Advanced Regression Assignment

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?

Ans – The optimal value of thee hyperparameter alpha for LASSO regression came to 0.001 and the value of alpha for Ridge regression came to 0.5.

Doubling the value of alpha in LASSO (L1) regularisation, reduced the number of features from 47 to 28. Also, the train and test R-squared values changed from 0.88 and 0.84 respectively to 0.86 and 0.81 respectively. For Ridge (L2) regularisation, the train and test R-squared values change from 0.88 and 0.8 respectively to 0.9 and 0.8 respectively.

For both LASSO and Ridge regression, the most important predictor variables remain the same even after doubling the optimal alpha value.

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?

Ans – From the results, we have found that the performance of both the L1 regularised and the L2 regularised models are fairly well (greater than 80 percent in both the cases). However, due to the presence of such a large number of features (after creating the dummy variables), the choice of Lasso regression seems more viable in this case, since Lasso regression automatically performs feature selection/elimination. That step has to be done separately in Ridge regression. Since the model performances are similar (a little better test performance using LASSO), therefore, I would choose to go with Lasso regression.

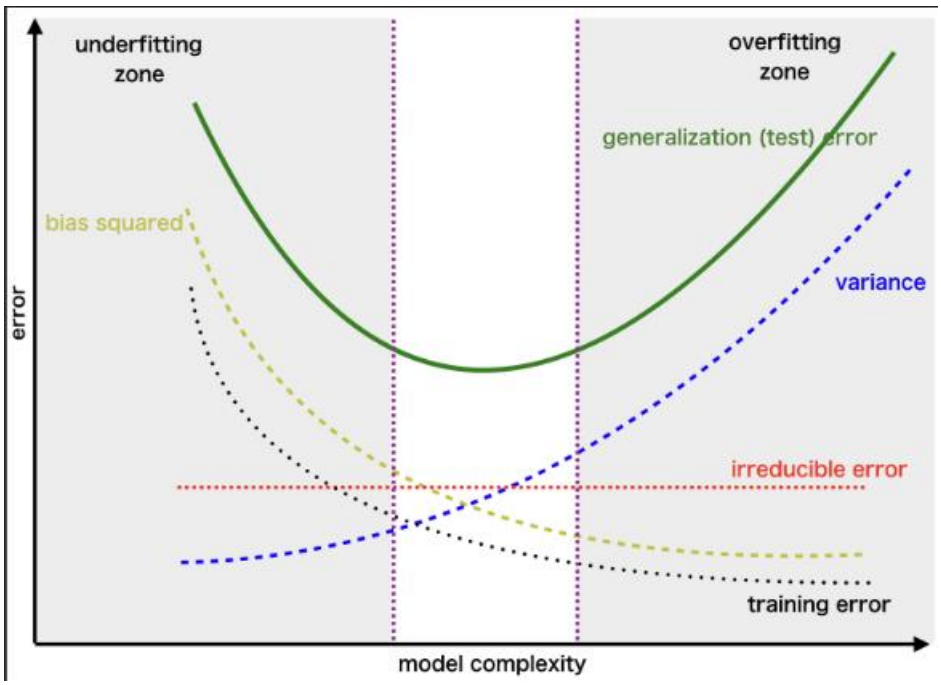
3) After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables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

Ans – The 5 most important predictor variables after removing the original top 5 features are:

a) SaleType_New - Home just constructed and sold
b) ExterQual - Quality of the material on the exterior
c) BsmtQual - height of the basement
d) 2ndFlrSF - Second floor area in sqft
e) 1stFlrSF - First floor area in sqft

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?

Ans – To make a model robust and generalisable, we have to make sure to follow Occam's Razor, i.e., make a model only as complex as it needs to be. We need to make a model such that it picks up on the general trends and patterns in the data without learning the noise. To do this, we have to take care of something called the Bias-Variance Tradeoff.



As a model becomes complex, its bias decreases. However, its variance increases. Similarly, a simple model has a high bias but low variance. We have to achieve a balance or an optimal point which ensures that the model is just right i.e., neither too simple, nor too complex. This can be achieved using various techniques, one of which is regularization.

This will cause the overall accuracy of the model on the training set to decrease, but will reduce the variance and hence, the model will be more generalized and robust.