REPORT

Submitted by:Ishita Mallecherry

Machine Learning Module used: sklearn
Time: 2 weeks
Collaborator: Aparna Jaiswal
Programming language: Python 3

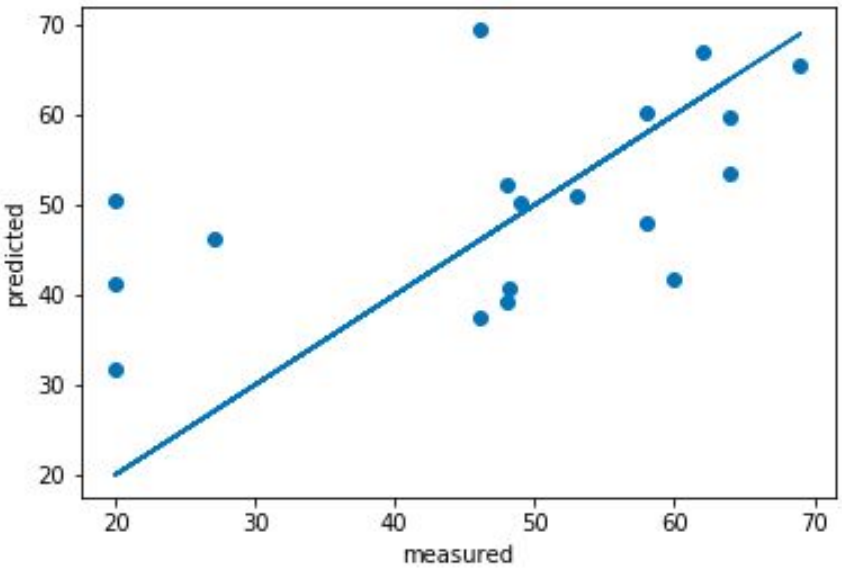
TASK 0

In this assignment Jupyter Notebook has been used and sklearn for the machine learning library

In task one the given data had to be imported to the machine learning environment. This is done by creating a csv file and importing the data.
Then the next task is to split the data into train and test sets. Fours sets of data X_Train,y_train,X_test,y_test is created such that it randomly distributes data into training set containing 85 samples and the test set containing 18 samples.
Next job is to apply cross validation to the train and test sets. We perform a 5-fold cross validation which is iterated 10 times. Every time the performance is tested on the test sets in order to calculate the mean squared error and the cross validation score.
The minimum mse value comes around 107.

TASK 1

1.1: Un-regularized Linear Regression: the task was to perform Linear Regression on the data. The regression model is created and the X_Train and Y_train are fit on the model and then this is tested on y_Test to get the predicted data. Then mean squared error is calculated to be 181.1185086749484. The graph between predicted and measured values:

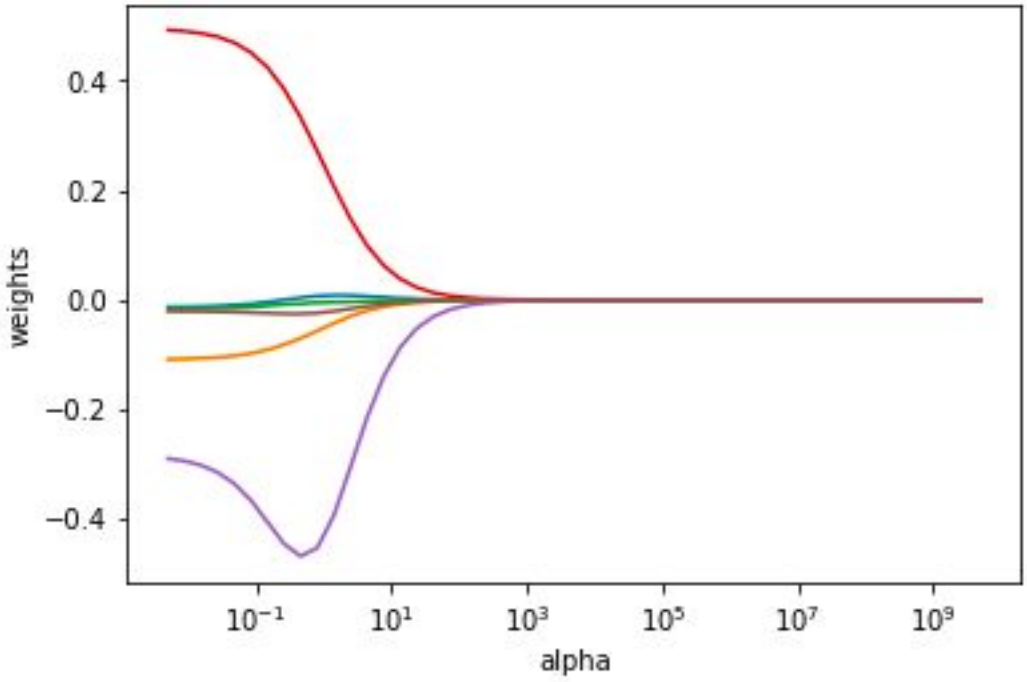


1.2 Ridge Regression: A ridge regression model is created and the job is to estimate the regularization coefficient which produce minimum mean squared error. The RidgeCV classifier is used for the task. An array for value of coefficients and an array for different values of alpha is created to ranging over a broad set of values. These values are used for task 2 as well the alpha value corresponding. The best value of alpha is computed and the mean squared error comes around 184, which is more than the unregularized linear regression.

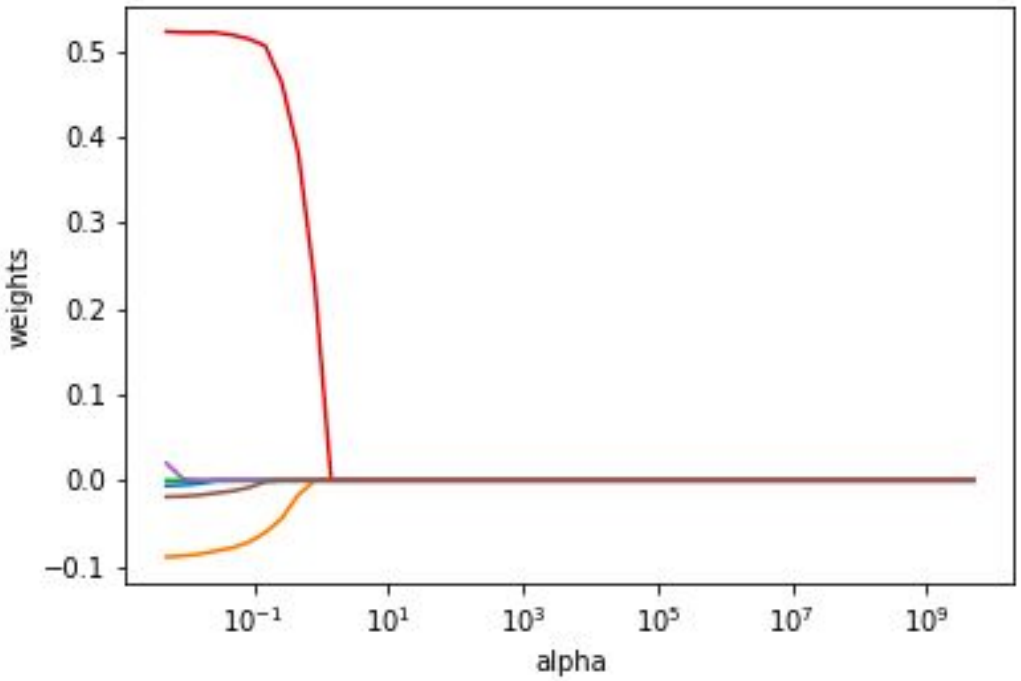
1.2Lasso Regression: A lasso regression model is created and the same steps as ridge regression are used. We get a mean squared error around 177 which is less that in 1.1. So here regularized lasso regression proves to be better than un-regularized linear regression.

TASK 2

In this task the regularization paths for both Lasso and Ridge regression are plotted. The array for alpha created in task one for both Ridge and Lasso regression are used to create a loop and estimate the coefficients. The coefficients referred to as weights and the value of alpha are used to plot the regularization paths.



Ridge Regression



Lasso Regression

THE END