CS F320 – Foundations of Data Science

Assignment – 2

Submission deadline: 24-10-2020, 23:59

General Instructions:

- This assignment is a coding project and is expected to 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 by the IC. All group members are expected to be present during the demo.

Problem Statement:

- In this assignment, you will be implementing Linear Regression using all three methods, i.e. a) Solving by normal equations (i.e. finding inverse of design matrix) b) Gradient Descent c) Stochastic Gradient Descent as taught in class. 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. 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.
- The dataset consists of three features i.e. age, bmi and number of children of an individual. Using these features, you are expected to predict the insurance amount for that person. Try to write a clean, modularized and vectorized code which can solve the above problem. Please refrain from hardcoding any part of your code, until unless it is absolutely necessary.
- With 20 random 70:30 splits of the data set into training and testing data, build 20 regression models. Find the mean and variance of prediction accuracies of these models.

What needs to be documented in your report:

- Write a very brief overview of your pre-processing, model and implementation of each algorithm.
- Mention the minimum training and testing error achieved using each algorithm and justify your observations.
- Plot a graph of error vs epochs using various learning rates (ex. 0.1, 0.01, 0.001) for GD and SGD.

Questions to ponder on:

- Do all three methods give the same/similar results? If yes, Why? Which method, out of the three would be most efficient while working with real world data?
- How does normalization/standardization help in working with the data?
- Does increasing the number of training iterations affect the loss? What happens to the loss after a very large number of epochs (say, ~10^{10})

- What primary difference did you find between the plots of Gradient Descent and Stochastic Gradient Descent?
- What would have happened if a very large value (2 or higher) were to be used for learning rate in GD/SGD?
- Would the minima (minimum error achieved) have changed if the bias term (w_0) were not to be used, i.e. the prediction is given as $Y = W^T X$ instead of $Y = W^T X + B$.
- What does the weight vector after training signify? Which feature, according to you, has the maximum influence on the target value (insurance amount)? Which feature has the minimum influence?

Link for the dataset:
https://drive.google.com/file/d/1plkY18tLL4P0DLlUwYB6l1P4YH24_lOQ/view?usp=sharing