Advanced Regression Assignment Subjective

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?

Answer (1)

- Optimum value for lambda in ridge regression is 10
- Optimum value for lambda in lasso regression is 100

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?

Answer (2)

We found the optimal values of lambda for lasso and ridge to be 100 and 10 respectively.

Now, which one we choose to apply depends on the data we have. Ridge regression is bit easier to implement and faster to compute and on the other hand Lasso regression have feature selection property.

So, if we have less computation power then we would go for Ridge regression and if we want to reduce the number of feature we would go for the Lasso regression. Lasso regression is generally the model of choice because feature selection is most important in most cases.

From the output of both models, it is evident that Lasso does feature selection and its output is much simpler than Ridge's, without any compromise in the accuracy over the test data. We will select lasso regression with this dataset.

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?

Answer (3)

five most important predictor variables in the lasso model :

RoofMatl_WdShngl	62021.528773
Neighborhood_NoRidge	42745.803789
Neighborhood_NridgHt	28043.296456
Neighborhood_Crawfor	24322.458606
2ndFlrSF	20244.091551

we create another model excluding the five most important predictor variables after that five most important predictor variables now are

BsmtExposure_Gd	20139.763146
Neighborhood_Somerst	20079.126199
Neighborhood_StoneBr	18303.842819
Exterior1st_BrkFace	16038.774978
HouseStyle_1Story	15740.604142

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 (4)

Robustness

- **Robustness** means the property of the model according to which whether the model is tested on the training set or test set, the performance is same.
- The **robustness** is the characteristic describing a model's ability to effectively perform while it's variables or assumptions are altered.

Generalisability

- **Generalizable** means the model has learnt with few variables but are more likely to predict on the unseen data.
- The **generalisability** is the characteristic describing a model's ability to find abstract pattern in data and which can be applied to a large variety of data it can encounter.

To make a model more robust and generalizable it should be simpler, remove the outliers etc., so that the model can predict well for the unseen data. By making model more robust and generalizable the accuracy of the model gets reduced. Since it is measured by how much the model is able to predict the training data. And since the model is trying to be more robust, it will try to avoid the noise in the data and try to be more generic this leads to lesser accuracy. This can also be understood using the Bias variance trade-off.

Advantages of simpler model are:

- Models are more generic: if the model has learnt every feature then that model can answer the question which is similar. But if the question is unfamiliar the complex model is more likely to make error. The simpler model just learns the basic principles and when the unfamiliar question arises they are less likely to make error as compared to the complex model.
- Models are robust: Complex models are more sensitive to the data set. As the training set changes they are more likely to swing, which is not in the case of simple models.

Thus we can say**, Complex models have high variance and low bias and Simple models have high bias and low variance.**