Shahid Bahonar
University of Kerman

Machine Learning Course
Autumn semester
By Dr. Mahdi Eftekhari
Project #2

In this assignment, you're provided with 3 identical datasets, one normal and one with some outliers which include information about height, weight and sex of some random people.

The 3rd one is a Glass Identification dataset from UCI which contains 10 attributes including id. The response is glass type (discrete 7 values). The last column contains label of the glass.

Part A: Linear Regression

Datasets: Data_Normal.txt, Data_With_Outlier.txt

For the first part of the assignment, you need to implement stochastic gradient descent (2000 epochs) and batch gradient descent (2000 epochs) algorithms for linear regression on heights (X) and weights (Y). Please perform the following tasks:

1. Normalize the datasets so that the values of each feature change between 0 (for min. value of the feature) and 1 (for max. value of the feature). Note that this task is very important for the desired results of the upcoming tasks.
2. Train each model separately on the normalized datasets and plot the datasets alongside with the obtained regression model. For these plots, the X axis should be the height feature and the Y axis should be the weight feature. In addition, you should discriminate males and females by using different colors. (4 figures)
3. Report your choice of the parameter "Learning Rate" for each model and explain the effects of changing this parameter.
4. One of these datasets have some outliers. Does it affect the robustness of the model? Explain.
5. Explain what does the normalization process do? When would it be useful to normalize the data?

Part B: Classification

Datasets: Data_Normal.txt, Glass.txt

1. first divide the data into two categories: training data (70%) and test data (30%).
2. For test data, report all classification performance metrics. You are not allowed to use the sklearn library to create the network, but you can use this library to report.

Notes:

- Prepare your full report in PDF format and include the figures and the answer of the asked questions.
- The allowed programming language is Python.
- Assume each row to be a sample in your implementation.
- Your codes should be self-commented.
- Submit your assignment using a zipped file with the name of "Name_Familyname_ StdNum.zip".

You can access the datasets via the link below:
https://drive.google.com/drive/folders/1U3OaSLCJrEsG2pvE8UnATec6AO-PgT1P?usp=sharing