
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: Optimal alpha values found from ridge and lasso methods is Ridge : 5 and Lasso : 1

If they are doubled,they become Ridge : 10 and Lasso : 2

Before doubling alpha values :

R2 score for train data – 0.900 and test data – 0.875 for Ridge

After doubling alpha values :

R2 score for train data – 0.892 and test data – 0.874 for Ridge

Before doubling alpha values :

R2 score for train data – 0.938 and test data – 0.830 for Lasso

After doubling alpha values :

R2 score for train data – 0.938 and test data – 0.830 for Lasso

---

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:

Since lasso helps in feature shrinkage and provides a significant sub set of predictors variables since it has the ability to mitigate irrelevant columns in the dataset making them close to zero where they become insignificant, so it is good to go with Lasso as final model.

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:

Before dropping important predictor variables on Lasso :

- BsmtUnfSF
- LotArea
- BsmtFinSF1
- MSSubClass
- TotalBsmtSF
- 1stFlrSF
- YearBuilt
- OverallQual
- YearRemodAdd
- OverallCond
- BsmtFinSF2

After dropping above significant values, lasso predictors obtained:

- BsmtFullBath
- BsmtHalfBath
- HalfBath
- BedroomAbvGr
- FullBath
- LowQualFinSF
- GrLivArea
- KitchenAbvGr
- TotRmsAbvGrd

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:

- Robust implies strong model and generalisable means which is less affected if any new data is included.
- Model should also be with good score both for training and test data.
- If outliers can be removed to the best effort then the model is least affected by new data.
- Model should be kept as simple when compared to complex ones since they might lead to overfitting and may fail on test data.
- If model is robust, accuracy would be high which might lead to test data failing.