Advanced Regression - Assignment Part-II

Submitted By Richa Goel

Problem Statement - Part II

Question-1:

Rahul built a logistic regression model having a training accuracy of 97% while the test accuracy was 48%. What could be the reason for the seeming gulf between test and train accuracy and how can this problem be solved.

Answer:

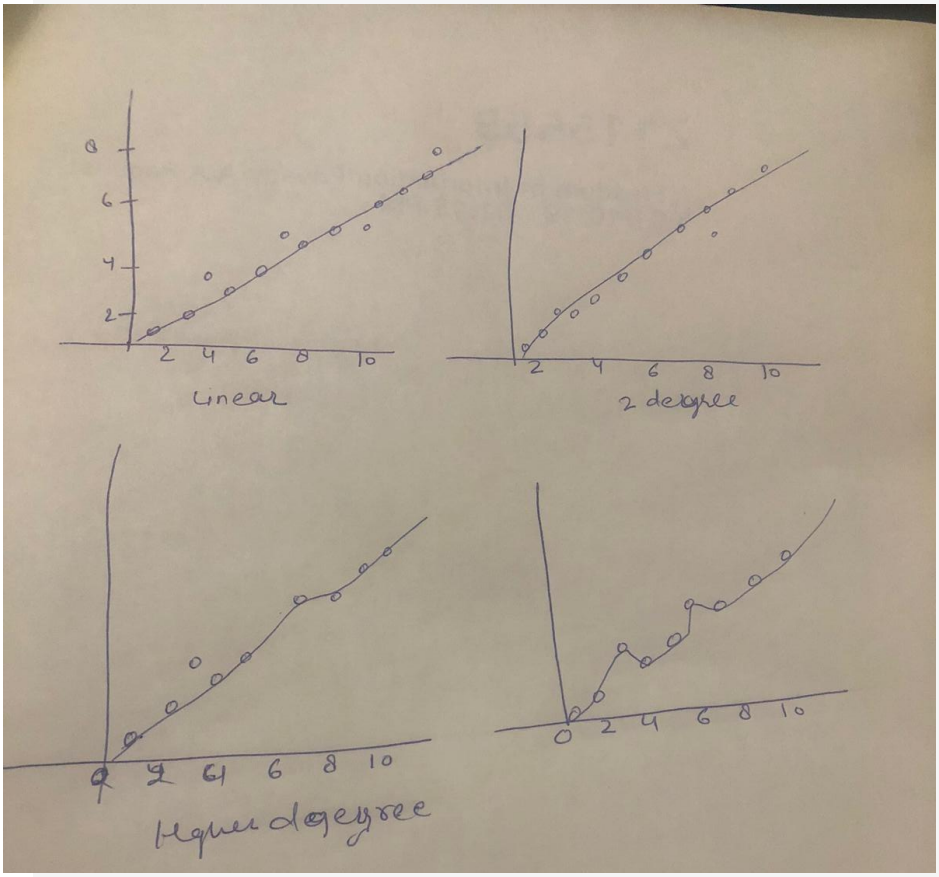
This can be a scenario of Overfitting and Complex model . An over-fitted model is too complex for any data set . It kind of memorize the training data set that's why accuracy in training set is very high but when we try to generalize these kind of model , it gets failed on test data set and accuracy is very low .

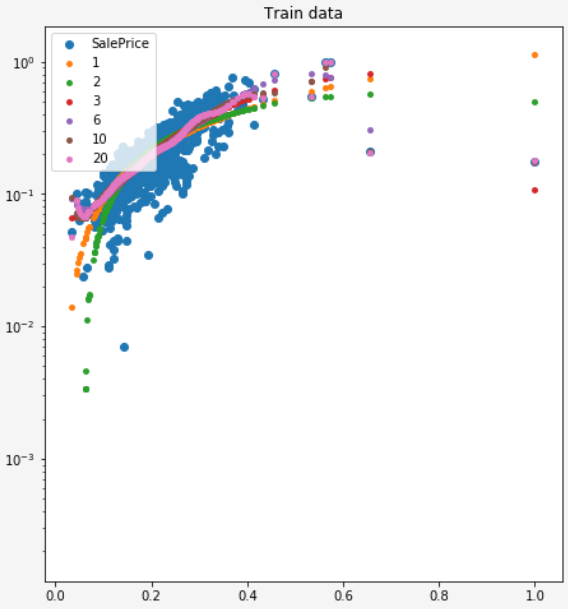
Once over-fitting occurs the regression coefficients represents the noise than the actual relationship between the variables.

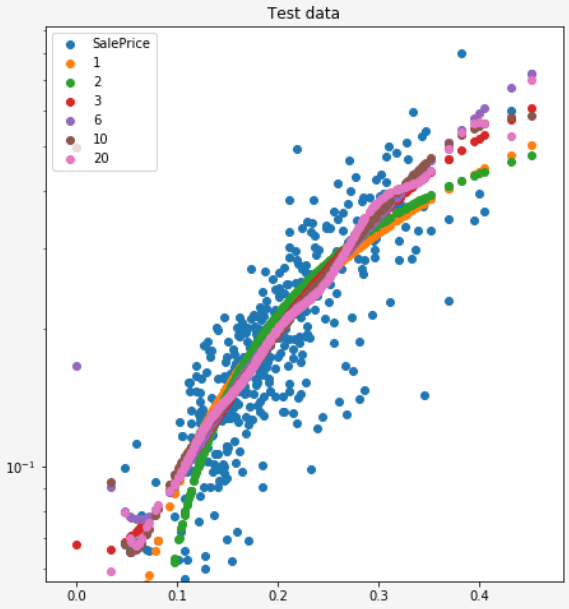
In over fitting model, If data set is change a little, this model will need to change drastically. The model is, therefore, unstable and sensitive to changes in training data, and this is called high variance.

Overfitting model the noise rather than the true data distribution.

For example :- In below diagram it seems that higher degree polynomial it best model but its not correct , it is good for known data set but if we try to predict value for unknown data then its get failed and liner modal works best in that case


linear
2 degree
Higher degree


Train data
SalePrice
1
2
3
6
10
20


Test data
SalePrice
1
2
3
6
10
20

Over-fitting are also called as overtraining. Some of the characteristics of Over-fitting in linear regression are:

1) Too many Features are present in model , basically model is too complex.
2) R2 and Accuracy is very high on Training data but very less on test data as Over-fitted model try to memorize the data and failed to generalize on unseen data .

<u>It can be resolved by using below approach/Technique</u> :-

1> Using Cross – validation also multiple run of cross validation (K-Fold Cross validation)
2> By keeping model as simple as possible
3> Regularization:- (Regularization is a process used to create an optimally complex model, i.e. a model which is as simple as possible while performing well on the training data.)
4> Bias variance tradeoff

The basic idea to avoid Over-fitting is to either explicitly penalize complex models for use of more features (hyper-parameters) or to test the ability of the model to generalize by evaluating model performance on unseen data.

Question-2: -List at least 4 differences in detail between L1 and L2 regularization in regression.

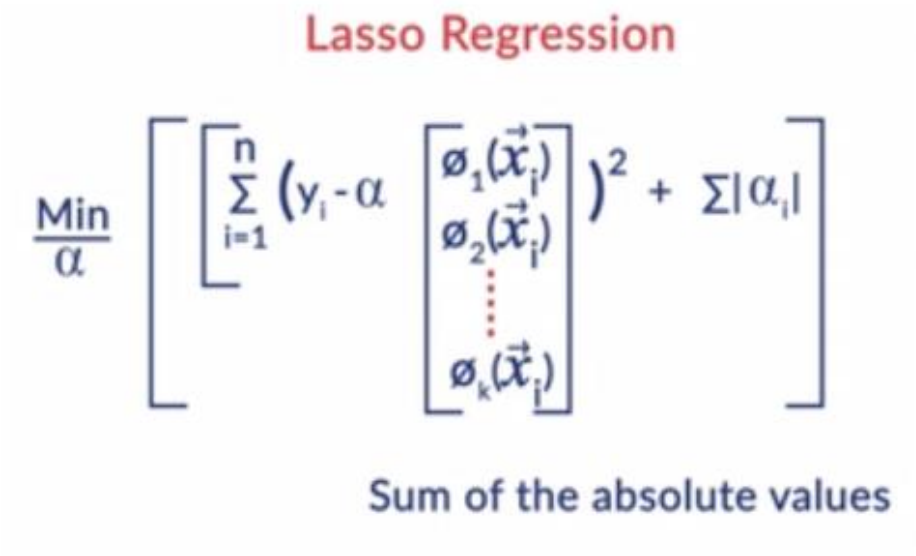
Answer:

L1 and L2 are two type of mechanism, which is used to avoid overfit, and called regularization. Regularized model contains a another element that should be minimized.

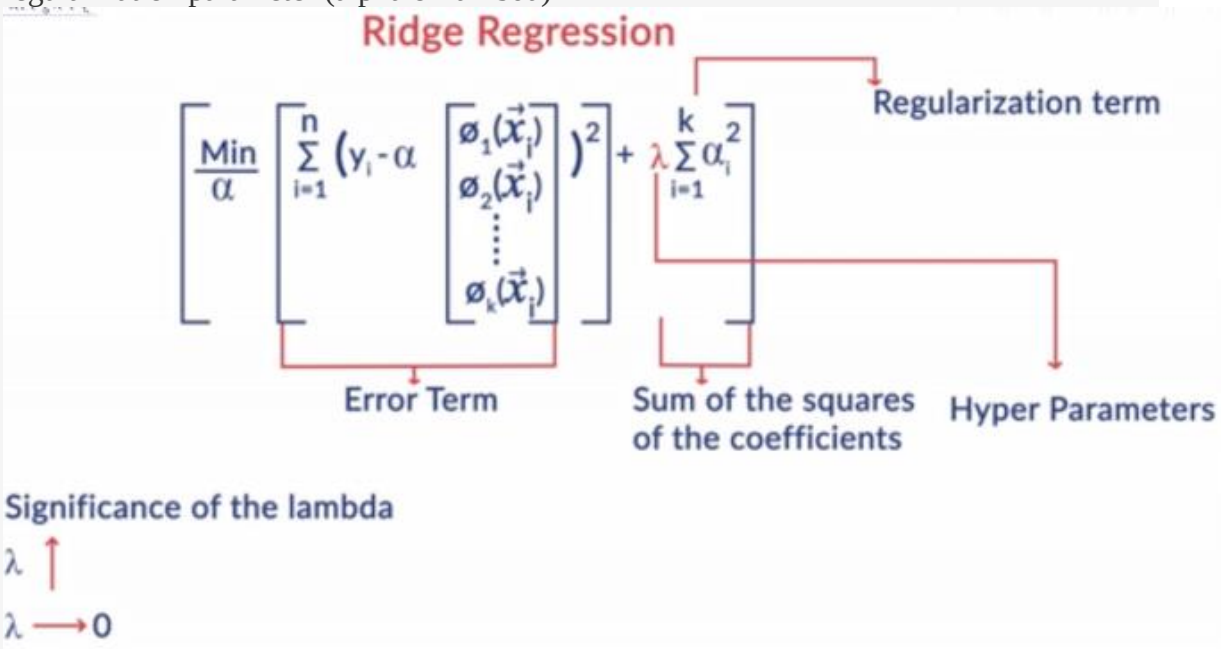
. Below are the key differences between them:

1> L1 Regularization is used in Lasso Regression while L2 regularization is used in Ridge Regression to impose a penalty on the size of coefficients.

2> L1 Regularization term are sum of the absolute value of magnitude of coefficient multiplied by the regularization parameter (alpha or lambda)



L2 Regularization are sum of squared magnitude of coefficient multiplied by the regularization parameter (alpha or lambda)



So basically both of them differs with respect to the penalization term.

3> Lasso Regression results into features selection. If we have redundant features included in our model, then L1 Regularization term (Lasso Regression) reduces the coefficient of those redundant features to zero. While Ridge(L2 Regularization) Regression results into the same number of features as provided although it reduces the coefficients of not important features to a lower value but does not make it completely zero. Thus Lasso Regression results into a Sparse Model.

4> Ridge regression has always matrix representation for the solution while Lasso requires multiple iterations to reach to the final solution. Thus Ridge regression is computationally faster than the Lasso regression.

5> L1 Regularization term as represented in the equation represents a three dimensional Rhombus (Diamond) while error term can be represented using an ellipsoid. L2 Regularization (Ridge) term as represented in the equation represents a Sphere.

6) Ridge regression is computationally less intensive than Lasso regression.

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 dataset, which one would you prefer and why?

Answer:

The Model *L2: y = 43.2x + 19.8* is preferred over the L1 model because of the Model simplicity reasons.

L2 and L1 model though both are Linear models L2 model has much lesser representational complexity and is much simpler than the L1 as in both the terms of the equation we have simpler coefficients.

The constant term in L2 model is only 3-bits in total and one bit after the decimal point while in L1 model it is total 9 bits and 7 bits after the decimal. Also the coefficient beta-1 in L2 Model is just 3 bits in total and one bit after the decimal point while the same coefficient beta-1 in L1 Model is 4-bits in total and has two bits after the decimal point. Due to this reason L2 Model will take much lesser space and will use much lesser resources than the L1 model.

Everything else being same we will always prefer the simpler model (L2) as it is likely to generalize far better than the complex model. Just to bring the error term low if the coefficient is becoming complex it should not be allowed as it will be tough to get minimum solution for the equation. Due to this reason regularization, term is introduced in Linear Regression to keep the coefficients simple.

Question-4:

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

Answer: A model is robust and generalized if its performance and adaptability are same when applied to new conditions while maintaining the same basic set of explanatory. Some of the features and methods to get a robust and generalized models are discussed below:

a> Simplicity – Simplification usually produces more precise models. When we have several models with similar predictive power, we should opt for the simplest because it is the most likely to be the best model.

b> Multicollinearity – It occurs when independent variables in a regression model are correlated. Due to this even a small change in the equation can produce dramatic changes in the coefficients and p-values. It also reduces the statistical significance for relevant variables.

c> Avoid Under-fitting / Over-fitting -There should be correct number of independent variables in the regression model. Too few that is underspecified or Under-fitted models tend to be biased. While too many or Over-fitted models tend to be less precise. So we need to find the right balance and correct terms for a precise and un-biased model. The performance scores of a training and test data should be similar to avoid overfitted model.

d> Outliers treatment - We should properly treat the outliers in Training dataset. We can impute outliers properly by using proper methods like mean, median etc or removing outliers completely by performing EDA.

e> Adjusted R-squared – A regular R-squared always increases when we add an independent variable and this is tempting to specify a model that is too complex producing misleading results. To avoid this we should use Adjusted R-squared and Predicted R-squared. Adjusted R-squared increases only when a new variable improves the model. A low quality variable can decrease Adjusted R-squared value.

f> Cross validation – Cross validation partitions the data to determine whether the model is generalizable outside the provided data set. We can have K-Fold cross validation and compute predicted R-squared which can decrease for low quality varaibles.

g> P-values for independent variables – p-values less then significance level indicate that the term is significant. In regression we use this for reducing the model. We include all the candidate variables and then repeatedly remove single term with the highest non-significant p-value until the model contains only significant term.

h> Stepwise Regression and Best Subset regression – These are automated model selection algorithms that pick the variables to include in the regression equation. These are helpful when we have many independent variables. These gives statistics to help us balance the tradeoff between precision and bias.

Question-5:

As you have determined the optimal value of lambda for ridge and lasso regression during the assignment, which one would you choose to apply and why?

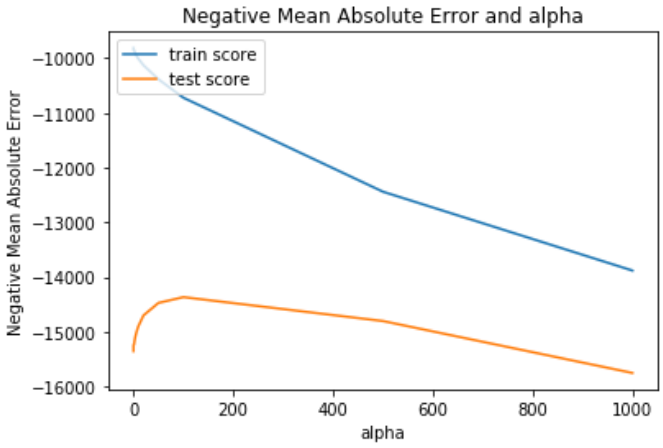
Answer:

Lasso regression has feature selection property that is it makes the coefficient of irrelevant and not so important variables to zero. Thus we have decided to take the benefit of Lasso feature selectivity properties first and thus we performed Lasso first to find the optimal value of lambda or alpha and then perform the Ridge regression. After Lasso we get the optimal value of Lambda as 500.

But after Ridge we have got the optimal value of Lambda as 100.

We will be using the final value as 100, which we have received finally after Ridge Regression because of the following reasons-

Larger Negative Mean Absolute Error is better. As alpha increases we actually decrease accuracy.



At low value of alpha accuracy is more. As alpha increases the average test performance scores becomes better till a point but beyond that the test performance scores starts decreasing. Here as per the above graph we can see that at point where alpha = 100 we test performance has reached maximum value and after that the test performance is constantly decreasing.

Thus alpha =100 is the optimal value to be used, which we got finally after the Ridge regression.