Assignment 2

CSL7620: Machine Learning
AY 2022-23, Semester – I

Due on: October 11, 2022

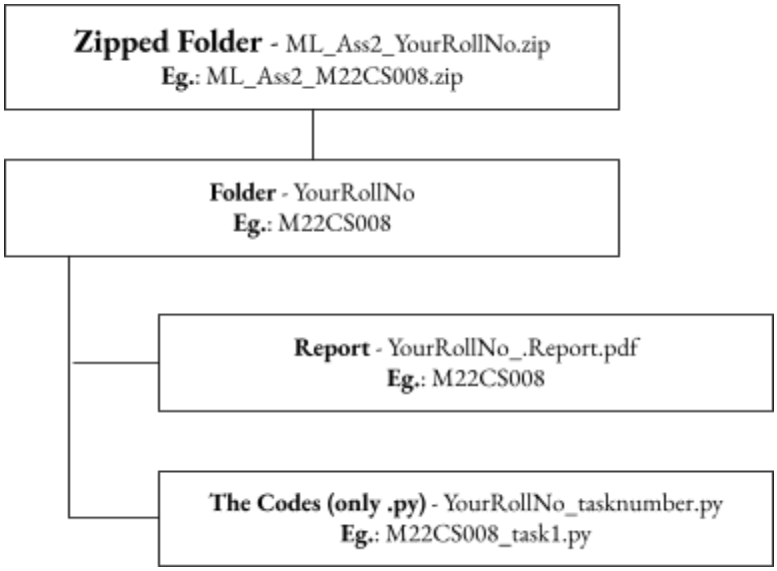
M.M: 200

General Instructions:

1. Clearly, mention the assumptions you have made, if any.
2. Clearly, report any resources you have used while attempting the assignment.
3. Any submission received in another format or after the deadline will not be evaluated.
4. Make sure to add references to the resources that you have used while attempting the assignment.
5. Plagiarism of any kind will not be tolerated and will result in zero marks.
6. Select your dataset correctly. If found otherwise, your assignment will not be evaluated.

Submission Guidelines:

1. Prepare separate Python code files for each task and name them YourRollNo_task1.py, YourRollNo_task2.py, and so on.
2. Also, provide your colab file link in the report. Make sure that the file is sharable.
3. Put both the codes and a report in a folder named <YourRollNo>, create a zip folder named ML_Ass2_YourRollNo.zip, and upload to google-classroom. See attached image to get better clarity.
4. Do not download the .ipynb file, rename it as .py, and upload it. .ipynb files are not exactly in a readable form, so uploading it will only result in you receiving 0 marks for the same. You have an option to download .py file in google colab. Use it to get the .py format.
5. Submit a single report depicting all tasks' methods, results, and observations. There is no need to add theory behind the concepts. Preparing a report is mandatory; failing it will lead to non-evaluation of the assignment.
6. Do not copy-paste code or screenshots, etc. in the report. The report should look like a technical document, containing plots, tables, etc. whenever necessary.



Task 1: Unsupervised learning [100 marks]

Select your dataset [10 marks]

You are given a vector G = [v1 v2 v3 v4], where v1, v2, v3 and v4 can take either of 0|1 values.
Let your IITJ roll number be (M|D|P) *yy* (Branch Code) *abc*.

The values v1 to v4 are as calculated:
v1 = 1 if yy>= 20
v2 v3 v4 is the binary equivalent of the mod operation of *abc* with 4, where v2 is the MSB and v4 is the LSB.

Let Count represent the count of 1s in the vector G, as calculated from your IITJ roll number.
The value of Count represents your dataset number.

Count = 1: 1. Link to your dataset
Count = 2: 2. Link to your dataset
Count = 3: 3. Link to your dataset
Count = 4: 4. Link to your dataset

(a) On the dataset above, apply the K-means clustering algorithm to perform the task of clustering. Vary the values of K and report your observations. The implementation of K-means should be from scratch. **[30 marks]**
(b) On the same dataset, find the gaussian clusters using GMM. The parameters of GMM are to be estimated using EM (Expectation-Maximization) Algorithm, which has to be coded from scratch. **[50 marks]**
(c) When K-means is working well to cluster your dataset, why do you think you are using the GMM algorithm? **[10 marks]**

Task 2: Using Dimensionality Reduction Technique[90 marks]

Click here to download the dataset.

(a) Use Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. Inbuilt libraries can be used for the same. **[20 marks]**
(b) Apply K-means clustering and GMM on the set of uncorrelated features so obtained in (a). **[50 marks]**
(c) Plot the cluster results so obtained in (b). **[20 marks]**

References:

1. **Reference for GMMs** (Opens in Incognito if not a tds user)
2. **PCA documentation**

Task 3: Report [10 marks]