CSE-XXXX Introduction to Machine Learning Lab
4th Year 1st Semester, 2024 (Held in March, 2025)
Department of Computer Science and Engineering, University of Dhaka
Course Instruction: Dr. Muhammad Ibrahim, Associate Professor

Lab 1 (10-4-2025)

In this lab, you shall implement the linear regression algorithm.

Linear regression with one variable

With synthetic data
Generate 100 samples for the function y = 3 + 5x with standard Gaussian noise. The values of feature x will be from 1 up to 100. Plot the data points. Write the data in a csv file named `lab01_data.csv` for future use. Divide the dataset in 80%-20% for training and validation sets. Plot training set and validation set separately. Do add a dummy feature (x_0) with all 1's as per the standard practice of the machine learning community, and then use dot product with this feature vector when needed. Vectorize your code as much as possible as per the standard practice of the machine learning community.

Implement linear regression with gradient descent. Plot the training error curve and validation error curve in the same plot. Print the final training error and validation error, and the best validation error and corresponding training error. Also, print the values of the parameters of the learnt model. Do feature scaling if necessary, but before that, tweak the values of the learning rate to see if desired performance is achieved without feature scaling. Plot the entire dataset with the learned regression line.

Your code must have the following functions:

- `load_data()`
- `process_data()`
- `train()`
- `evaluate()`
  These comprise the basic modules of a machine learning system. In addition, your code must have the following two functions:
  - `gradient_descent()`
  - `compute_cost()`

You can have other functions as well based on your coding practice.

IMPORTANT: You will be evaluated on the cleanliness and modularity of your code. This is the code that you'll use throughout the entire course, so make it as standard as possible.

With real data
Repeat the entire process for the supplied data file `data_01.csv`.