Assignment-1 Problem Statement:

Deadline: Sunday, 4 September 2022, 11:59 PM

Attached are two datasets.

Classification: data_bank_note_authentication.txt

Regression: garments_worker_productivity.csv

Metrics to be used to validate the model:

For classification: F1 score, ROC-AUC score and Accuracy Score

For regression: Mean Square Error

Classification Dataset:

Description:

Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images.

Attributes:

1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (integer)

Regression Dataset:

Description:

The Garment Industry is one of the key examples of the industrial globalization of this modern era. It is a highly labour-intensive industry with lots of manual processes. Satisfying the huge global demand for garment products is mostly dependent on the production and delivery performance of the employees in the garment manufacturing companies. So, it is highly desirable among the decision makers in the garments industry to track, analyse and predict the productivity performance of the working teams in their factories. This dataset is to be used for regression purpose by predicting the productivity range (0-1).

Attributes:

01 date : Date in MM-DD-YYYY
02 day : Day of the Week
03 quarter : A portion of the month. A month was divided into four quarters
04 department : Associated department with the instance
05 team_no : Associated team number with the instance

06 no_of_workers : Number of workers in each team
07 no_of_style_change : Number of changes in the style of a particular product
08 targeted_productivity : Targeted productivity set by the Authority for each team for each day.
09 smv : Standard Minute Value, it is the allocated time for a task
10 wip : Work in progress. Includes the number of unfinished items for products
11 over_time : Represents the amount of overtime by each team in minutes
12 incentive : Represents the amount of financial incentive (in BDT) that enables or motivates a particular course of action.
13 idle_time : The amount of time when the production was interrupted due to several reasons
14 idle_men : The number of workers who were idle due to production interruption
15 actual_productivity : The actual % of productivity that was delivered by the workers. It ranges from 0-1.

Other Important Information:

1. You may have to perform basic preprocessing as taught in the last week.
2. Write the code for the models and training on your own using Numpy/Scipy**. The use of Scikit-Learn or any automatic differentiation package is forbidden**. Note: You can't use numpy.gradient either.
3. You will be expected to explore univariate + multivariate linear regression in closed-form + gradient descent , logistic regression using gradient descent, and Naive Bayes models.
4. Investigate whether selecting a few columns instead of selecting all features yields a better result.
5. For classification, explore if there are any columns due to which Multivariate Gaussian models would be more suitable. (No need to implement Multivariate Gaussian, just mention which columns and how you figured this out).
6. For the purposes of debugging, you may check your implementations on any randomly generated data.
7. Keep in mind that **the aim is not to maximize the training metric, but the metric on test data** that the model has not seen during training. Take appropriate measures for the same.

What you need to submit is a ZIP file with your name as the roll number:

1. A Jupyter Notebook with well-documented code. The code needs to be in working condition without any modifications to be done by the TA to get the results.
2. The datasets are in the same directory so that we can just open-and-run without modifying.
3. A PDF version of the notebook with all cell outputs printed.

Note that different groups have different datasets and will get different results. Do not compare with them -- that will only lead to more anxiety.

Plagiarism from your friends' work or from online will not be tolerated and will invite harsh penalties. Discussion is permitted -- sharing of code is not.

We will be conducting a viva-voce for all students so that we can accurately gauge your sincerity during the assignment as well as your competence.