**BIO542: Machine Learning for Biomedical Applications Monsoon 2018**

**Assignment 2**

**Plagiarism:** All submitted codes are expected to be the result of your individual effort. You should never misrepresent someone else’s work as your own. In case any plagiarism is detected you will get one grade reduction in the final examination. Cite the resource wherever using other’s work.

**Instructions:**

1) Allowed programming language is Python.Name of the input and output files should be given as argument and proper documentation is required inside the code.

2) You must submit your working solution on the backpack on the deadlines page. No extensions on the deadline. If you fail to submit within the time limit, then your solution will not be evaluated.

3) Mention your roll number and name at starting of each file. Write a clean code with proper comments at appropriate places as it will be checked.

4) Store each problem with rollno\_questionX.py where X is the question no. and upload a zip folder with rollno\_name\_assignmentX.zip containing all the codes. Codes won’t be checked if they don't follow the guidelines.

5) You all need to submit only zipped files. RAR files or file with any other extension will not be evaluated.

**Link for the dataset:**

You all can download the data from the given link

<https://drive.google.com/drive/folders/1-kbW_ent3tmCHslGZYtaVsfgqOCOgq99?ogsrc=32>

At the provided link, there is a folder named zip\_data\_mlba, which contains two zip files, one is expression.zip and another is peptide.zip.

Unzip both the files, expression.zip and peptide.zip. Post unzipping, you will get two folders named expression and another peptide. Both folders contain 55 different folders.

There is a folder for each student with their name and Roll no mentioned clearly.

Download your respective folder.

**(A) Questions on Expression Dataset**

Expression folder contains two files; one for positive samples and one for negative samples file. Each file contains the expression data of the genes corresponding to each sample, where rows represents the sample and columns represents the gene expression value. These values are different for each student.

For a given dataset you need to answer the following questions:

**Question 1.** Write a python code for splitting the given datasets into internal (training) and external (test) dataset in the ratio of 80:20 respectively. Write two files; one for training contain 80% data (80% of each positive and negative samples) and one file for testing contain 20% data (20% of each positive and negative samples).

**Question 2.**Write a python code for developing a method for predicting positive samples using SVM and ANN techniques. You may use python library Scikit(<http://scikit-learn.org/stable/>) to develop prediction models. Train your model on training dataset (created in Question 1) and predict on testing dataset. Compute performance of your best models on testing dataset in term of Accuracy, MCC, Specificity, Sensitivity and AUC-ROC Value.

**(B) Questions on peptide dataset**:

Peptide folder contains one positive and one negative file. Each file contains the peptides in the single amino acid residue format.

For a given dataset you need to answer the following questions:

**Question 1:** Write a python code for calculating number of sequence in each file (file contain positive and file contain negative samples). Percent amino acid composition of all peptides in each file. Write results in a file.

**Question 2:** Write a python code for computing percent amino acid composition of each peptide in both files.Write results in a file.