Practical 4
19BCE248
2CS501

AIM:- Linear Regression using Gradient Descent & Normal Equation Method with Regularization (without using sklearn or equivalent library for both)

About Dataset (Boston):
Boston is a pre-built dataset for practising Regression problems.This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass.
Rows: 506
Columns: 13
The target column describes the target variable i.e. cost of house according to parameters.

Preprocessing:
- Firstly dividing the entire dataset into Training and Testing dataset.
- Training rows: 400 Testing Row:106
- Next preprocessing is done on data.
- The StandardScaler method is used to normalize the data.

Equations used:

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n}\theta_j^2\right]$$

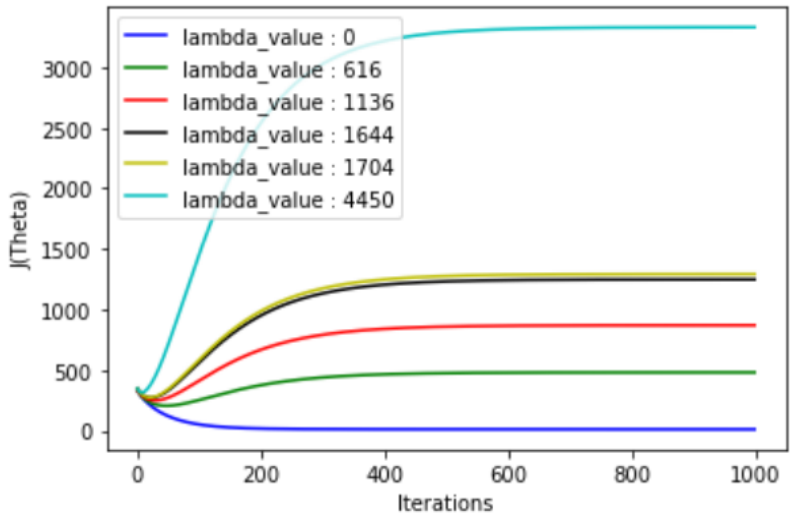
$$\min_\theta J(\theta)$$

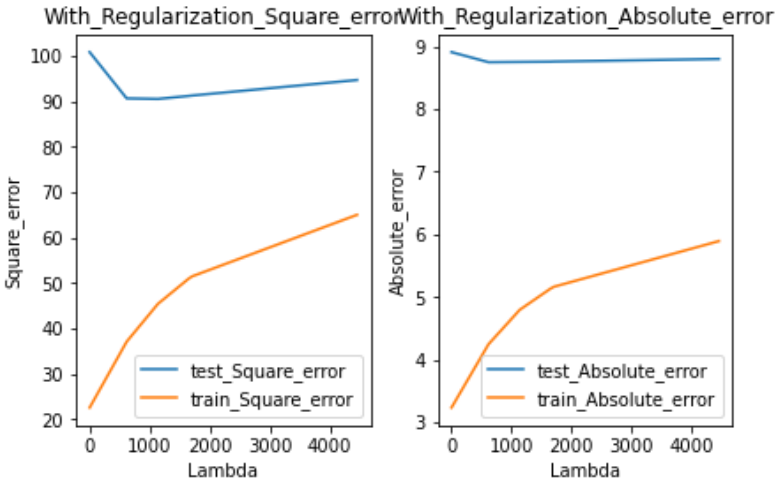
Normal Equation:

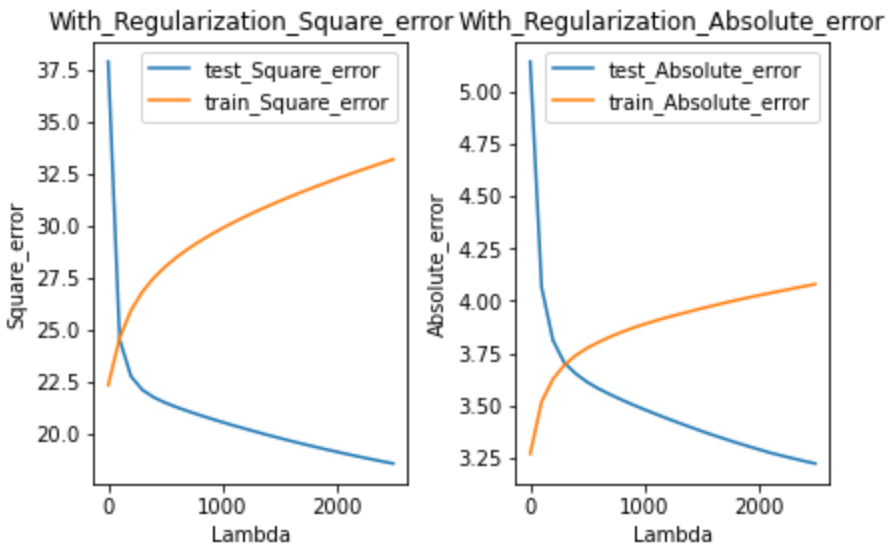
If $\lambda > 0$,

$$\theta = \left(X^T X + \lambda \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & 1 & \\ & & & \ddots \\ & & & & 1 \end{bmatrix} \right)^{-1} X^T y$$

Conclusion:







So here from the above figure we can clearly see that as the value of lambda increases the testing error decreases and training error increases. Such variation is seen because due to overfitting as we train the data it will show us very minimal error but at testing time it drastically changes to bigger values. So knowingly we add regularization to increase the cost function so it will keep diverging and training can be done in an efficient manner.