ADVANCED REGRESSION ASSIGNMENT

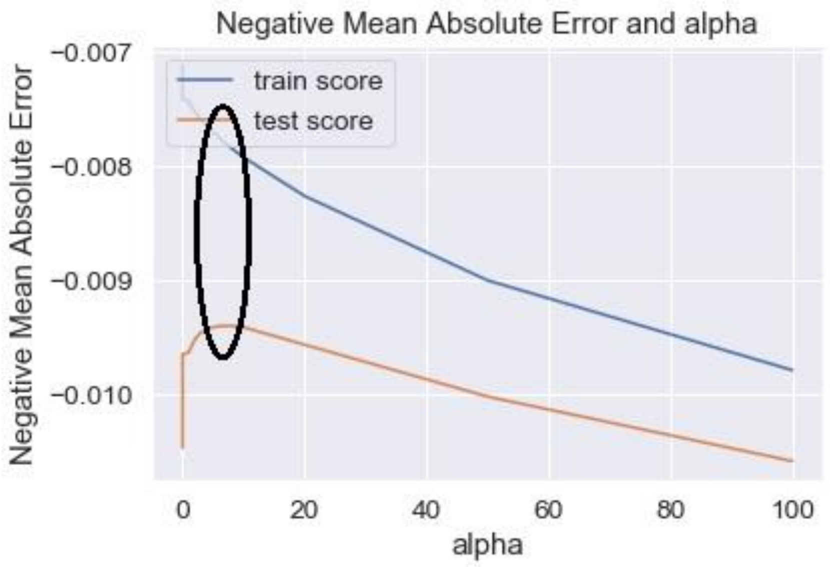
Question 1

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?

OPTIMAL VALUES OF ALPHA FOR RIDGE AND LASSO REGRESSION-

- The optimal value for alpha is the point in the parameter space, when we have less train score(R2) value and but not more test score(R2) value, The plot defines the theoretical way of deciding the optimal value of alpha.
- First, the training error will be more, and the test also will be more. As we add the more parameters to the model, it starts to predict well on the train data due to which the plot is score is decreasing as we add more coefficients.
- After some point, the training error will be less, but the test error will be quite large. This is a start point for this underfit. Hence we shouldn’t attain this state, so we use a lasso and ridge regression. In both the case deciding of optimum value is same.

- This is a kind of trial and error iterative method where we find out a negative mean square for each test and train dataset, Then we plot alpha against both train and test scores as shown in the plot. The very starting point at which we feel we have less distance between test and train score is the point of having optimum value for alpha.



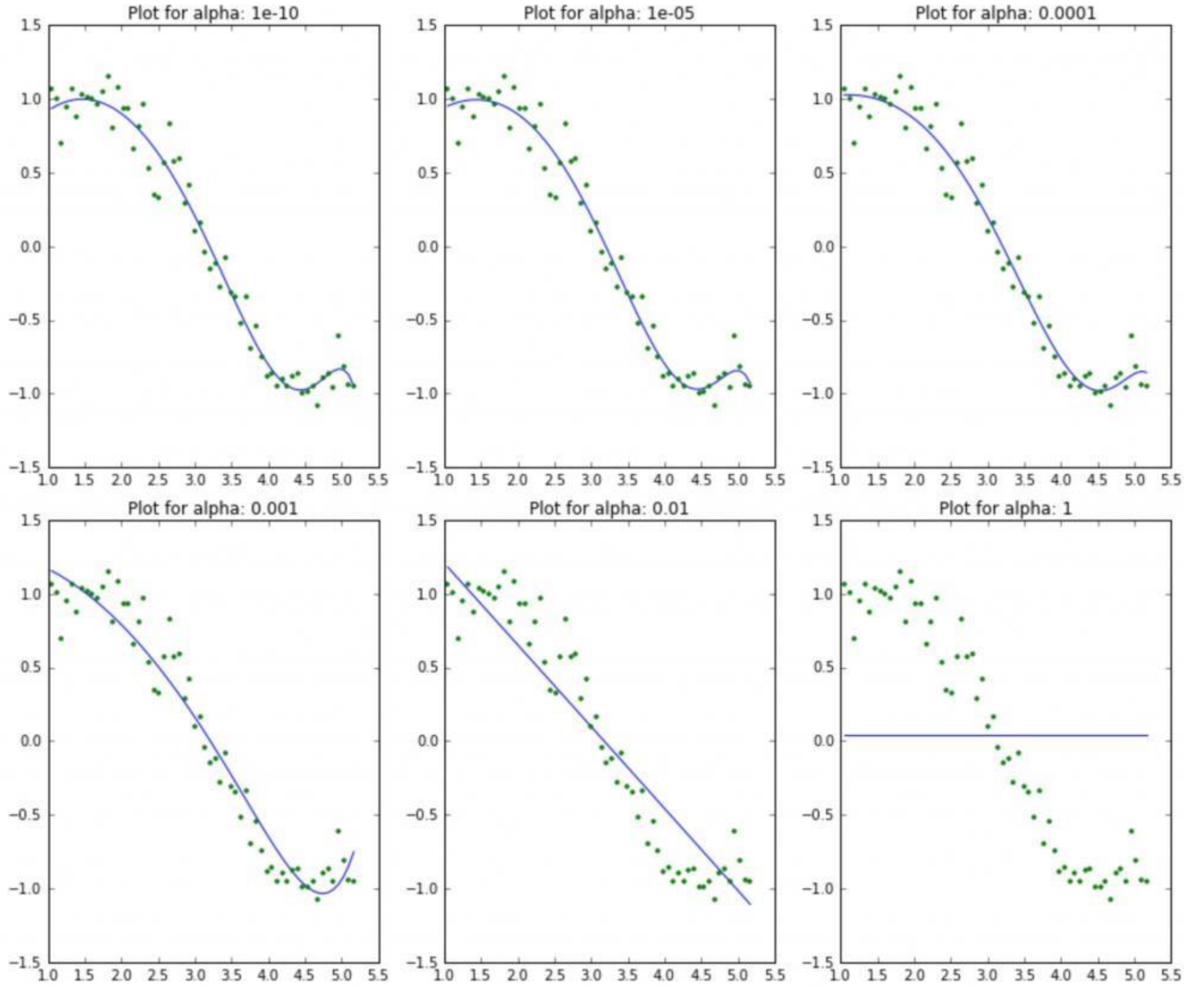
In the above plot, as we can see the range of 0 - 20 we can see they been having optimum value for alpha. The point which has highlighted with black color is the point of optimum value which we have to consider in both the cases of ridge and lasso.

THERE ARE CHANGES IN THE MODEL IF YOU CHOOSE DOUBLE THE VALUE OF ALPHA FOR BOTH RIDGE AND LASSO RERESSION

Various alpha values:

1. α = 0:
   - The objective becomes the same as the simple linear regression.
   - We’ll get the same coefficients as simple linear regression.
2. α = ∞:
   - The coefficients will be zero, Because of infinite weightage on the square of coefficients, anything less than zero will make the objective infinite.
3. 0 < α < ∞:
   - The magnitude of α will decide the weightage given to different parts of the objective.
   - The coefficients will be somewhere between 0 and 1 for simple linear regression.

From the above definitions, we can say that as we double the value of alpha, the coefficients become zero. The variables dependency on model decreases to a large extend and equation will become equal to intercept.



The above plot shows a condition for increasing value of alpha, at first we had taken a very low value for alpha, the model complexity is more. But as we increase alpha the model will become simpler and it would reduce the possibility of overfit. This again tells us that the model complexity decreases with increase in the values of alpha. But notice the straight line at alpha=1.

THE MOST IMPORTANT PREDICTOR VARIABLES AFTER THE CHANGE IS IMPLEMENTED?

- Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression.

- Ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity

$$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} w_j^2 \tag{1.3}$$

- when λ → 0, the cost function becomes similar to the linear regression cost function

- Hence lower the constraint (low λ) on the features, the model will resemble linear regression model.

- As it shrinks the coefficients. The variables which has more coefficient value are the one which are important predictors of the resultant.

- **Lasso regression** not only helps in reducing over-fitting but it can help us in feature selection

$$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} |w_j| \tag{1.4}$$

- When λ → 0, the cost function becomes similar to the linear regression cost function by making coefficients to zero.

- Hence lower the constraint (low λ) on the features, the model will resemble linear regression model.

- With the regularization, the model will perform the feature selection also, it will make the coefficients zero, if the variable has no dependency on the resultant. If there are 200 columns, in ridge regression the coefficients will not be zero for those which has less dependency on variables. But when we do a lasso regression, it will make coefficients as zero if they have no dependency on resultant.

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?

Ridge Regression

- Removing predictors from the model can be seen as settings their coefficients to zero. Instead of forcing them to be exactly zero, let's penalize them if they are too far from zero, thus enforcing them to be small in a continuous way. This way, we decrease model complexity while keeping all variables in the model.

- We use this when we don’t want to feature selection in the data that is when we don’t want to remove the dependency of those variables on the model, we use the ridge regression.

-we not only minimize the sum of squared residuals but also penalize the size of parameter estimates.

Lasso Regression

- Lasso, or Least Absolute Shrinkage and Selection Operator, is quite similar conceptually to ridge regression.

- It also adds a penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squared coefficients (the so called L2 penalty), lasso penalizes the sum of their absolute values (L1 penalty)

-Lasso can set some coefficients to zero, thus performing variable selection, while ridge regression cannot.

Both methods allow to use of correlated predictors, but they solve multicollinearity issue differently:

- In ridge regression, the coefficients of correlated predictors are similar;
- In lasso, one of the correlated predictors has a larger coefficient, while the rest are (nearly) zeroed.

Difference between Lasso and Ridge

- Lasso tends to do well if there are a small number of significant parameters and the others are close to zero (ergo: when only a few predictors influence the response).
- Ridge works well if there are many large parameters of about the same value (ergo: when most predictors impact the response).

Question 3

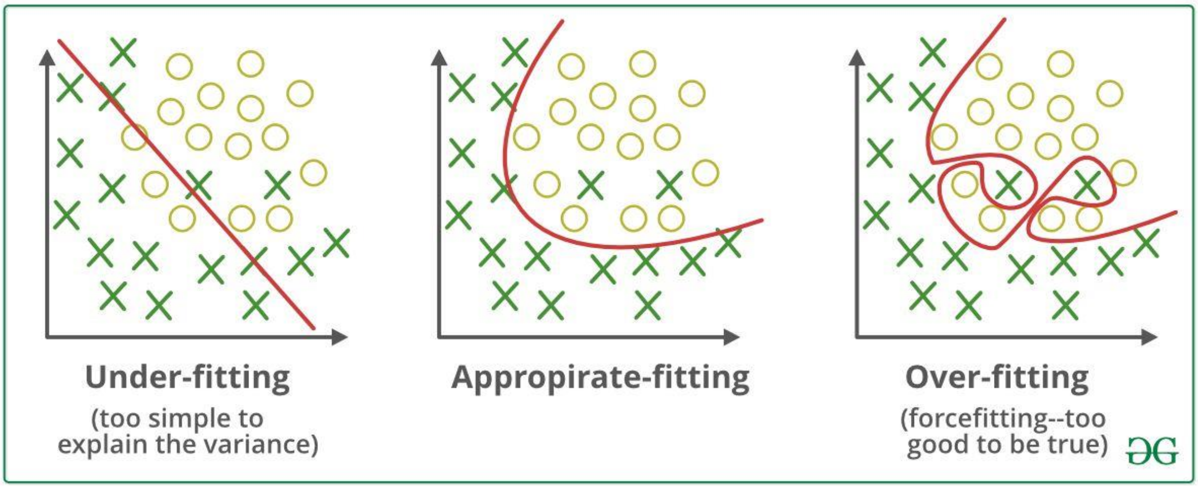
After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

- The question of important predictor variables is the one that has a high dependency on the model.
- If top predictors refer to the one which has high coefficiency in a positive and low negative value, which means they have a dependency on the final model.
- If the high dependency variables are categorical, then we create a dummy variable. If previously shared data had this categorical/dummy variable, if it doesn't have on present data then, we can remove that categorical /dummy column with that model (If we had that column previously, now if we don't have that column, then we can just simply drop that column.)
- In the above question, we are referring top 5 as the predictors which have high absolute coefficients (includes both positive and negative coefficiency) in the model.
- If we obtained them in original data, but when we are testing if we don't get that column, it is better to remove it, instead of populating that value with wrong information into data.

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?

- The robust and generalize is a point in which the model is not either overfitted or underfitting.


Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

- By looking at the graph on the left side we can predict that the line does not cover all the points shown in the graph. Such a model tends to cause the underfitting of data. It also called High Bias.

- Whereas the graph on the right side shows the predicted line covers all the points in the graph. In such a condition, you can also think that it’s a good graph which covers all the points. But that’s not true, the predicted line into the graph covers all points which are noise and outlier. Such a model is also responsible to predict poor results due to its complexity. It is also called High Variance.

- Now, looking at the middle graph it shows a pretty good predicted line. It covers the majority of the point in the graph and also maintains the balance between bias and variance.

- To attain the middle stage, we have to see look into the model r2 value for both the train and test dataset.

- Ridge and Lasso's regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression. As we have to decide on alpha to perform both the regularizations.

- Ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} w_j^2 \qquad (1.3)$$

- Lasso regression not only helps in reducing over-fitting but it can help us in feature selection

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} |w_j| \qquad (1.4)$$

Various lambda (α) values:

1. α = 0:
   - The objective becomes the same as the simple linear regression.
   - We'll get the same coefficients as simple linear regression.
2. α = ∞:
   - The coefficients will be zero, Because of infinite weightage on the square of coefficients, anything less than zero will make the objective infinite.
3. 0 < α < ∞:
   - The magnitude of α will decide the weightage given to different parts of the objective.
   - The coefficients will be somewhere between 0 and 1 for simple linear regression.

Using both methods we can prevent it from overfitting.

Below defines the three scenarios of the accuracy of the model how it will be affected as by model:

- If train r2 is 0.90 and test r2 value is 0.88 – Good Model
- If train r2 is 0.99 and test r2 value is 0.70 – Overfitted
- If train r2 is 0.5o and test r2 value is 0.44 – Underfitted

as we can see from the above r2 scores, the model must be a Good Predictor. It shouldn’t overfit or underfit the model.