Foundations of Data Science (CS F320)

Assignment - 1

Submission Date: TBD in class 14th Oct

Background:

Bob and Lisa are primary school students. They are given a home assignment in which they are given the latitude and longitude values of a few places. Their task is to surf the internet and find the altitude of those places. Unfortunately they have lost the network connection because of bad weather. Your task is to help them find the altitude values through linear regression using the dataset you were given in the attachment.

Task:

You will be developing linear regression models to predict the altitude values using the following three methods on the data set specified in towards end of this document. You are not supposed to use direct python APIs to build regression models but to write python code to build models using numpy, pandas and matplotlib libraries in python.

Part A - Gradient Descent Method:

You will be developing linear regression model by minimizing the loss function i.e., half of the sum of squares error over the train set using gradient descent method. You need to write python code to implement gradient descent method. Choose an appropriate initialization for the weights, learning rate and stopping criteria.

Part B – Stochastic Gradient Descent Method:

You will be developing linear regression model by minimizing the loss function i.e., half of the sum of squares error over the train set using gradient descent method. You need to write python code to implement stochastic gradient descent method. Choose an appropriate initialization for the weights, learning rate and stopping criteria.

Part C - Gradient Descent Method along with regularization:

As an extension to Part A, you will be implementing L1 and L2 regularization. You can use a part of the train data as the validation set. Also, determine the best value for the regularization coefficient.

Part D - Normal Equations Method:

You will be developing linear regression model by the method of solving normal equations as discussed in the class. You will be implementing this in python.

Dataset:

The dataset and its description can be found in the following link
https://archive.ics.uci.edu/ml/datasets/3D+Road+Network+%28North+Jutland%2C+Denmark%29
Drop the first column. The next two columns are the latitude and longitude values and the fourth column is the target attribute.

Report:

- ✓ The report should have all the details of the models that are developed in the assignment.
- ✓ Comparative study of all models that are developed should be recorded with appropriate measures like R^2 and RMSE.
- ✓ For Part C, make plots of validation loss against the regularization coefficient for both L1 and L2 regularization.
- ✓ Plot the loss over the train set for every 20 iterations of the gradient descent algorithm.
- ✓ Comment on the differences between L1 and L2 regularization w.r.t the regularization coefficients.
- ✓ Put all the code in a single file, zip the code and this document and name it with your ID numbers.

For Queries:

Itiyala Sonika (f20160099@hyderabad.bits-pilani.ac.in)