Foundation of Data Science

Assignment-1 : Linear Regression

Members :

Garvit Jain- 2016A7PS0080H

Rahul Midha - 2016B3A70923H

Aman Mehta - 2016A7PS0066H

Date : 20/11/2019

Linear Regression

OVERVIEW:

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable. Here we apply linear regression of given longitude and latitude of given places to predict altitude. Data is divided into training and test data. Training data is 70% of the data. Test data is further divided into validation data, which is 5% of the tedata.

GOALS:

1. Apply linear regression of given data for degree 1.
2. Apply stochastic gradient
3. Use L1 and L2 regularization and find appropriate regularization constant for both
4. Normal Equation

Language used to implement : Python3

Corpus used : Latitude, longitude and altitude data for 434874 points

We then split the data into training, testing and validation.

70% training data

30% test data

5% of test data is the validation set which we will use for L1 and L2 regularization.

Packages used :

1. numpy : To perform Matrix multiplications,to slice arrays, find eigen-pairs, pseudoinverses, perform dot-products etc.

2. matplotlib : To plot graphs of given problem

3. pickle : To manage data between different code.

Concepts and formulae:

When selecting the model for the analysis, an important consideration is model fitting. Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R^2). However, overfitting can occur by adding too many variables to the model, which reduces model generalizability.

R2=1-SSE/TSE

Loss Function: ½ Σ((y(x,t)-t)^2)

We try to find differences between different regression techniques. Following table shows the results obtained by using different techniques on our data.

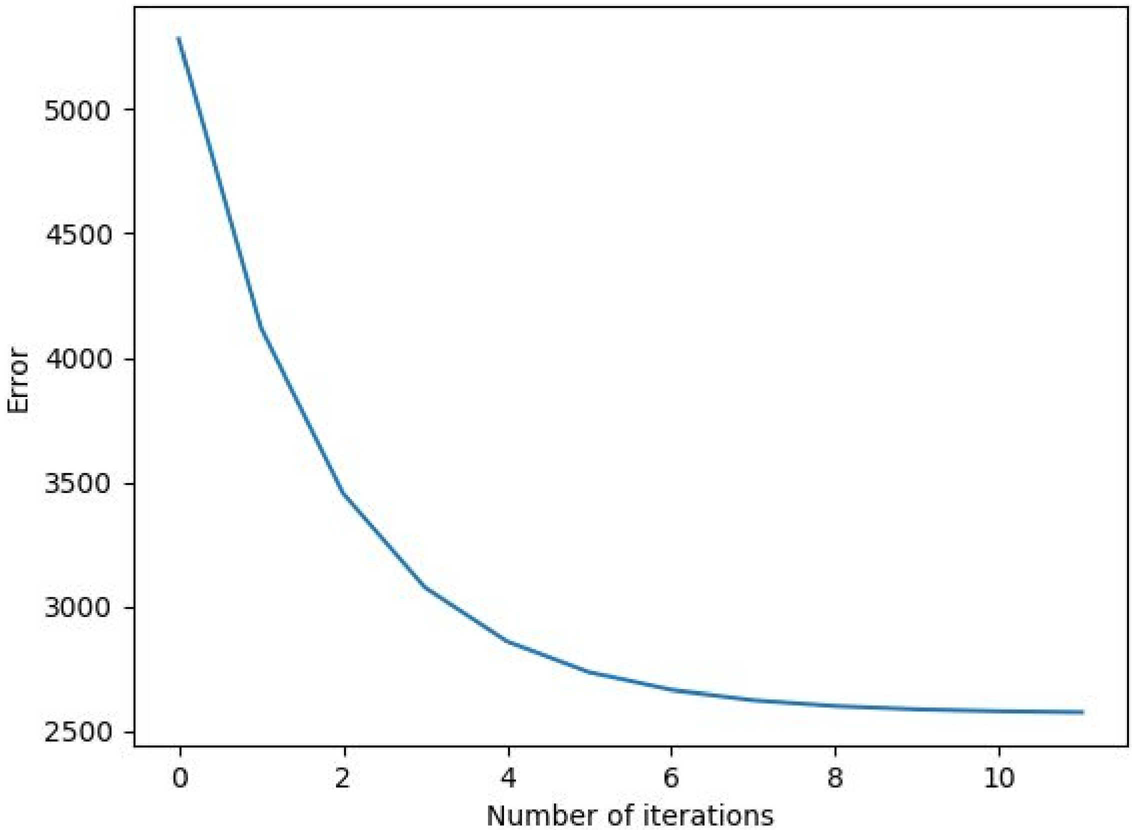
	Root Mean Squared Error(RMSE)	**R2 Value**	**Time Taken (In Seconds)**	**Training Error**
Linear regression of deg 1(normalized)	0.1472	-0.029	2.61	2598.88
Linear regression with l1 reg (lamda = 100)(normalized)	.148	-0.045	3.09	2611.303
Linear regression with l2 reg (lamda = 100)(normalized)	0.147	-0.029	4.296	2479.004
Stocastic (normalized)	0.1405	-0.036	140	3622.49
Normal form	0.024	0.32	0.89	8505.85

Leaning Rate = 8 * 10^-7   Stopping Constant = 0.1   Initial w's =[1,1,1]

+

PLOTS

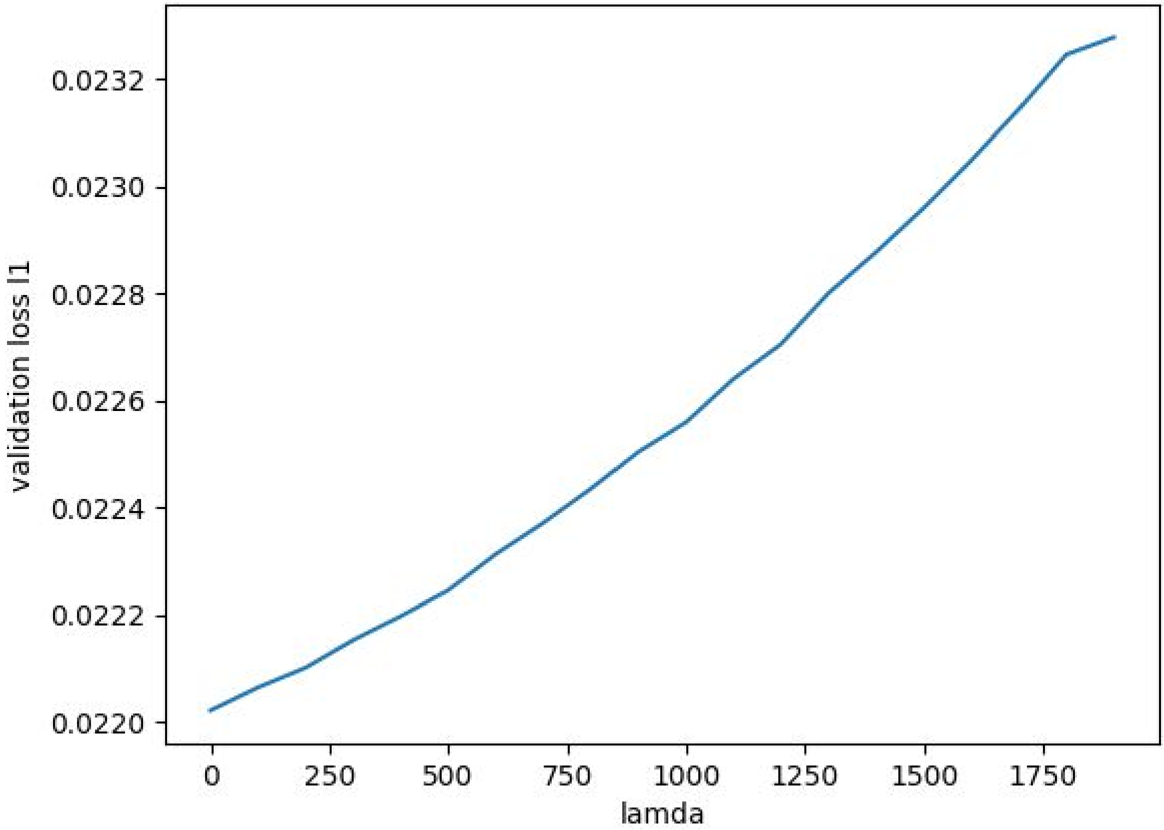
Gradient Descent



Plot of Error vs no. of iterations for gradient descent

We can see the error reducing with iterations as expected.

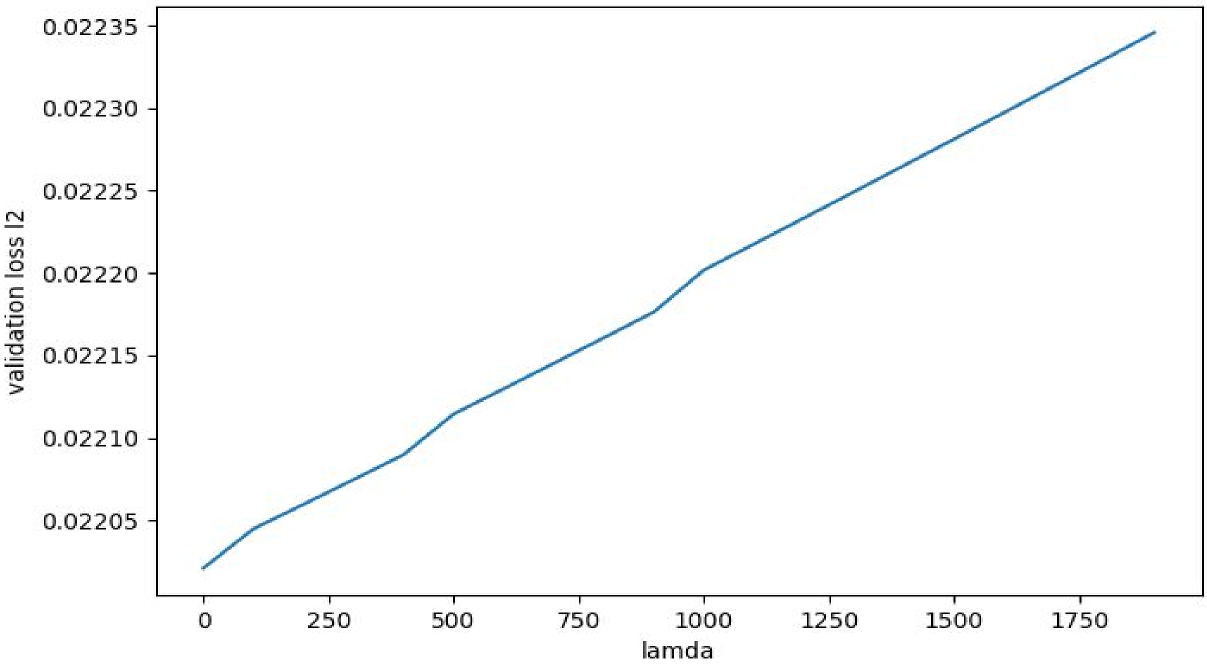
L1 Regularization



Lamda approximation for regression with l1 regularization (Validation Loss vs Iterations)

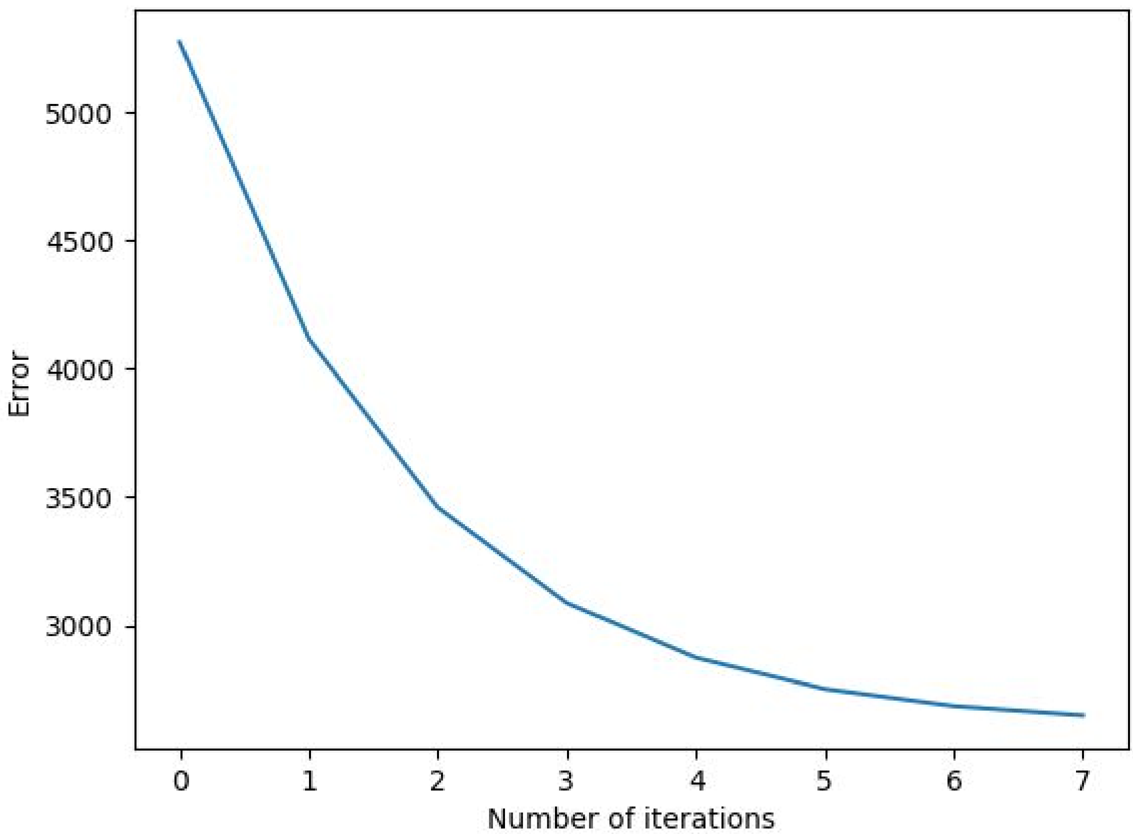
We plot validation loss for different values of lambda. We see that we get the minimum validation loss at lambda = 0, this shows that there is no overfitting in the data.

L2 Regularization

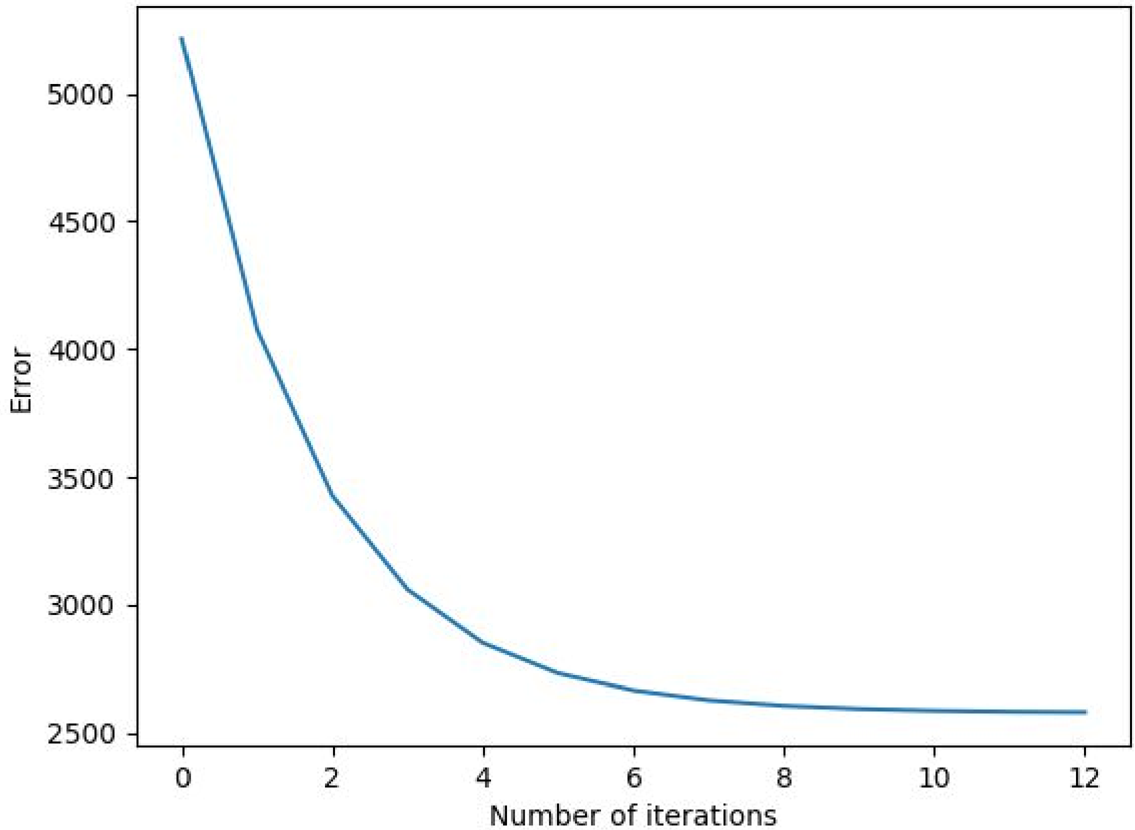


Lamda approximation for regression with l2 regularization (Validation Loss vs Iterations)

We plot validation loss for different values of lambda. We see that we get the minimum validation loss at lambda = 0, this shows that there is no overfitting in the data.



Linear Regression with l1 regularization (lamda = 100)



Linear Regression with l2 regularization (lamda = 100)

The main difference between l1 and l2 regularization:

L2 regulariztion	**L1 regulariztion**
Computationally efficient	Less efficient
Makes weights close to zero but not zero	Makes weights close to zero
No feature selection	Helps in feature selection

Since our model doesn't have many features we can't see the difference between two regularization techniques

Stochastic Gradient Descent