Assignment-3: Logistic Regression Classifier

Problem Statement:

A floriculture research team X is studying the use of multiple measurements to distinguish three different iris flower species. The dataset contains a set of 150 records under five attributes: sepal length, sepal width, petal length, petal width and species (see Fig. 1). Develop a logistic regression that classifies the species according to the above measurements.

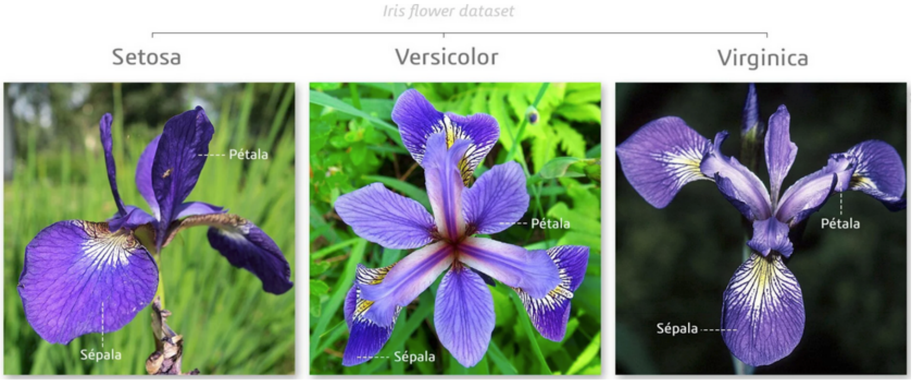

Figure 1: Different iris flower species and their attributes

Implementation: [3+2=5]

- Implementation of gradient descent approach towards logistic regression for multiple classes [LOG_MUL_GRAD] with Minibatch
- Evaluate the model using (a) Accuracy; (b) confusion matrix; (c) precision, recall and f1-score.

****Implement [LOG_MUL_GRAD] from scratch**. You may make use of the numpy library to perform matrix operations.

**In general, you may use libraries to process and handle data.

**For training [LOG_MUL_GRAD], use minibatch size of 30 and total no of epochs 50.

**Perform feature scaling before feeding the data in your model.

Experiments: [3+3+2=8]

The dataset will be split into Train:Validation:Test with 60:20:20 ratio.

1. **Experiment 1:** Report the effect of varying learning rate in [LOG_MUL_GRAD] on validation data. Choose learning rate values from [1e-5, 1e-4, 1e-3, 1e-2, 0.05, 0.1]. Plot Percentage Accuracy vs learning rate. Find the best value of the hyperparameter learning rate.

2. **Experiment 2:** With the optimal parameters found in the earlier experiments, plot the average class probability for each class in the training data after every epoch. Specifically,
    a. Segregate the training data into three different sets according to their true class label.

b. For a particular set, find the mean probabilities for different classes using the updated weights after every epoch. Then, plot the probabilities vs epochs in a single figure.

c. Repeat the previous step (Step b) for all three sets.
(For the 3 class classification problem, there will be 3 plots corresponding to 3 different sets, with each plot having 3 probability curves corresponding to 3 different classes.)

Try to match the experimental observation with the theory.

3. **Experiment 3:**
    a. Analyse the performance of the two models [LOG_MUL_GRAD] with the optimal hyperparameters found in the earlier experiments using the following:
    1. Confusion matrix as a matrix and heat map.
    2. Precision, Recall, and F1-score for individual classes. What characteristics of the dataset will be reflected in this class-specific information (e.g., if data-size for a particular class is relatively less or has a large amount of noise)?

    Report your observations with appropriate explanations.

Datasets:

This dataset comprises three iris species with 50 samples each as well as some properties about each flower. You can find the dataset here.

- ID: Identification number of the flower
- Sepal length: Length of sepal in cm (in real numbers)
- Sepal Width: Width of flower sepal in cm (in real numbers)
- Petal length: Length of flower petal in cm (in real numbers)
- Petal Width: Width of flower petal in cm (in real numbers)
- Species: Three iris flower species (iris-setosa, iris-versicolor, and iris-virginica)

Problem: Predict the species of an iris flower

Submission:

A .zip file containing the python source code and a PDF report file. The final name should follow the template: <Assign-No>_<Your Roll No>.zip. For example, if your roll no is 15CE30021, the filename for Assignment 3 will be: `Assign-3_15ce30021.zip`

1. A single python code (.py) containing the implementations of the models and experiments with comments at function level. The first two lines should contain your name and roll no.

2. A report [PDF] containing **[2 points]**
    a. Experiment 1: Performance values for models with and without feature scaling.
    b. Experiment 2: Percentage Accuracy vs learning rate plot. Also the best choice for the learning rate value.
    c. Experiment 3: Plots of the probabilities vs epochs. For the three class classification problem, there will be 3 plots with each plot having 3 probability curves.
    d. Experiment 4: Tables that present confusion matrix, and precision/recall/f1-score for each class for [LOG_Mul_GRAD] followed by your observations. The template for the tables are shown below.

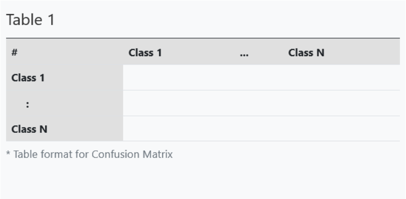

Table 1

#	Class 1	...	Class N
Class 1			
:			
Class N			

* Table format for Confusion Matrix

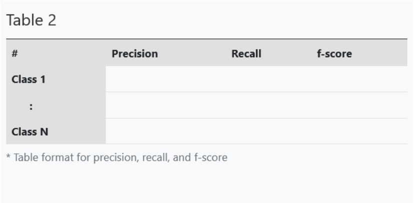

Table 2

#	Precision	Recall	f-score
Class 1			
:			
Class N			

* Table format for precision, recall, and f-score

Responsible TAs:

Please write to the following TAs for any doubt or clarification regarding Assignment 3
Ravi Teja: sista.raviteja@kgpian.iitkgp.ac.in
B Trishaa: b.trishaa@iitkgp.ac.in

Deadline:

The deadline for submission is **2nd SEPTEMBER (Saturday), 11:55 PM, IST**. Irrespective of the time in your device, once submission in moodle is closed, no request for submission post-deadline will be entertained. No email submission will be considered. So, it is suggested that you start submitting the solution at least one hour before the deadline.