Mickell Als
Date: 2023/04/15

Subjective Questions for Regularisation

Question 1

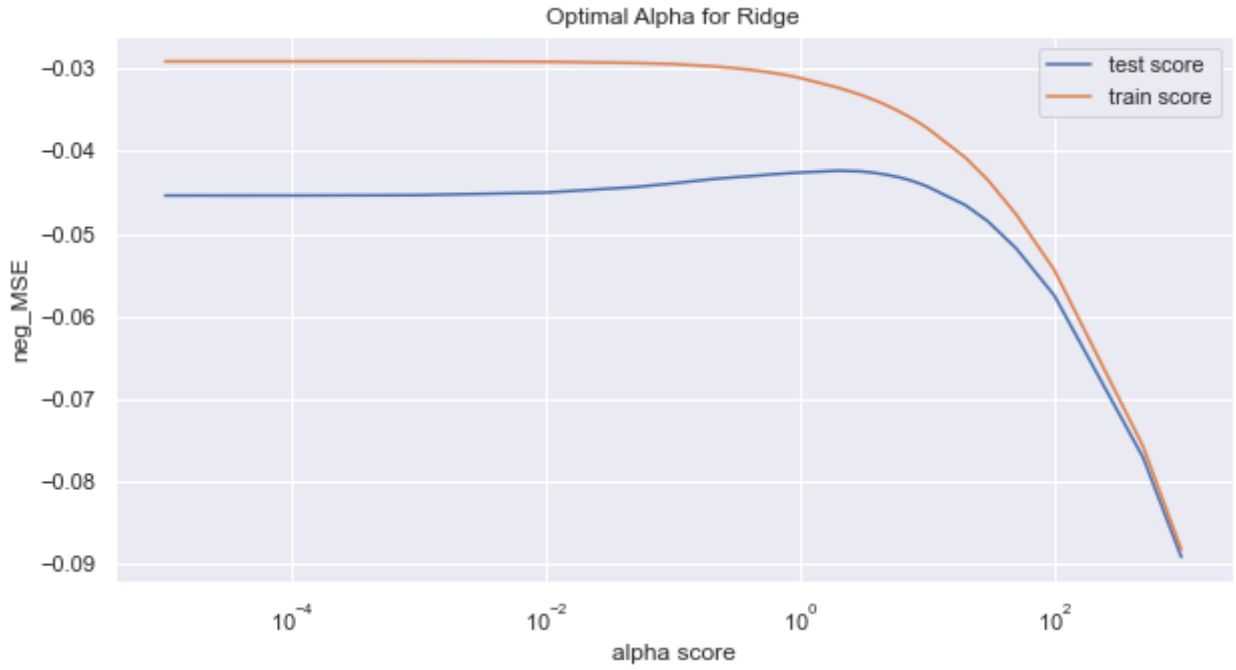
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Question 1 Response

Optimal Values for Alpha

Ridge Regression	2.0
Lasso Regression	0.0001

In doubling the value of alpha to 4.0 for ridge regression, the model's performance did not fall significantly. And in fact this can be expected, in the cross validation conducted a list of possible values for alpha were selected and the best from that list was chosen (2.0). However, if we plot each value of alpha out and their performance on the test and training sets using cross validation we see that there is actually a narrow window of values that produce similar results. Therefore, I submit that as long as the values of alpha stay within a set interval then performance of the model will not significantly change. This can be seen from the graph below.



This can also be seen with Lasso's doubled alpha, a value of 0.0002. The predictive power of the model remained stabled around and r-squared value of 0.907.

However, while both models performed similarly, there were small changes to the values of the beta coefficients, the intercept, and the most important features.

Mickell Als
Date: 2023/04/15

For Ridge regression the changes were observed to the top predictor features

Initially, these were the top predictors and their coefficient values; [***GrLivArea*** *0.235465,* ***TotalBsmtSF*** *0.156934,* ***OverallQual-9*** *0.115909,* ***HomeAge*** *-0.089490,* ***BsmtFullBath-2*** *0.074884*]

These were updated and changed to [***GrLivArea*** *0.195815,* ***TotalBsmtSF*** *0.146191,* ***OverallQual-9*** *0.103882,* ***GarageArea*** *0.072098,* ***LotArea*** *0.070077*]. Note that GarageArea and LotArea became more important features than the HomeAge and BsmtFullBath-2.

For Lasso regression the changes were

From - [***GrLivArea*** *0.327264,* ***OverallQual-9*** *0.179723,* ***OverallQual-10*** *0.156059,* ***TotalBsmtSF*** *0.149738,* ***HomeAge*** *-0.138712*]

To - [***GrLivArea*** *0.316012,* ***OverallQual-9*** *0.179589,* ***TotalBsmtSF*** *0.150049,* ***HomeAge*** *-0.114761,* ***OverallQual-10*** *0.107628*]

The set of predictors remained the same, but their order was changed with OverallQual-10 dropping to 5th from 3rd.

If we compare the Ridge and Lasso predictors we also see some overlap in the top predictors but also a few differences BUT even with the differences, all four versions (2-Ridge and 2-Lasso) still performed similarly indicating that their effects on the response hold similar weight BUT not significantly so as to alter model performance negatively.

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?

Question 2 Response

I would opt for a Lasso regression with a lambda value of 0.0001. I opt for this form of regularization because it sufficiently answers the business questions of which predictors are most significant and described the unseen test data just as well as the Ridge model. While both models performed similarly, the business wants to know the best predictors for housing price in Australia and Ridge regression only shrinks coefficients to near 0 but almost never 0 meaning that while noising variables may be minimised, it may still make it hard to interpret the model. Lasso however, shrinks less important feature coefficients to 0 therefore that feature would have no influence on the response variable. With these fewer features, we can more easily understand the model and more readily use it as we only need to collect data on the features deemed important.

Therefore, Lasso regression provided a simpler model that was just as robust and generalizable as the Ridge model thanks to its fewer predictors variables and similar performance to the Ridge model.

Mickell Als
Date: 2023/04/15

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?

Question 3 Response

If the model is rebuilt without the initial top 5 features, we lose some predictive power but it still remains high. The features that are deemed significant are all binary categorical fields with negative coefficients meaning their act significantly to lower the sale price. These features are:

Feature	Coefficient
OverallQual-3	-0.160400
OverallQual-2	-0.159466
OverallQual-4	-0.150425
OverallQual-5	-0.146578
OverallQual-6	-0.131202

r2 train: 0.9094561725498274

r2 test: 0.8701825372917817

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?

Question 4 Response

To make a robust and generalizable model, we can use techniques such as Regularization and Cross-validation. In regularization, the cost is not solely determined by the cost function BUT also the penalty term added to it. Therefore, if the penalty term has too many values supplied to it, it added to the cost. For example, in Lasso regression, we seek to shrink coefficients to 0, therefore their impact on the response is eliminated. The features that remain are those for which the MSE and penalty remained lowest. We then have a model which has identified the predictors deemed important to predicting the response. This gives us a model that is not more complex than it needs to be while retaining the valuable features and dropping noisy ones. Another method is cross validation, here we split our training set into two, a training and validation set. We then train the model on the new training set and validate it on the validation set. This is repeated on the same training set many times, this allows us to not have just a single dataset but many and test the model on multiple. After we have validated the model on many of these sets, we can get an idea of model stability and performance. If the model performed very well on the full training set but poorly in cross validation then that could mean we have an overfitting or underfitting problem and therefore the model is not very stable and will likely perform poorly in the real world. Cross validation is a great way to check how well your model may perform in the real world. Cross validation can also be used in conjunction with other methods to improve model robustness and generalisability as it can be used to select the hyperparameters that provided the best performance, essentially removing the guess work out of hyperparameter tuning.

Mickell Als
Date: 2023/04/15

So, while a robust and general model may have a lower training accuracy, because it has identified the underlying pattern in the data, the test accuracy will also be similar.