Foundations of Data Science (CS F320)

Assignment - 2

Submission Date: TBD on 30th Oct class
Points: 25

Background:

You have successfully helped Bob and Lisa by building Simple Linear Regression models. Now you want to verify if you can build a better model by implementing Polynomial Regression on the same dataset.

Your task:

You will be developing polynomial regression models to predict the altitude values using the following methods on the data set specified towards the 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 polynomial regression models of degrees 1, 2, 3, 4, 5 and 6 by minimizing the loss function i.e., half of the sum of squares of error over the train set using gradient descent method.
For instance, the model that you will be using for a degree 2 polynomial regression will be

$Y = w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2$

Similar models are to be used for polynomial regression of degree 3 and 4.

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:

Compare the models by computing the sum of squares of error, R^2 and RMSE for each of the four polynomial models that have been developed in PART A.
Hence find the model that best fits the data and the one that overfits.

Part C - Gradient Descent Method along with regularization:

You will be implementing L1 and L2 regularization for the regression model built with degree 6.

Note that a portion of the test set can be used as the validation set.

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:

- ✓ Prepare a document containing your answer for PART B.
- ✓ Comment on the effect of regularization on the loss over the test set.
- ✓ Also comment on the effect of regularization on overfitting.
- ✓ Compare the loss of a model with a degree 6 polynomial with regularization applied to that of a lower degree polynomial without regularization.
- ✓ 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)