**Instructions Lab 1.1**

In Machine Learning we have to deal with raw data and make some sense out of it. Most often, we need to input data (in various formats), visualize it and then try to model it. Quite often, the data is available in repositories like UCI Machine Learning Repository, Kaggle etc. In this lab you have to become familiar with the process of reading in the data which is present in .csv (comma separated) format. A very convenient way to read in the data is using the Pandas library in Python. You can do some basic visualization using the plotting libraries (like matplotlib).

1. Download the data from

<https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/>

It is data regarding breast cancer diagnostics.

2. First look at the data *very closely*. There are 11 columns. The first column is just an index and does not carry any information. The next 9 columns are the data that we want to visualize. The last column is the data (actually a class label) that we want to predict based on the data of the earlier 9 columns. You will learn to do the prediction (classification) later in this course. Notice that the last column has only two possible values – either 2 or 4. The data in columns 2 – 10 are called feature values and the last column is called a class label.

3. Input the data using the Pandas library.

***Caution: The file has a .data extension and not a .csv extension. It is actually a text file. Use the appropriate options of Pandas library. Alternately, you can save the file in .csv format in your current working directory.***

4. Plot histograms of each feature.

***The data points with class label 2 should have a different color from the data points with class label 4.***

A possible approach is that you can collect all data points with class label 2, build and plot the histogram with values in column 2, using green color. Similarly, collect all data points with class label 4, build and plot the histogram with values in column 2, using red color. Repeat the above for all column from 3 – 10.

You might find the following code snippet useful. It is reading from a different file (**'Placement\_Data\_Full\_Class.csv')** and builds scatter plots. The head( ) and tail( ) methods give you the first few and last few rows. The plot( ) method with the “scatter” argument produces scatter plots. The other arguments, “hsc\_p” and “ssc\_p” refer to specific features (i.e. columns) in the input data file. Look up matplotlib and see how you can create histograms from each column of the data and then complete the assignment.

**import pandas as pd**

**import matplotlib.pyplot as plt**

**df = pd.read\_csv('Placement\_Data\_Full\_Class.csv')**

**print(df.head(10))**

**print(df.tail(10))**

**df.plot(kind = 'scatter', x = 'ssc\_p', y = 'hsc\_p')**

**plt.show()**

**print(df[df.status == "Not Placed"])**

**df\_unplaced = df[df.status == "Not Placed"]**

**df\_placed = df[df.status == "Placed"]**

**print(df\_unplaced.head(10))**

**df\_unplaced.plot(kind = 'scatter', x = 'ssc\_p', y = 'hsc\_p')**

**df\_placed.plot(kind = 'scatter', x = 'ssc\_p', y = 'hsc\_p')**