

Problem Statement - Part II

Assignment Part-II

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:

The optimal lambda value in case of Ridge and Lasso is as below:

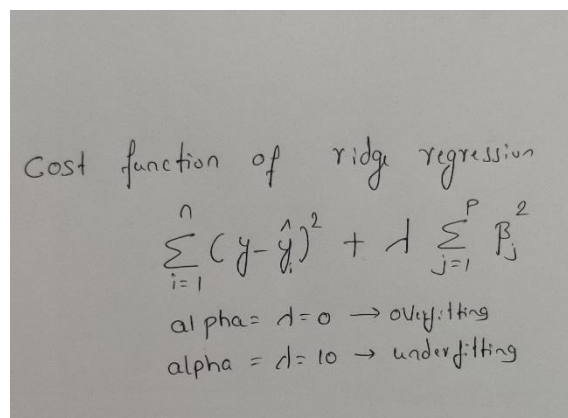
- Ridge - 0.0001
- Lasso - 0.0001

When you double the value of alpha for both ridge and lasso i.e Ridge (alpha=0.002) and lasso(alpha=0.0002)

You can see that as we increase the value of alpha the magnitude of the coefficient decreases.

Ridge

- Double the value from 0.0001 to 0.002
- mean_squared_error will be 0.015904
- After minimizing the features using RFE MSZoning_RL is the important predictor variables after the change is implemented



Cost function of ridge regression

$$\sum_{i=1}^n (y - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

alpha = $\lambda = 0 \rightarrow$ overfitting
alpha = $\lambda = 10 \rightarrow$ underfitting

Lasso

- Double the value from 0.0001 to 0.002
- mean_squared_error will be 0.015799
- after minimizing the features using RFE MSZoning_RL is the important predictor variables after the change is implemented

Cost function of lasso regression

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Mathematics behind lasso is quite similar to that of ridge only difference being instead of adding square of β we will add absolute value of β_j

Mathematics behind lasso regression is quite similar to that of ridge only difference being instead of adding squares of theta, we will add absolute value of beta(j)

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: Outcome of Ridge and Lasso regression

- The optimal lambda value in case of Ridge and Lasso is as below:
 - o Ridge - 0.0001
 - o Lasso - 0.0001
- The Mean Squared error in case of Ridge and Lasso are:
 - o Ridge - 0.015904
 - o Lasso - 0.015843

The MSE - Mean Squared Error of Lasso is slightly lower than that of Ridge.

Lasso has a better edge over Ridge, as it helps feature reduction because one of the coefficient value became 0.

- Based on Lasso the factors that generally affect the price are the following:
 1. Zoning classification - Residential Low Density, Medium Density and Floating Village Residential
 2. Above grade (ground) living area square feet
 3. Overall quality and condition of the house

4. Foundation type of the house - Poured Concrete
5. Number of cars that can be accommodated in the garage
6. Total basement area in square feet and the Basement finished square feet area

Hence the variables predicted by Lasso in the above bar chart as significant variables for predicting the price of a house.

Question 3

After building the model, you realized 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:

Excluding the Five most important predictor variables in the lasso model

- LotFrontage
- MSZoning_RM
- BsmtFullBath
- Neighborhood_Crawfor
- MSZoning_RH

After creating another model by excluding the above-mentioned predictor variables. Following are the new predictor variables. alpha value=0.0001

- LotArea
- Fireplaces
- OverallCond
- BsmtFinSF1
- GrLivArea

Question 4

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?

Answer:

To make model robust and generalizable, make the model 'simple' with less error during on training data.

According to 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 errors on the test data.

When a model performs really well on the data that is used to train it, but does not perform well with unseen data, we know we have a problem: overfitting.

Such a model will perform very well with training data and, hence, will have very low bias; but since it does not perform well with unseen data, it will show high variance.

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.
- In other words, bias in a model is high when it does not perform well on the training data itself, and variance is high when the model does not perform well on the test data.

Regularization helps with managing model complexity by essentially shrinking the model coefficient estimates towards 0. This discourages the model from becoming too complex, thus avoiding the risk of overfitting.

If the model is not robust, it cannot be trusted for predictive analysis.

