Prog Asg 1: Regression Model for Structural Strength Prediction

Start Assignment

Due 3 Sep by 23:59 **Points** 200 **Submitting** a file upload
Available 18 Aug at 0:00 - 4 Sep at 23:59

The aim of this assignment is to learn how to apply various regression techniques to real-life problems. You can submit your solution as a Jupyter Notebook with comments and discussions on the results obtained in each step.

Train Data: **https://github.com/gagan-iitb/CS550/blob/main/Labs_M23/MaterialStrength_Train.csv** (https://github.com/gagan-iitb/CS550/blob/main/Labs_M23/MaterialStrength_Train.csv)

Test Data: **https://github.com/gagan-iitb/CS550/blob/main/Labs_M23/MaterialStrength_test.csv** (https://github.com/gagan-iitb/CS550/blob/main/Labs_M23/MaterialStrength_test.csv)

[40 marks] EDA and Feature Engineering

Apply various EDA techniques to visualize, pre-process, and clean the data. Study the correlations amongst attributes, perform feature transformations etc. and prepare your dataset for modeling (machine learning). Prepare your validation set OR cross-validation approach that would be used in the remaining part of the assignment for hyper-parameter tuning and/or model selection.

[30 marks] Exact Solution

Apply the method(s) discussed in class: Normal Equations and Pseudo-inverse to compute the optimal parameters. You may test out various features/combinations in this stage. Please provide the equation for the response variable and its interpretation.

[20 marks] Statistical Analysis

Using OLS library, study the statistical properties of your model(s). Which attributes/features are significant? What are the confidence intervals for each of them? Write your interpretations and comparison of at least 2 good models.

[30 marks] Gradient Descent

Write your own code to perform gradient descent and experiment with the learning rate hyper-parameter. Plot the loss and validation curves. Code your strategy for the convergence criterion.

[20 marks] KNN

Use KNN (non-parametric approach). Experiment with various values of K. What do you observe? How does its validation accuracy compare to the parametric approaches? What are the pros and cons?

[30 marks] Generalized Linear Models

Build at least 2 GLM models with different link functions and distributions. Provide justifications for your choice and interpret your results.

[30 marks] Test Output

Use your validation approach to select the best model and provide the predictions for the test set as a .csv file with only a single column (the output). There should be no header to the file. We would set up a leaderboard and your marks here would be proportional to the accuracy of your predictions.

Submission instructions

a. No zip files are allowed. No Colab files are allowed. Multiple files can be submitted.

b. No copying allowed. A plagiarism check will be performed. This will lead to severe penalties.

c. Naming convention for every file you submit: Roll number_First Name_Asg1_...

d. Provide justifications and interpretations of your results.

e. No late submissions are allowed. Start early and submit often (incrementally, unlimited attempts are allowed).

f. If you are seriously sick (for more than a week) or have an emergency submit a medical certificate and get approval from the instructor., skip the assignment (we will re-scale marks for your other assignments).