

## Problem Statement - Part II

**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?

**Ans:**

For ridge regression, when we plot the curve between negative mean absolute error and alpha we see that as the value of alpha increase from 0 the error term decreases. The train error shows an upward trend when value of alpha increases.

When the value of alpha is 2 the test error is at its minimum, so we decided to go with value of alpha equal to **2** for our ridge regression.

the R2 values for Train and Test (0.9364594823911133, 0.9077597079466581) matches well, indicating an optimum model (RMSE 0.11485785595060997).

For lasso regression we have decided to keep very small value that is **0.01**, when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value 0.

At 0.01, the R2 values for Train and Test (0.8854624158407248, 0.8894603158029368) matches well, indicating an optimum model (RMSE 0.12573595366706636).

The most important predictor variables after the changes have been implemented for ridge regression are mentioned below:

1. MSZoning\_FV
2. MSZoning\_RL
3. Neighborhood\_Crawfor
4. MSZoning\_RH
5. MSZoning\_RM
6. SaleCondition\_Partial
7. Neighborhood\_StoneBr
8. GrLivArea
9. SaleCondition\_Normal
10. Exterior1st\_BrkFace

The most important variable after the changes has been implemented for lasso regression are mentioned below:-

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontage

**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?

**Ans:**

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

Ridge regression uses a tuning parameter called lambda as the penalty. Residual sum of squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values get 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. As the lambda value increases Lasso shrinks the coefficient towards zero and it makes the variables exactly equal to 0. Lasso also does feature 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.

**In this case, ridge regression will be preferred over lasso.**

Ridge regression (lambda = 2):

the R<sup>2</sup> values for Train and Test are 0.9364594823911133 and 0.9077597079466581 while RMSE is 0.11485785595060997.

Lasso regression (lambda = 0.01):

the R<sup>2</sup> values for Train and Test are 0.8854624158407248 and 0.8894603158029368 while RMSE is 0.12573595366706636.

**Clearly, Ridge is better** as the R<sup>2</sup> is higher.

**Q3.** 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?

Those 5 most important predictor variables that will be excluded are :-

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. GarageArea

**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?

**Ans:**

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.

**Bias:** Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

**Variance:** Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.

It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.