Advanced Regression 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 value of alpha for Ridge is 200 whereas for Lasso is 0.001 with scores as below:

	Ridge(alpha=200)	Lasso(alpha=0.001)
R-Squared (Train)	0.938726	0.944033
R-Squared (Test)	0.919145	0.914382
RSS (Train)	7.760769	7.088602
RSS (Test)	4.411701	4.671624
MSE (Train)	0.007601	0.006943
MSE (Test)	0.010072	0.010666
RMSE (Train)	0.087185	0.083323
RMSE (Test)	0.100361	0.103275

The top 5 predictors for Ridge and their coefficients are as below:

- OverallQual_Excellent --> 0.020271
- OverallQual_Very Good --> 0.024191
- 1stFlrSF --> 0.027782
- TotalBsmtSF --> 0.028167
- GrLivArea --> 0.044738

The top 5 predictors for Lasso and their coefficients are as below:

- RoofMatl_WdShngl --> 2.875691e-02
- OverallQual_Very Good --> 3.267919e-02
- RoofMatl_CompShg --> 3.686069e-02
- TotalBsmtSF --> 4.365792e-02
- GrLivArea --> 1.113557e-01

By choosing double value for alpha for Ridge alpha=400 and for Lasso alpha=0.002 the scores are as below:

	Ridge(alpha=200)	Ridge2(alpha=400)	Lasso(alpha=0.001)	Lasso2(alpha=0.002)
R-Squared (Train)	0.938726	0.931748	0.944033	0.937398
R-Squared (Test)	0.919145	0.918906	0.914382	0.923050
RSS (Train)	7.760769	8.644639	7.088602	7.929026
RSS (Test)	4.411701	4.424768	4.671624	4.198625
MSE (Train)	0.007601	0.008467	0.006943	0.007766
MSE (Test)	0.010072	0.010102	0.010666	0.009586
RMSE (Train)	0.087185	0.092015	0.083323	0.088125
RMSE (Test)	0.100361	0.100510	0.103275	0.09

We can see a drop in both train and test score for Ridge whereas for Lasso we can see a drop in train scores, but test scores are better. This can be explained as Ridge keeps all the predictors moving the coefficients closer to 0 but not making them absolute 0 but Lasso for some predictors makes the coefficient absolute 0.

The top 5 predictors and their coefficients for Ridge with alpha=400 are as below:

- OverallQual_Below Average --> -0.015389
- YearRemodAdd_Age --> -0.015380
- Neighborhood_MeadowV --> -0.013645
- OverallCond_Below Average --> -0.013055
- OverallQual_Average --> -0.012634

The top 5 predictors and their coefficients for Lasso with alpha=0.002 are as below:

- BsmtFinSF1 --> 2.571180e-02
- OverallQual_Excellent --> 2.634795e-02
- OverallQual_Very Good --> 3.345817e-02
- TotalBsmtSF --> 3.870722e-02
- GrLivArea --> 1.101390e-01

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:

The model scores are almost similar

	Ridge(alpha=200)	Lasso(alpha=0.001)
R-Squared (Train)	0.938726	0.944033
R-Squared (Test)	0.919145	0.914382
RSS (Train)	7.760769	7.088602
RSS (Test)	4.411701	4.671624
MSE (Train)	0.007601	0.006943
MSE (Test)	0.010072	0.010666
RMSE (Train)	0.087185	0.083323
RMSE (Test)	0.100361	0.103275

so the model we will choose to apply will depend on the use case.

- If we have too many variables and our primary goal is feature selection, then we will use Lasso.
- If we don't want to get too large coefficients and want to keep all the predictors along with reduction of coefficient magnitude is one of our primary goals, then we will use Ridge Regression

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:

The top 5 predictors and their coeficients in lasso models are:

- RoofMatl_WdShngl --> 2.875691e-02
- OverallQual_Very Good --> 3.267919e-02
- RoofMatl_CompShg --> 3.686069e-02
- TotalBsmtSF --> 4.365792e-02
- GrLivArea --> 1.113557e-01

Removing these and rebuilding the model gives us following findings:

	Lasso(alpha=0.001)	Lasso(alpha=0.001,without top 5)
R-Squared (Train)	0.944033	0.935253
R-Squared (Test)	0.914382	0.912402
RSS (Train)	7.088602	8.200712
RSS (Test)	4.671624	4.779618
MSE (Train)	0.006943	0.008032
MSE (Test)	0.010666	0.010912
RMSE (Train)	0.083323	0.089622
RMSE (Test)	0.103275	0.104462

There is a slight drop in both train and test scores.

The new top 5 predictors and their coefficients as below:

- LotArea --> 0.023137
- Neighborhood_NridgHt --> 0.026026
- BsmtFinSF1 --> 0.043300
- 2ndFlrSF --> 0.087073
- 1stFlrSF --> 0.094959

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:

A robust model is one where any variation in the data does not affect its performance much.

A generalizable model successfully adapts to data sets other than the one used for training and testing.

To make sure a model is robust and generalizable, we must make sure model doesn't overfit on training data as overfitting model has very high variance and change in data degrades the accuracy of the model. An overfitted model will identify all the patterns of a training data but fail to pick up the patterns in unseen test data.

In general, balance between model accuracy and complexity is required. This can be achieved by Regularization techniques like Ridge Regression and Lasso. Also using cross validation methods like K-Fold Cross validation with GridSearch cross validation helps to find the best fit hyperparameters for Regularized Regression models.