Advanced Regression Assignment – Part 2

Question 1

Rahul built a logistic regression model with a training accuracy of 97% and a test accuracy of 48%. What could be the reason for the gap between the test and train **accuracies**, and how can this problem be solved?

Answer

Since the training accuracy is high (97%), it means that the model has fitted well on the training dataset. But, the test accuracy has dropped to 48%. This suggests that the model has lead to **'Overfitting'**. It has tried to memorize the data points and thus unable to generalize on the test set or unseen data. This leads to high variance and low bias.

This problem can be solved by using regression methods that strongly handles 'Overfitting'. Two such methods are Ridge Regression and Lasso Regression. Both these techniques add a regularization term to the feature coefficients and make the regression model simpler while balancing the 'bias-variance' trade-off.

Question 2

List at least four differences in detail between L1 and L2 regularisation in regression.

Answer

S.No	L1(Lasso)	L2(Ridge)
1	This leads to feature selection by bringing down the coefficients of the insignificant features to 'Zero'	Ridge does not lead to feature selection, but assigns some value as coefficient to all features

2	This is computationally more intensive	This is computationally less intensive than Lasso
3	Adds the ‘absolute value of magnitude’ as the penalty term to the loss function	Adds the ‘squared magnitude’ as the penalty term to the loss function
4	A simple matrix function cannot yield the solution	A simple matrix function can yield the solution

Question 3

Consider two linear models:

L1: y = 39.76x + 32.648628

And

L2: y = 43.2x + 19.8

Given the fact that both the models perform equally well on the test data set, which one would you prefer and why?

Answer

The 2nd Model L2 should be preferred since, it is Simpler and generalisable than L1. Also, it will occupy less memory space since, more the float points, more the byte size.

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer

To make sure the model is robust and generalisable, select a model which is simpler. This may result in lower accuracy, but will work well with unseen or test data. It can be chosen using Bias-Variance tradeoff. A simple model may have high bias but has low variance. It makes sure that it has not learnt the data as is and hence, magnitude of variance on the test set is minimal.

Question 5

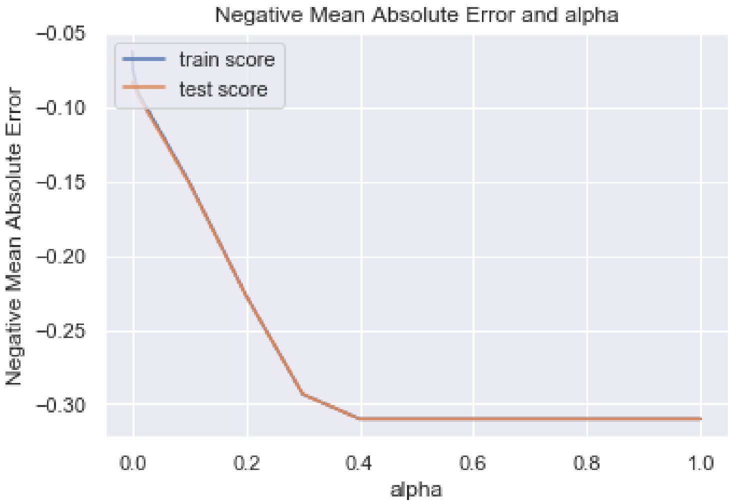
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer

In case of the **Lasso**, the objective is:

RSS + α * (sum of absolute value of coefficients)

Lower value of α indicates a lower penalty for complexity to the feature coefficient. It also indicates that when α = 0, it is similar to the Simple Linear Regression, without any regularization term. Whereas, a very high value of α indicates that most of the features will be zero.



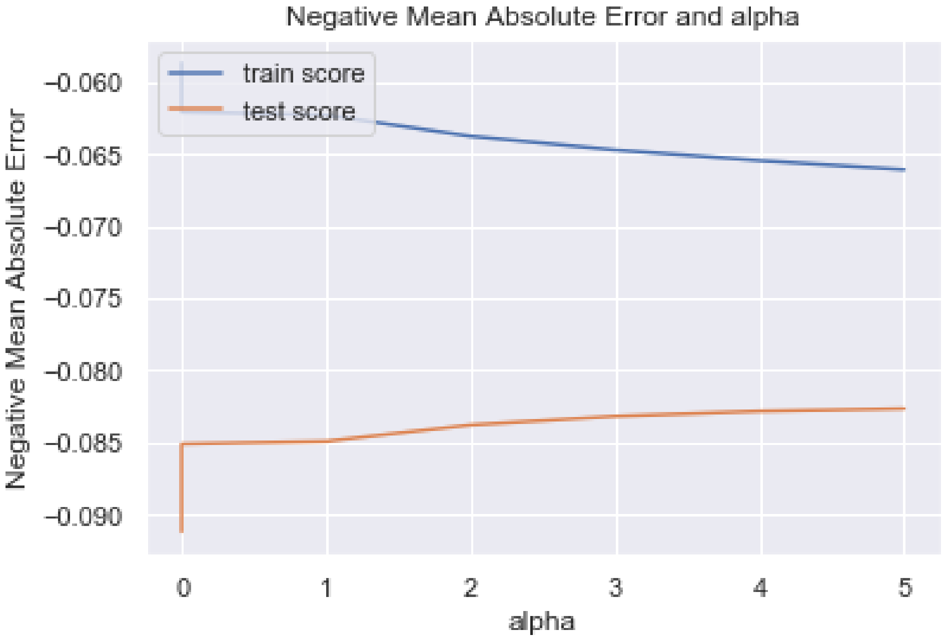
The plot of Negative of Mean Absolute Error indicates that when α =0.4, the error term stabilizes. But at this value of α, most of the coefficients become zero.

We have chosen a value **α = 0.01**, where, the significant variables are assigned a coefficient without making the model too complex.

In case of the **Ridge**, the objective is:

RSS + α * (sum of square of coefficients)

The effect of α is same as in case of Lasso.



To balance between the complexity and variance of the model, we have chosen **α = 2** in case of Ridge regression.