Questions & Answers

Part II : Assignment Questions

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:

To obtain the optimal value of alpha for both ridge and lasso regression, a set of parameters was set. The parameters were predefined by using the numpy library logspace function and this function returns numbers spaced evenly on a log scale. The number of hyper parameters generated was 200.

K-fold cross validation known as **GridSearchCv** was then applied on the 200 parameters. This method will search through the best parameter values from the given set and the k-fold was set to 5. So 5 folds were fit for each of 200 candidates, which resulted in 1000 fits.

After performing the above, the optimal value of alpha for ridge regression was 15.70 while that of lasso was 0.01.

If we double the value of alpha for both ridge and lasso, the quantity of bias being introduced into the algorithm's output will also double. In other words, the impact of the shrinkage penalty will double.

Below is an illustration of the difference in terms of r-squared, RSS and mean squared error.

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.891304	0.844459
1	R2 Score (Test)	0.815782	0.816528
2	RSS (Train)	106.413698	152.274674
3	RSS (Test)	85.636591	85.289553
4	MSE (Train)	0.329691	0.394387
5	MSE (Test)	0.451549	0.450633

Figure 1.1: Original alpha value

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.849508	0.940894
1	R2 Score (Test)	0.811331	0.542816
2	RSS (Train)	147.331967	57.864646
3	RSS (Test)	87.705431	212.528450
4	MSE (Train)	0.150492	0.059106
5	MSE (Test)	0.208822	0.506020

Figure 1.2: Double alpha value

For ridge regression, the r2 score for the training set has dropped when the alpha value has been doubled when compared to the original alpha value. It can be said that the model has generalized more and complexity has lessened because the difference between the train and test set r2 score is lesser with doubled alpha value.

However the Residual Sum of Squares in the training set has increased more due to more generalization. There is also a slight increase in terms of RSS with respect to the test set when the alpha has been doubled.

Additionally, the Mean Squared Error has dropped significantly for both train and test set with doubled alpha value. This is because the model is able to do proper prediction on unseen data.

While for Lasso regression, we can notice that the model performance has actually dropped with doubled value of alpha. This is due to more features being assigned a coefficient of 0. The latter has removed 198 features and only considered 48.

This explains the major gap between the r2 value of train and test set. The RSS as well as the MSE for the train set is relatively low, however for the test they are very high because the model is unable to predict on unseen data.

The five most important predictor variables for ridge will be (with their coefficient) :

- OverallQual : 0.161681
- GrLivArea : 0.101319
- KitchenQual : 0.088121
- ExterQual : 0.083546
- GarageArea : 0.077637

The five most important predictor variables for Lasso will be (with their coefficient) :

- RoofMatl_CompShg : 5.743132
- RoofMatl_Tar&Grv : 5.741194
- RoofMatl_WdShake : 5.569610
- RoofMatl_WdShngl : 5.553336
- RoofMatl_Roll : 5.45833

To conclude, the accuracy of the ridge model on the test set when the value of alpha has been doubled is better when compared to that of 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:

I will choose to apply lasso regression for this case study because the model is more generalized and the r-squared difference between the train set and test set is lesser when compared to that of ridge, it can be seen from the diagram below. This was a result of lasso performing feature selection, the lasso penalty has forced some the coefficient to 0 and thus eliminating certain features from being considered. Our model had a total of 241 features and lasso is considering 43 of them. The Lasso regression model has favored subsets of features that have less collinearity therefore making the model more accurate and interpretable.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.430428e-01	0.891304	0.844459
1	R2 Score (Test)	-1.860349e+21	0.815782	0.816528
2	RSS (Train)	5.576113e+01	106.413698	152.274674
3	RSS (Test)	8.648102e+23	85.636591	85.289553
4	MSE (Train)	2.386571e-01	0.329691	0.394387
5	MSE (Test)	4.537700e+10	0.451549	0.450633

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:

After building the new lasso model, the five most important predictor variables is listed below together with their respective coefficients :

- HeatingQC : 0.259237
- 1stFlrSF : 0.251741
- MasVnrArea : 0.180830
- KitchenQual : 0.151136
- Neighborhood_Mitchel : 0.131792

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:

Model robustness refers to the effectiveness of the algorithm when being tested on new and independent dataset and generalization refers to the model's ability to react to new data. To ensure that a model is both robust and generalizable, we have to address the issue of overfitting and underfitting. Overfitting is when the model fits exactly against training data while underfitting is when the model is unable to capture the trend of the data.

Regularization is one to ensure that a model is robust and generalisable. Regularization is used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting. The important step is regularization is choosing an optimal lambda which can be achieved by using cross validation.

Another important step to make sure that a model is robust is to divide the data set into training, validation and test set. The validation set helps to improve the model performance by fine-tuning the model after each epoch. The test set helps to check for model robustness and accuracy. Additionally, outliers have to be tackled to increase a model robustness

A robust and generalized model will compromise on the training accuracy but improves the validation and testing accuracy, this is because the model has not memorized the data patterns in the training dataset and successfully generalizes unseen data.

Moreover, a robust and generalized model strikes a nice balance between accuracy and interpretability.