2022

Advanced Regression Subjective Questions Solutions

SUBMITTED BY:

BALABHASKAR ASHOK KUMAR – MLC35

Assignment-based Subjective Questions

Q1) 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?

Soln.)

Ridge Regression

1. The optimal value for the Ridge Regression was clocked at 6. This resulted in an R2 Score value of 0.896 for the train data set and a value of 0.824 for the test data. The most important predictor variable (based on the magnitude of coefficient) in this case was OverallQual followed by GrLivArea and 2ndFlrSF
2. Assuming the scenario where we double the alpha value (i.e 12) – the resultant R2 Score of train and test data was R2 score was 0.889 (a decrease of 0.78%) and 0.819 (a decrease of 0.60%) respectively. The important predictor variable in this was OverallQual followed by GrLivArea and TotRmsAbvGrd.

   Hence by an increase of the alpha value there was a slight negative effect on the R2 Score. Although the top two variables stayed the same – the third important variable swapped position.

   Note: As we move further and check there is significant difference in other variables – but their coefficient power is negligible compared to the top three attributes

Lasso Regression

1. The optimal value for the Lasso Regression was clocked at 0.001. This resulted in an R2 Score value of 0.879 for the train data set and a value of 0.812 for the test data. The most important predictor variable (based on the magnitude of coefficient) in this case was GrLivArea followed by OverallQual and GarageCars
2. Assuming the scenario where we double the alpha value (i.e 0.002) – the resultant R2 Score of train and test data was R2 score was 0.856 (a decrease of 2%) and 0.819 (a decrease of 2%) respectively. The important predictor variable in this was GrLivArea followed by OverallQual and GarageCars

   Hence by an increase of the alpha value there was a slight negative effect on the R2 Score. An interesting result was the top three variables remained the same in spite of doubling the alpha value.

Q2) 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?

Soln.) The optimal value for ridge and lasso regression was 6 and 0.001 respectively as per the analysis.

The better approach would be to move forward with Lasso Model – Although this has slightly lesser accuracy than the Ridge model – the number of variables have dropped to 34 to 112. This will create a simpler model and has better coefficient magnitude compared to Ridge.

Q3) 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?

Soln.) The five most important variable for the Lasso Regression in the order of decreasing importance was as follows:

1. GrLivArea
2. OverallQual
3. GarageCars
4. KitchenQual
5. BsmtQual

As a path forward with respect to the question - these 5 drops where dropped and Lasso model was run. The result is described below:

1. R2Score on Train Set : - 0.853
2. R2Score on Test Set : - 0.785

The top 5 important variable then becomes:

1. 1stFlrSF
2. GarageArea
3. 2ndFlrSF
4. TotRmsAbvGrd
5. ExterQual

Interestingly the number of variables that is non zero in both the models is 34.

Q4) How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Soln.)

A robust and generalizable model will be simple even though if there is a trade off in the accuracy score. The model should predict any unseen dataset with accuracy close to the one it was trained on. Hence in order to make the model more reliable and robust – more investigation into feature engineering must be done. If there is a chance to further eliminate some of the variables and still maintain a good enough predictive power – it should be leveraged. Additional path forward could be better data imputation strategy. The implication on the accuracy would be that it could go down (although not significantly).

Hence even though the model might have a bit high bias – the variance should be minimal. This bias –variance trade-off is the key to a robust and generalizable model