Advance Regression

House Price Prediction

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:

Below are the optimal values of alpha for ridge and lasso regression:

1. Ridge regression : alpha = **100**
2. Lasso regression : alpha = **0.01**

Change if we choose double the value of alpha for both ridge and lasso i.e:

1. Ridge regression alpha from 100 → 200

- R2 score for train set dropped from 0.882 to 0.878
- R2 score for test set dropped from 0.867 to 0.864
- MSE error value increased from 0.0218 to 0.0223
- RMSE error value increased from 0.147 to 0.149
- Thus, after doubling all the metrics degraded.

2. Lasso regression alpha from 0.01 → 0.02

- R2 score for train set dropped from 0.868 to 0.849
- R2 score for test set dropped from 0.861 to 0.844
- MSE error value increased from 0.0227 to 0.0256
- RMSE error value increased from 0.15 to 0.16
- Thus, after doubling all the metrics degraded.

The predictor variables are not changed after doubling the alpha Value for Ridge and Lasso.

The most important Variable in Predicting House Sale Price in **Ridge model** are:

1. OverallQual
2. GrLivArea
3. OverallCond
4. GarageCars
5. 1stFlrSF

Definations:

1. OverallQual : Rates the overall material and finish of the house
2. GrLivArea : Above grade (ground) living area square feet
3. OverallCond : Rates the overall condition of the house
4. GarageCars : Size of garage in car capacity
5. 1stFlrSF : First Floor square feet

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:

It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance, and making the model interpretably.

Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

The **Optimal Value** of alpha for ridge and lasso regression:

- Ridge: 100 (R2 score = 0.882 & RMSE = 0.147)
- Lasso: 0.01 (R2 score = 0.868 & RMSE = 0.15)

- The Ridge Regression has better scoring metric (R2 score) on train and test dataset.

- However, the predictor variables we get using lasso helps in determining House Prices much more effectively.

- We can use Lasso Regression as Business demands to identify Predictor Variables that has a significant impact on the House Prices.

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:

If we have top 5 predictor variables missing, we consider next 5 variables with a significant impact on the Response Variable in **Lasso model**.

These features could be:

~~1. OverallQual~~
~~2. GrLivArea~~
~~3. GarageCars~~
~~4. OverallCond~~
~~5. Fireplaces~~

1. CentralAir : Central air conditioning
2. Electrical : Electrical system
3. FireplaceQu : Fireplace quality
4. BedroomAbvGr : Bedrooms above ground floor
5. BsmtFinSF2 : Type 2 finished square feet

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:

The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

The concept of **Occam's Razor** comes handy while selecting a Generalized model.

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.

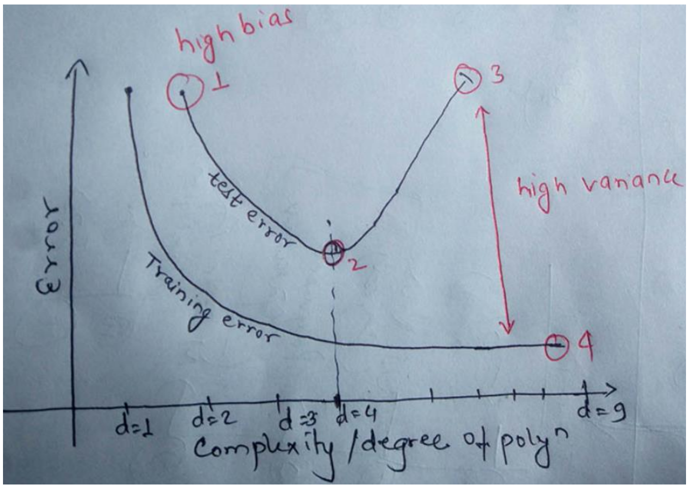
Also, Making a model simple leads to Bias-Variance Trade-off:

- 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.

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


high bias
high variance
test error
Training error
Error
d=1
d=2
d=3
d=4
d=9
Complexity / degree of polyn