Advanced Regression: Assignment Part – II

Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Solution:

Optimal Values of alpha for Ridge Regression is: 5 and

Optimal value of alpha for Lasso Regression is: 10

Let us first understand what the significance of alpha is.

Alpha is a regularization coefficient which is designed to penalize model complexity. Therefore, higher the alpha, less complex is the model, decreasing the error due to variance (i.e. overfit).

On the other hand, if alpha is too, it also increases bias error because of underfitting. So, we choose that value of alpha which helps to balance the error in both directions.

Below is the summary report for both Ridge and Lasso Regression:

Regression Type	Alpha	Train R2 score	Test R2 score	Most Important Predictor Variables and their Coefficients	Effect on model after choosing double value of alpha
Ridge	Best **alpha = 5**	0.939	0.921	**Top 5 Features:** BsmtHalfBath: 0.233 TotalPorchSF: 0.222 SaleCondition_Others: 0.18 RemodelAge**:** -0.24 LotFrontage: -1.894 **Total Number of features to be considered: 180**	As we increased the value of alpha, R2 score of train dataset has been decreased and that of test dataset has been increased. Also, with increase of alpha, penalty has been increased and so coefficients of our features have been reduced. As we increase alpha, variance decreases and bias increases.
	Double Value = 10	0.9367	0.9225	**Top 5 Features:** TotalPorchSF: 0.223 BsmtHalfBath: 0.222 SaleCondition_Others: 0.16 RemodelAge: -0.206 LotFrontage: -1.895	

				Total Number of features to be considered: 180	
Lasso	**Best alpha = 0.001**	0.903	0.909	**Top 5 Features:** TotalPorchSF: 0.26 BsmtHalfBath: 0.15 OverallCond: 0.14 RemodelAge: -0.25 LotFrontage: -2.07 **Total Number of features to be considered: 24**	As we increased the value of alpha, R2 score of train dataset has been decreased and that of test dataset has been increased. Also, wit increase of alpha, penalty has been increased and so coefficients of our features have been reduced. As we increase alpha, variance decreases and bias increases.
	Double alpha = 0.002	0.893	0.902	**Top 5 Features:** TotalPorchSF: 0.24 OverallCond: 0.16 BsmtHalfBath: 0.14 RemodelAge: -0.209 LotFrontage: -1.97 **Total Number of features to be considered: 15**	

Question 2

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?

Solution:

After carefully examining both Ridge and Lasso Regression techniques, I found that

- Using Ridge Regression with best alpha, my model can explain 94% of variance in train dataset but only 92% of variance for test dataset.
- But using Lasso Regression with its best alpha value, model can explain 90% of variance for both train and test dataset.

Here, Ridge is giving us an impression of slight overfitting, which may be because of presence of high number of features involved in the model. As Ridge regression doesn't give coefficient value of 0 even for highly insignificance variables, I will go with Lasso Regression. Reason of choosing lasso regression is: Even though r2 score produced lasso model is less in comparison to Ridge model, but using lasso model, I just need to consider 24 variables but with Ridge model, I will have to consider all 180 variables.

Lasso also helps in feature reduction.

And if I tweak the value of alpha a little (for example I doubled the value of alpha in my case study), I got almost same r2 score with number of features even reduced to 15.

Question 3

After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Solution:

There are various techniques to deal with the model. Some of them are:

1. One of which is – we can recalibrate the model.
2. We can also replace those predictor variables with those variables which have a strong relationship with those varaibles.
3. We can rebuild the model.

In our case, if five most important predictor variable in lasso are no more a part of file, we can either replace those predictor variables with those variables which have strong relationship with them.

If that is not possible, we will need to recreate the model. After recreating the model, Top 5 important predictor variables are:

- BuiltAge
- LowQualFinSF
- OverallQual
- BedroomAbvGr
- HeatingQC

Question 4

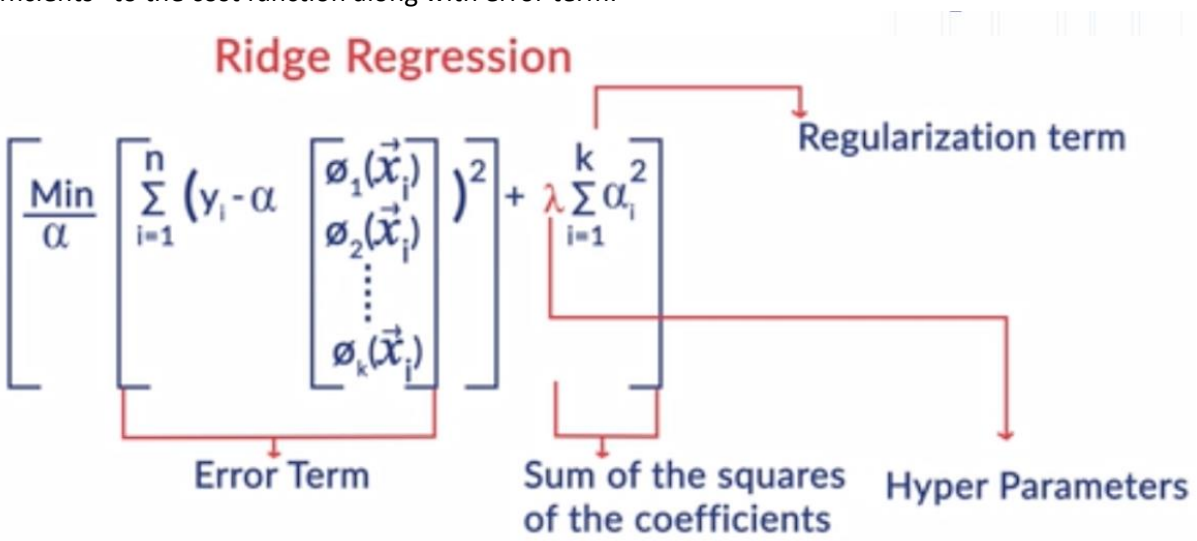
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Solution:

A model is said to be robust and generalizable when it works well on unseen data. In our real-life scenarios, we prefer a model to be more robust and generalizable because it always have to work on unseen data. A simple model is not sensitive to the specifics of training dataset as their more complex counterparts are.

To make a model more robust, we can make use of below techniques:

1. **Eliminate features with strong relationship**
2. **K-fold Cross-Validation:** In this technique, training data is divided into k subsets. In each iteration k-1 subsets are chosen as training dataset and kth subset is chosen as validation dataset. In this way, our model is not able to take a peak into test dataset and will be more robust.
3. **Regularization:** Regularization is the simplification done by the training algorithm to control model complexity. It is a very popular method to reduce overfitting (and reduce the problem of high variance). Under this method, we penalize the model for being too complex.
   Most popular regularized regression models are:
   a. **Ridge Regression:** In this technique, we add an additional term of "sum of squares of coefficients" to the cost function along with error term.



   b. **Lasso Regression:** In this technique, we add "absolute value of coefficients" into cost function along with error term.
   Cost function can be represented as:

Lasso Regression

$$\underset{\alpha}{\text{Min}} \left[\left[\sum_{i=1}^{n} \left(y_i - \alpha \begin{bmatrix} \phi_1(\vec{x}_i) \\ \phi_2(\vec{x}_i) \\ \vdots \\ \phi_k(\vec{x}_i) \end{bmatrix} \right)^2 + \sum |\alpha_i| \right] \right]$$

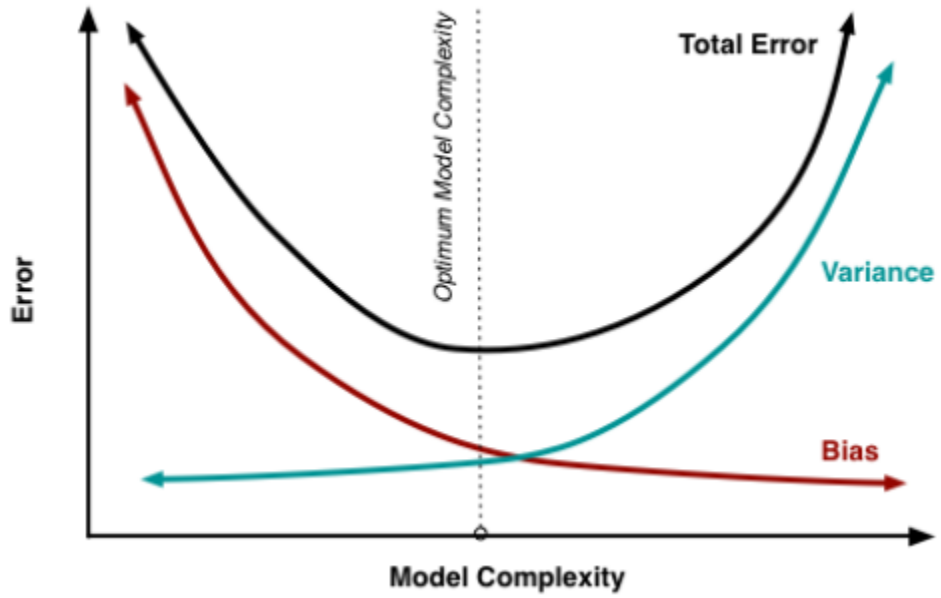
Regularization Term

Sum of the absolute values

Implications on Accuracy of model:

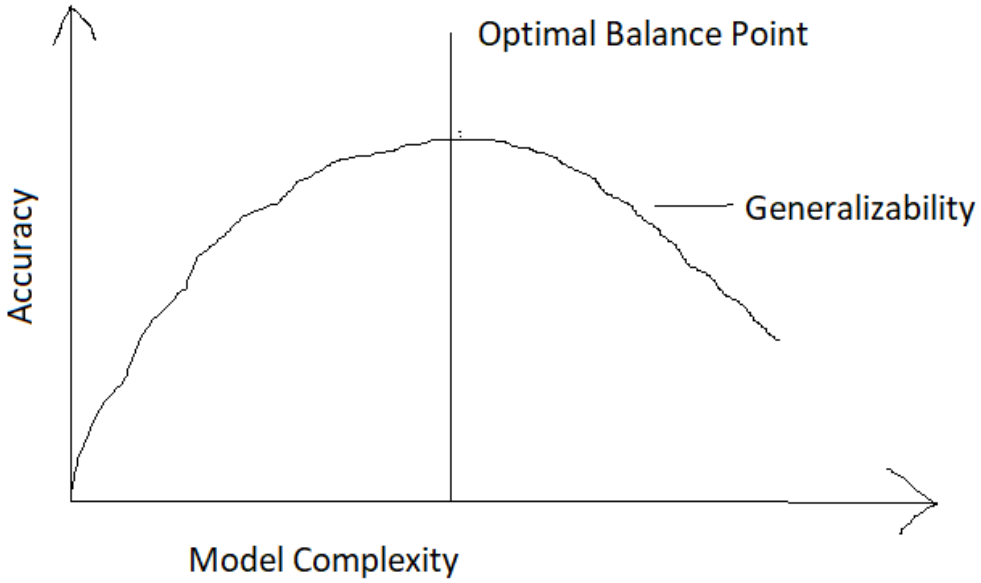
Having said that we should always keep in mind that a model should not be too naïve that it is not able to give us correct result at all (low variance and high bias). There should be an optimal balance between model complexity and accuracy of a model so that it can balance out both variance and bias error.

Below is the diagram which represents relationship between Model complexity and error term:



It shows that less complex is the model, it will result into underfitting (high bias) and more complex is the model, it will result into overfitting i.e. high variance.

Let’s look at below graph to see impact of model complexity on accuracy of a model:



Above graph shows that a with increase in complexity of a model, its generalizability will reduce and hence the accuracy with unseen data.

But our model should not be too simpler because that also will not be able to provide us good result.