COMP3314 Machine Learning

Programming Assignment 1:

Logistic Regression and Random Forest

Start date: March 1, 2021

Due date: 11:59pm, March 26, 2021

Task:

This assignment is about the implementations of multiclass logistic regression AND random forest algorithms. Students are required to follow the lectures, re-implement these two methods, and test them on 2 datasets provided by UCI (url: https://archive.ics.uci.edu/ml/index.php) to verify the validity of implementations.

Datasets:

Students should train and test their implementations on two datasets: Iris dataset [1] and Car evaluation dataset [2]. Each dataset contains multiple classes and has been split into training and testing subsets for evaluation purpose. **Students may download the corresponding .csv files in Moodle to train and test their models**.

[1] Iris dataset

(url: https://archive.ics.uci.edu/ml/datasets/Iris)

The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The 3 plants are Versicolour (class 0), Virginica (class 1) and Setosa (class 2) respectively.

There are 4 input attributes: sepal length, sepal width, petal length and petal width.

There are 100 training samples and 50 testing samples. The training attributes and testing attributes are stored in '**iris_X_train.csv**' and '**iris_X_test.csv**' respectively. And the training and testing labels are stored in '**iris_y_train.csv**' and '**iris_y_test.csv**' respectively.

[2] Car evaluation dataset

(url: https://archive.ics.uci.edu/ml/datasets/Car+Evaluation)

This model evaluates cars according to their status and classifies them as unacceptable (class 0), acceptable (class 1), good (class 2) and very good (class 3).

The 6 input attributes are buying price, price of the maintenance, number of doors, number of persons to carry, size of luggage boot, safety of the car.

There are 1209 training samples and 519 testing samples. The training attributes and testing attributes are stored in '**car_X_train.csv**' and '**car_X_test.csv**' respectively. And the training and testing labels are stored in '**car_y_train.csv**' and '**car_y_test.csv**' respectively.

Guidelines:

[1] Students are required to implement the following algorithms: logistic regression and random forest. Students must implement its core functions, including but not limited to the regularization term in logistic regression and the random bagging strategy in random forest. Moreover, all implementations should be multiclass because each dataset has multiple classes.

[2] It is strongly recommended to use Python as the programming language because we will provide python interfaces in the next assignment. Nonetheless, C/C++/Matlab are also allowed in this assignment.

[3] It may be useful to use third-party numerical packages such as NumPy in python. But it is prohibited to simply call existing machine learning model interfaces of logistic regression and random forest. Also, it is forbidden to copy code from open-source projects.

[4] The submitted source codes must be self-contained. A README file regarding how to run your code should be provided so that we can compile and run your code on our machine.

[5] Students are required to try different parameters in their implementations (such as different learning rates in logistic regression and different numbers of trees in random forest), and examine the final performance under different parameter settings. Discuss their implementations, obtained results, advantages and limitations of the implemented methods in their reports.

Submission Instructions:

[1] One report in **PDF** describing the implementation and results of the model.
[2] Source codes and a README file packed in **ZIP** format that can be unzipped and compiled.