Foundations of Data Science (CS F320)

Assignment – 1

Submission Time & Date: 2359hrs on 29th Oct 2021
Max Marks:20

General Instructions:

This assignment is a coding project and is expected to be done in groups. Each group can contain at most three members. Make sure that all members in the group are registered to this course.

This assignment is expected to be done in Python using standard libraries like NumPy, Pandas and Matplotlib. You can use Jupyter Notebook. No other ML library like scikit/sklearn, TensorFlow, Torch etc. should be used.

Refrain from directly copying codes/snippets from other groups or the internet as all codes will be put through a plagiarism check.

All deliverable items (ex. .py files, .ipynb files, reports, images) should be put together in a single .zip file. Rename this file as A1_<id-of-first-member>_<id-of-second-member>_<id-of-third-member> before submission.

Submit the zip file on CMS on or before the aforementioned deadline. Please note that this is a hard deadline and no extensions/exemptions will be given. The demos for this assignment will be held on a later date which shall be conveyed to you.

All group members are expected to be present during the demo.

Problem Statement:

In this assignment, you will be implementing Polynomial regression (with degrees varying from 0 1, 2,. ., 9) using Gradient Descent and Stochastic Gradient Descent methods. But before implementing the algorithms, you are expected to pre-process your data which includes shuffling the data, standardizing/normalizing the values and creating a random 70-30 split to aid in training and testing respectively. Vectorize your algorithms as much as possible to efficiently carry out the computations. Try to print the error value after every 50 iterations during training for better visualization.

a. The dataset consists of two features i.e. ‘Strength’ and ‘Temperature’ applied to a certain piece of plastic. Using the features, you are expected to predict how much ‘Pressure’ that the plastic can stand by constructing matured polynomial features and optimizing the weights by using GD and SGD without any regularization. Do the same for degrees 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Determine the which degree polynomial provides the best fit to the data.

b. Using the same dataset, construct polynomial regression of degree 9 and implement ridge and lasso regression with gradient descent and stochastic gradient algorithms. Build both the models with for different values of lambda (at least 20values) and figure out the optimal models with the most appropriate lambda.

Try to write a clean, modularized and vectorized code which can solve the above problem. Please refrain from hardcoding any part of your code, unless it is absolutely necessary.

What needs to be documented in your report:

Give a brief description of your model, algorithms and how you implemented the regularization.

For part a), tabulate the minimum training and testing error achieved by your model by using polynomials of degree 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 to predict the output. Visualize the surface plots of your predictions (using matplotlib and Axes3D) that you obtained by using polynomials of varying degree and comment on how overfitting actually works.

For part b), tabulate the minimum training and testing error achieved by your model for 5 different values of lambda. Draw a plot of the root-mean square error vs the logarithm of lambda to figure out the optimal model.
(Refer Fig. 1.8 of Pattern Recognition and Machine Learning by Christopher. M. Bishop).

Compare between the models best model obtained in part a) and the best model obtained in part b).

Whom to contact for queries:
Please contact Mr.Achyuta Krishna V (f20180165@hyderabad.bits-pilani.ac.in)for any queries.

Link to the dataset:
https://drive.google.com/file/d/1HDs4i9jRWYo1Ov1xsobrX4vayPT8bw41/view?usp=sharing