CSE-4155 Introduction to Machine Learning Lab
4th Year 1st Semester, 2024 (Held in March, 2025)
Department of Computer Science and Engineering, University of Dhaka
Course Instructors: Dr. Muhammad Ibrahim, Associate Professor & Dr. Mosarrat Jahan, Associate Professor

Lab 3 (04-5-2025)

The students are required to bring their codes on the following (for binary classification only):
1) Perceptron: Code from scratch.
2) Logistic Regression: Code from scratch. For a fixed learning rate, implement it with only one *for* loop for the iterations. Use vectorized implementation as much as possible so that you can implement the entire algorithm (of one iteration) in 7-8 lines of code. Use *numpy* as the key.
3) Naive bayes: Here you can use existing libraries instead of coding yourselves such as *GaussianNB()*.
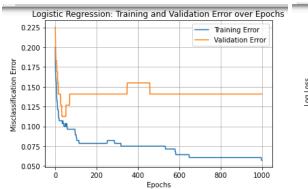
Only single-fold cross validation will be required. As before, you can use built-in functions for generating validation sets.
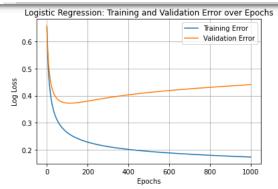Confusion matrix, accuracy, precision, recall and F1 score metrics are required to evaluate the models. You can use built-in functions here.

Print the following error curves:
1) For perceptron, misclassification rate versus the number of iterations.
2) For logistic regression, (a) misclassification rate versus the number of iterations, and (b) log loss versus the number of iterations.
   Example learning curve:

You can use the following codes for generating losses:
```python
def log_loss(y_true, y_pred):
    epsilon = 1e-15  # To avoid log(0)
    y_pred = np.clip(y_pred, epsilon, 1 - epsilon)
    return -np.mean(y_true * np.log(y_pred) + (1 - y_true) * np.log(1 - y_pred))


def misclassification_error(y_true, y_pred_probs, threshold=0.5):
    y_pred_labels = (y_pred_probs >= threshold).astype(int)
    return np.mean(y_pred_labels != y_true)
```

Data: Use the following code for loading the dataset
```python
def load_diabetes_data():
    url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv"
    columns = ["Pregnancies", "Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI", "DiabetesPedigreeFunction", "Age", "Outcome"]
    data = pd.read_csv(url, header=None, names=columns)
    print(data.head())
    return data
```

Regarding Deliverables
1) Do not upload any zip file.
2) Submit the code file(s) only. No data files need to be uploaded.
3) Filenames should have format: *<roll>_<name>.<ext>*.
   Example: *08_Rokonuzzaman.ipynb*

You shall not be allowed to submit your work after the lab hours, so make sure you prepare your deliverables in a timely manner.

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