# Assignment 2

## Foundations of Machine Learning (**CS564**)

## Department of CSE, IIT Patna

**Date:**- 28-Aug-2018
**Marks:-** 25 Marks

### Instructions:

1. All the assignments should be completed and uploaded by **7th September-2018, 5pm.**
2. Markings will be based on the correctness and soundness of the outputs. Marks will be ***deducted in case of plagiarism***.
3. Code should be done in ***Python*** or ***R***.
4. You should zip all the required files and name the zip file as ***roll_no*.zip**, eg. **1501cs11.zip.**
5. Upload your assignment (**the zip file**) in the following link:
 https://www.dropbox.com/request/sPldNavubkUTaFhQzaPY

---

- This assignment is related to clustering algorithm.

---

**Dataset Description:**
***Gene Expression Data Set***
Dataset link:
https://drive.google.com/file/d/1FMlg8OpfHEFwWiso0sdLrmv0IQjK1wcl/view?usp=sharing
Number of instances (gens) :- 4,655 (*N*)
Number of samples/features for each gene:- 21 (GSM45021, GSM45022, GSM45023, GSM45024, GSM45025, GSM45066, GSM45067, GSM45068, GSM45069, GSM45070, GSM45071, GSM45072, GSM45073, GSM45074, GSM45075, GSM45076, GSM45077, GSM45078, GSM45079, GSM45080, GSM45081)

Implement k-mean clustering step by step (**do not use in build packages**). In the clustering algorithm, the value of K (number of cluster centers) varies from 2 to $\sqrt{N}$. Finding a suitable *K* for any clustering algorithm is very much important. In this assignment, you run your implemented k-means clustering algorithm for 10 times where each time the value of K is randomly selected from 2 to $\sqrt{N}$ .

For each iteration, save your output in a folder named "***cluster_number_K***" and inside the folder create K text file (***cluster0.txt, cluster1.txt, …, clusterK.txt***) where each text file contains gene IDs(column 1 of the dataset) of particular cluster.

**Example:-** Suppose, the value of ***K*** in one iteration is 14. So create a folder(***using code***) named "***cluster_number_14***" and generate total 14 text file (***using code***) named ***cluster0.txt, cluster1.txt, …, cluster14.txt.*** Each text file contains the set of gene IDs of a particular cluster.