Advanced Regression Assignment

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?

Ans) Optimal Value of alpha for ridge and lasso regression are:

- Optimal Value of lambda for Ridge: 2
- Optimal Value of lambda for Lasso: 0.002

With these alphas the R2 of the model was approximately 0.784.

After doubling the alpha values in the Ridge and Lasso, the prediction accuracy remains around 0.7 but there is a small change in the co-efficient values. The new model is created and demonstrated in the Jupiter notebook. Below are the changes in the co-efficient.

Overall since the alpha values are small, we do not see a huge change in the model after doubling the alpha.

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?

Ans) The optimum lambda value in case of Ridge and Lasso is as follows:-

- Ridge – 1
- Lasso – 0.0002

The Mean Squared Error in case of Ridge and Lasso are:

- Ridge - 0.03505905306547901
- Lasso - 0.03433584007843537

The Mean Squared Error of both the models are almost same.

Since Lasso helps in feature reduction (as the coefficient value of some of the features become zero), Lasso has a better edge over Ridge and should be used as the final model

Advanced Regression Assignment

Question 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 five most important predictor variables in the current lasso model is:-

1. Neighborhood_NridgHt
2. Neighborhood_StoneBr
3. MSSubClass_2-1/2 STORY ALL AGES
4. SaleCondition_Partial
5. Fireplaces

We build a Lasso model in the Jupiter notebook after removing these attributes from the dataset.

The R2 of the new model without the top 5 predictors drops to .71

The Mean Squared Error increases to 0.045705353245799406

The new Top 5 predictors are:

	Lasso Co-Efficient
Neighborhood_NridgHt	0.215451
Neighborhood_StoneBr	0.214267
MSSubClass_2-1/2 STORY ALL AGES	0.159334
SaleCondition_Partial	0.102290
Fireplaces	0.098702

Advanced Regression Assignment

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?

Ans)

As Per, Occam's Razor— given two models that show similar 'performance' in the finite training or test data, we should pick the one that makes fewer on the test data due to following reasons:-

- Simpler models are usually more 'generic' and are more widely applicable
- Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
- Simpler models are more robust.
  - Complex models tend to change wildly with changes in the training data set
  - Simple models have low variance, high bias and complex models have low bias, high variance
  - Simpler models make more errors in the training set. Complex models lead to overfitting — they work very well for the training samples, fail miserably when applied to other test samples

Therefore, to make the model more robust and generalizable, make the model simple but not simpler which will not be of any use.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.

Also, Making a model simple leads to Bias-Variance Trade-off:

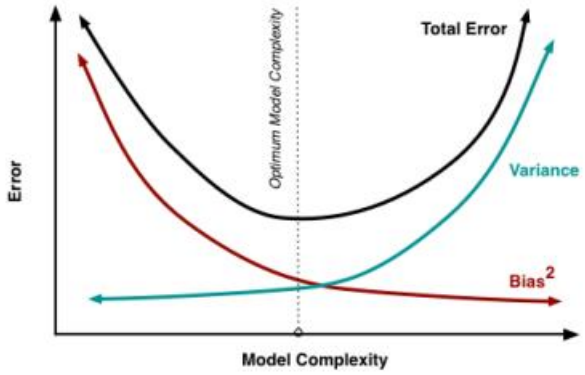
- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.

Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naïve, for e.g., one that gives same answer to all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high.

Advanced Regression Assignment

Variance refers to the degree of changes in the model itself with respect to changes in the training data.

Thus accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph



• Use a model that's resistant to outliers. Tree-based models are generally not as affected by outliers, while regression-based models are. If you're performing a statistical test, try a non-parametric test instead of a parametric one.

• Use a more robust error metric. switching from mean squared error to mean absolute difference (or something like Huber Loss) reduces the influence of outliers. I explain a bit about why this is the case at Why is the median a measure of central tendency? It doesn't have anything to do with any other values of the data set, so how does it "describe" the data set?

Here are some changes you can make to your data:

• Winsorize your data. Artificially cap your data at some threshold. See What are some applications of winsorization?

• Transform your data. If your data has a very pronounced right tail, try a log transformation.

• Remove the outliers. This works if there are very few of them and you're fairly certain they're anomalies and not worth predicting