the same number of clusters. Also plot a graph, where we can understand the change of silhouette score along with cluster centers.

b. From the previous question, we understand the best number of cluster centers. So for that particular iteration, show the gene names for each cluster. e.g Suppose the best possible cluster center is 5, then show store the gene names in 5 different files where each file represents each cluster.