Linear Regression Analysis

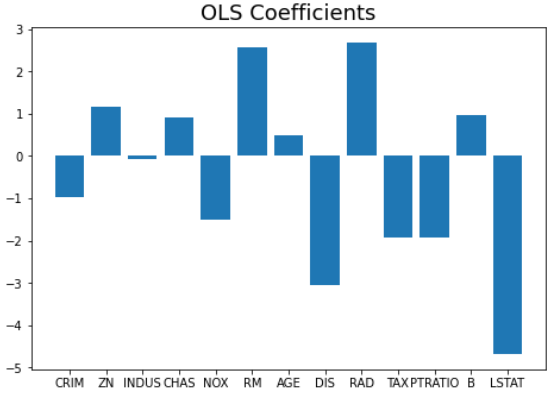
Vishawam Datta

Introduction

The Boston housing dataset is a collection of data about prices of houses in different areas of Boston along with providing various other features of that region (13 different features or attributes to be precise). The features are the following, crim-per capita crime rate by town, zn-proportion of residential land zoned for lots over 25,000 sq. Ft.,indus- proportion of non-retail business acres per town, chas-Charles River dummy variable, nox- nitrogen oxides concentration (parts per 10 million),rm-average number of rooms per dwelling.age-proportion of owner-occupied units built prior to 1940,dis-mean of distance to five boston employment centres,rad-index of accessibility to radial highways,tax-full value property tax rate per\$10,000, ptratio pupil-teacher ratio by town ,black-1000(B-0.63)^2, B is the propotion of blacks by town,istat-power status of population (percent),medv-median value of owner occupied homes in \$10,000.The inferences and the analysis of the observations is mentioned in each section along with the tasks.A brief summary of all the main and important conclusions is provided at the very end of the report(the major learnings).

1. OLS Regressor

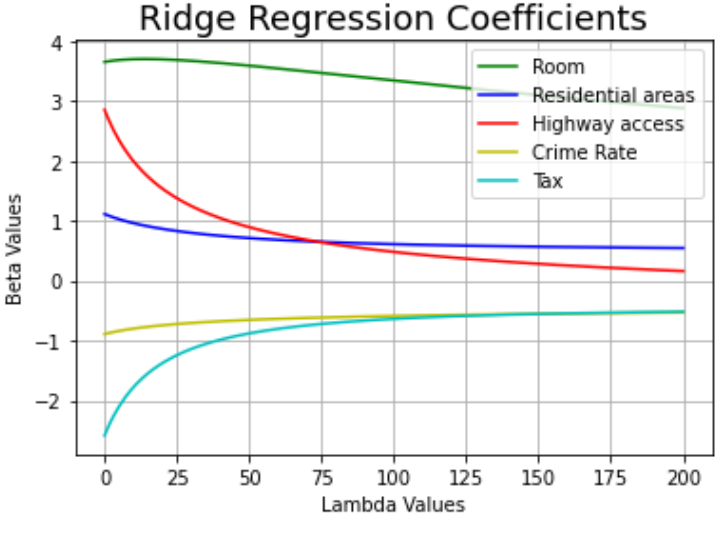
The regression line is obtained by minimising the RMSE (root mean square error) and the method of stochastic gradient descent is employed by the library to minimize the RMSE expression so as to find the intercept value and the coefficient corresponding to different predictor/feature variables. The plot for the regression coefficients obtained from this method is as follows: -



From the above plot we can clearly draw a qualitative as well as quantitative measurement as to how sensitive the prices of houses in Boston are to the different features provided in the dataset.Such as the price is least dependant on features such as AGE and INDUS but most sensitive to changes in features such as DIS and RM. However one must take the above inferences with a pinch of salt because when we use the method of OLS some of the features may start predicting(adjusting themselves to) the error in the training data (so that RMSE can be minimized) instead of predicting the prices of the houses. If this is the case then it becomes a classic case of overfitting which can mainly happen here if certain features which don't really predict the price have high value of coefficients/slopes to minimise the error(since these adjust themselves to the errors instead of the prices).Furthermore such models have a very low bias and seem to perform extremely well on the training data set but when these are evaluated with different testing datasets, such models prove to be quite inconsistent and hence result in high variance. However, we cannot conclusively say whether the OLS model is over fit or not, for this we will have to employ a different regression technique known as ridge regression.

2. Ridge Regression

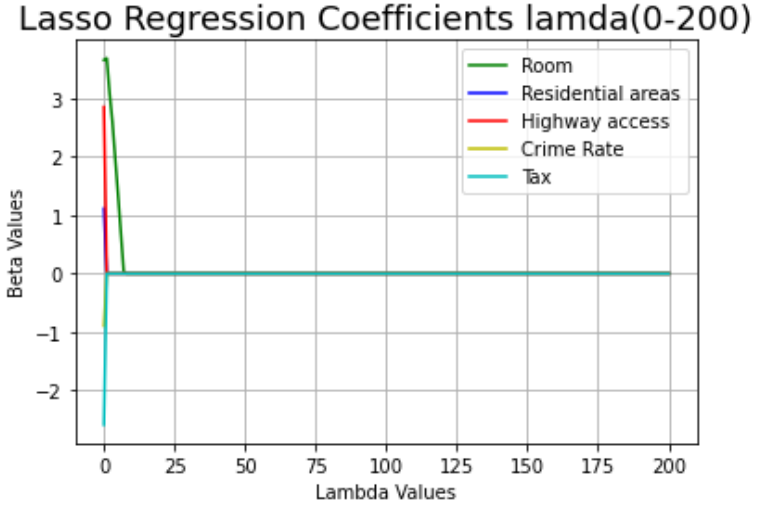
Here the regression line is obtained by minimising a different cost function. The cost function is same as RMSE but a penalty term lambda is added to the square of slope/coefficients of different features and that function is minimized. Note: before fitting the data in the ridge regression model ,the entire data set of features is standardized using the pre processing method from the library scikit learn, this was recommended in the documentation provided by the website and is done so that we can make different feature variables comparable to each other , that is to say that after pre processing all the variables are in a similar range and there will not be a case where one feature has a value of 2 units and the other has a value of 10,000 units, hence this greatly helps in the analysis of the performance of the model used(after visualizing data through various graphs).This pre processing may or may not be done for normal OLS model. The plot of various coefficients varying according to Lambda is as follows.



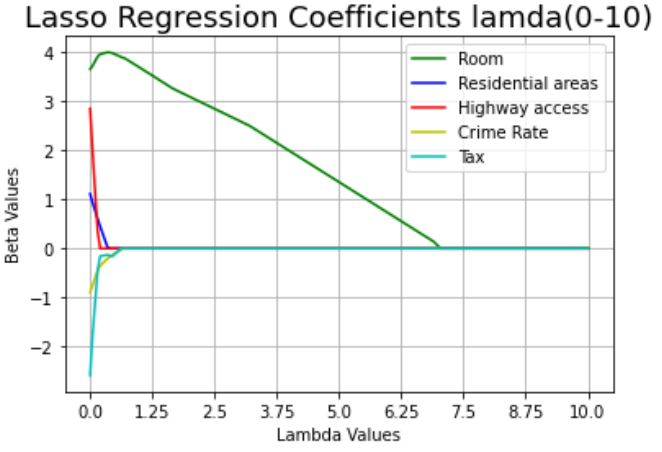
One thing that is clear from the above plot is that as the value of lambda increases the coefficients of different features approach zero but none of them become exactly zero, these coefficients/slopes only tend to zero as lambda tends to infinity. However, one important observation is that the rate of decrease of different coefficients is different and some decrease more aggressively as compared to others. These factors are a strong indication of the relative importance of different features. That is to say that the price of the houses in Boston is more sensitive(dependant on) to those features which don't approach zero very quickly or in mathematical terms the rate of decrease of such coefficients is less(as compared with others). Hence based on the above discussion and inference it can be easily observed that the prices of houses are more sensitive (strongly dependent) on features such as Room. On the other hand, seeing the curve of highway access (the rate of decrease is quite high, it approaches zero very fast) one can infer that the sensitivity of prices to the change of this feature variable is quite low and the prices are less dependent on it(note-the OLS regressor predicted a high value of coefficient for rad(highway) variable).This seems intuitive as well , that is to say that the price of a house should depend strongly on the number of rooms in it and the price shouldn't really fluctuate based on accessibility to highway unless the residents are ones who frequently travel to other areas which is not really common. The above model is surely an improvement on the model of OLS regressor but there is one clear drawback that I observed that it is incapable of reducing the coefficient values to an exact zero ,and to eliminate over fitting completely it is important to do so.Now we turn our attention to lasso regression which hopefully will solve this problem.

3. Lasso Regression

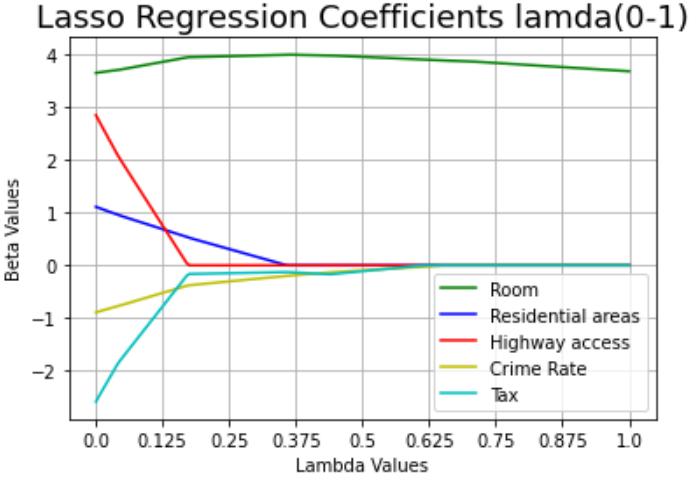
The lasso regressor is quite similar to ridge regressor albeit a subtle difference, here the sum squared coefficients are not penalized but the sum absolute values of the coefficients is penalized by the penalty term lambda. Here also the pre processing step is quite important as it'll allow us to get insights into the data when we plot the graphs. Plotting coefficients as lambda ranges from 0 to 200:-



It is clear that this plot isn't very informative hence zooming in and plotting for further analysis: -



Zooming further: -



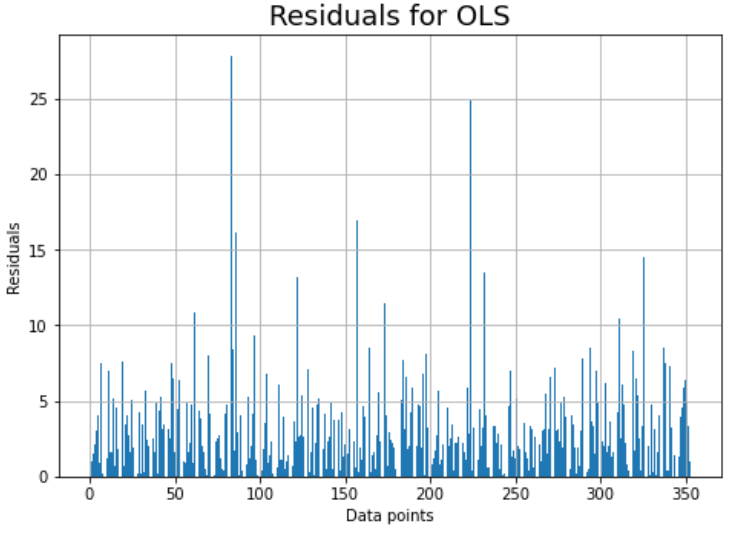
It is clear from the above plots that the value of the coefficients don't only decrease and tend to zero as lambda increase but become exactly equal to zero which is quite good as if we are able to find the correct value of lambda then we will be able to eliminate unimportant features without eliminating the important ones. As seen from the above graphs if the value of lambda crosses 7.5 the coefficient of room becomes zero despite it being an important feature, hence this is highly undesirable.

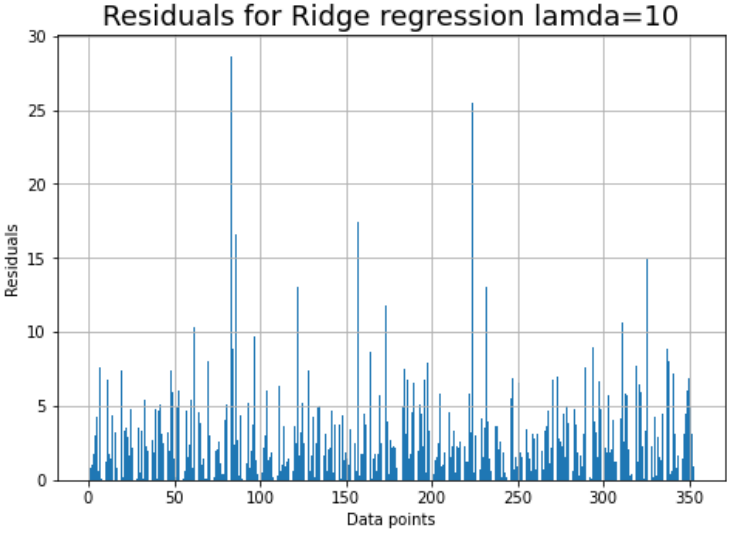
ESTIMATING LAMBDA: -

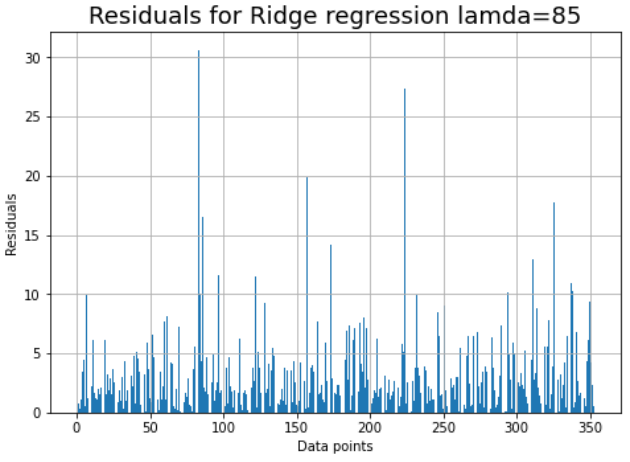
Here generally some random values are taken for lambda and then in cross validation the models which perform well (minimize errors) are chosen and we take that value of lambda.

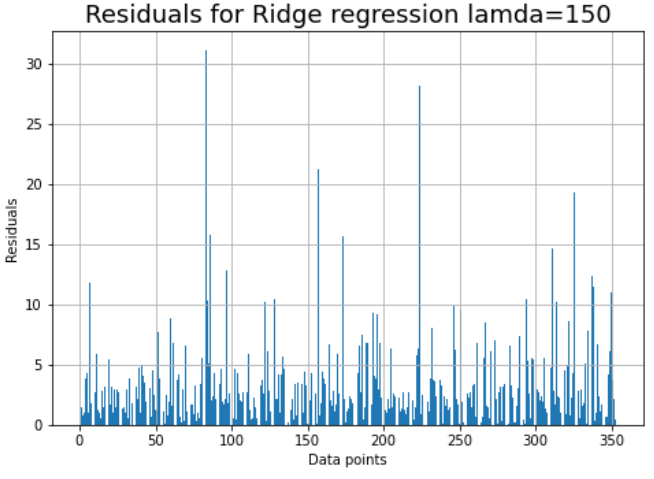
4. Analysing the performance.

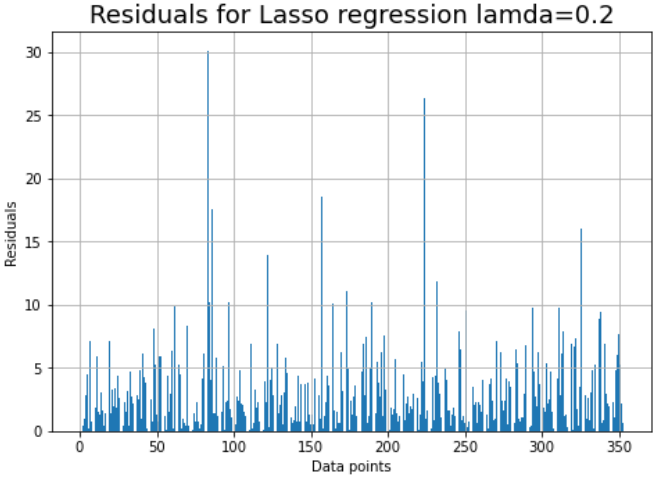
Plotting the residuals: -

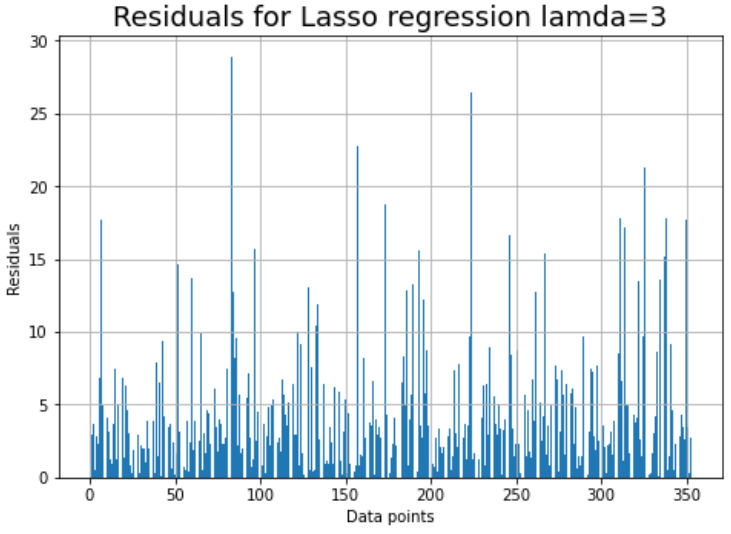


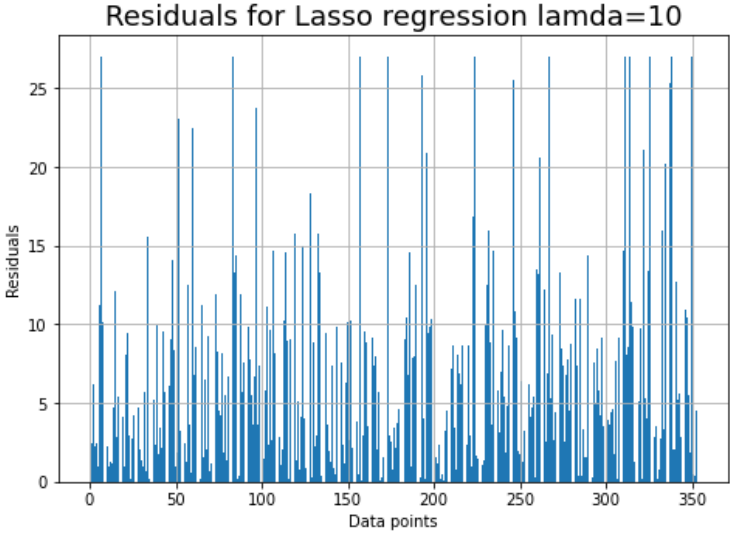



Residuals for Ridge regression lamda=85
Residuals
Data points
0
5
10
15
20
25
30
0
50
100
150
200
250
300
350


Residuals for Ridge regression lamda=150
Residuals
Data points
0
5
10
15
20
25
30
0
50
100
150
200
250
300
350


Residuals for Lasso regression lamda=0.2
Residuals
Data points
0
5
10
15
20
25
30
0
50
100
150
200
250
300
350





TABULATING THE MAEs (The mean absolute errors)

Sno.	The model used	MAE (training data)	MAE (testing data)
1.	OLS Regressor	3.23691	3.5792
2.	Ridge regression(lambda=10)	3.19769	3.50782
3.	Ridge regression(lambda=85)	3.26564	3.31368
4.	Ridge regression(lambda=150)	3.40143	3.26455
5.	Lasso regression(lambda=0.2)	3.24671	3.46759
6.	Lasso regression(lambda=3)	4.58152	3.92086
7.	Lasso regression(lambda=10)	7.01801	6.0541

Inferences from the plots and tabulations of task 4 and task 5:-

Looking at the table and the plots we can come at some really interesting observations. One observation that is extremely useful and interesting from the tabulated data is that the ridge regression and lasso regression increase the bias slightly so as to reduce the variance of the model, that is to say that now the model performs more consistently over various testing data sets. This is evident from entries at row no.(S.nos) 3, 4 and 5.For example looking at row 4 we see that the MAE for training data is ~3.40 and the MAE for training data for row1(OLS) is ~3.24 hence there is a slight difference that is to say that the ridge regression model in this case seems to perform poorly on the very data set that was used to train it ,BUT when we compare the MAE for testing data for row 4 ~3.26 and MAE for testing data for row 1 ~3.58 and now there is a significant difference , the ridge regression performs much better on the testing data than the OLS regressor. Similar is the case of lasso regression , if we perform a similar comparison between row 1 and row 5 we will observe that in this case also OLS performs better on the training data but fails to beat the lasso regressor when predicting the test data. Another extremely important observation, one that seems to be a drawback for lasso regression is that if lambda is too large the errors are quite high ,this is because the model also makes the coefficients of relevant features zero. This can be concluded by looking at the row 7 of the table , furthermore , any lambda value greater than 10 would have similar MAEs since on those lambda values almost all the coefficients would become zero, irrespective of the importance of the features. The bar plots of residuals (of training data) are slightly misleading because theses suggest that ridge and lasso regressors are bad models but once the testing data is analysed the picture becomes clear. Hence overall the ridge and lasso regression tend to perform better than OLS regressor but only if we are able to choose the appropriate value for lambda.

Summary of conclusions and learnings: -

1. The OLS regressor is prone to overfitting because the unimportant(less important) features adjust themselves to minimise the loss and this leads to high coefficient values for such features such as the highway(rad) feature in this example of the Boston house prices. One cannot simply conclude by using OLS which are the unimportant features which have simply adjusted to the error and not the price.

2. The ridge and lasso regressors perform surprisingly well by simply using the bias and variance trade off to their benefits. Both the models increase the bias slightly to decrease the variance by a significant amount, this leads to a bad performance on the training dataset but a significant performance boost while performing on the testing data sets by performing consistently which leads to low variance.

3. However there are some drawbacks for ridge regression,the major one being that it is unable to completely eliminate unnecessary features since the value of the coefficients can only approach zero but never be exactly equal to zero.

4. The lasso regression solves the above problem by allowing the coefficient to become exactly zero but if the wrong value of lambda is used then it’d make all

the coefficients zero which is not at all desirable. The correct value for lambda can be found by performing the cross-validation process.

Bibliography

[1] https://pandas.pydata.org/docs/

[2] https://numpy.org/doc/

[3] https://scikit-learn.org/stable/

[4] https://matplotlib.org/

[5] Stack overflow for some syntax.