Assignment Subjectives

By: Siddharth Singh

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?

Optimal value for alpha for Ridge is – 2.0.

Optimal value of alpha for Lasso is – 0.0035.

Changes If we use double the values for alpha/increasing Alpha:

For Ridge:

As we increase the alpha/double it will result in more shrinkage of coefficients.

For Lasso:

As we increase the alpha/double it will also result in shrinkage of coefficients but more coefficients become zero.

Most Important Variable Ridge:
Optimal alpha:

1. OverallQual_9 - 0.379492
2. MSZoning_FV - 0.275670
3. MSZoning_RL - 0.224364
4. OverallQual_8 - 0.190480
5. OverallQual_10 - 0.176134
6. OverallCond_9 - 0.160753
7. GrLivArea - 0.158244
8. GarageCars_4 - 0.150730
9. Neighborhood_ClearCr - 0.142735
10. Neighborhood_Crawfor - 0.134602

Doubling of alpha:

1. OverallQual_9 - 0.327648
2. MSZoning_FV - 0.206736
3. MSZoning_RL - 0.162133
4. OverallQual_8 - 0.159807
5. GrLivArea - 0.158041
6. OverallCond_9 - 0.139297
7. Neighborhood_ClearCr - 0.129488
8. Neighborhood_Crawfor - 0.128054
9. OverallQual_10 - 0.121593
10. GarageCars_3 - 0.115262

Most Important Variable Lasso:
Optimal alpha:

1. OverallQual_9 - 0.494490
2. MSZoning_FV - 0.376280
3. MSZoning_RL - 0.317962
4. OverallQual_10 - 0.293933
5. OverallQual_8 - 0.284265
6. MSZoning_RH - 0.227916
7. MSZoning_RM - 0.213152
8. GarageCars_4 - 0.208806
9. OverallCond_9 - 0.169564
10. GrLivArea - 0.159649

Doubling of alpha:

1. OverallQual_9 - 0.442373
2. MSZoning_FV - 0.256026
3. OverallQual_8 - 0.240928
4. OverallQual_10 - 0.218044
5. MSZoning_RL - 0.203115
6. GrLivArea - 0.161266
7. OverallCond_9 - 0.159821
8. GarageCars_4 - 0.152831
9. Neighborhood_ClearCr - 0.136456
10. GarageCars_3 - 0.127860

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?

- I would use lasso regression although its requires more computational power but in this case since data is not that vast Lasso is giving fast result. Main reason for choosing Lasso regression is its in built feature selection. Since Lasso regression drops the irrelevant features it makes our model more robust.

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?

New Top 5 Variables:

- GarageCars_4.
- OverallCond_9.
- GarageCars_3.
- GrLivArea.
- Neighbothood_ClearCr.

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?

By reducing the overfitting in our train data we can make our model more robust and generalisable. Some Methods to make our model more robust and generalisable are as follows:

- Using CrossValidation.
- Train with more data.
- Removing irrelevant features.
- Regularisation Technique(L1 and L2 norm).
- Ensembling (Bagging and Boosting).

A Model needs to be robust and generalisable so that they are not impacted by the outliers and variance in dataset. A model also needs to be generalisable so that **accuracy** does not drop on test set. Usually drop in **accuracy** indicates that model is overfitting on training set. If the model is robust and generalisable it means that model is not overfitting and can produce good result on dataset outside training and test set.