Assignment: Advance Regression

Question 1:

Rahul built a logistic regression model with a training accuracy of 97% and a test accuracy of 48%. What could be the reason for the gap between the test and train accuracies, and how can this problem be solved?

Answer:

The reason for such a vast gap in test-train accuracy and extremely poor accuracy in the test set while performing well in train set is due to **overfitting of model**. The model has learnt the train dataset very well. Hence, it's performing quite well with training data. However, the model has captured noise with the underlying pattern of the data. The model has low bias and high variance. Hence, the model has not generalized well from our training data to unseen data.

To get rid of the overfitting, we can do following

1. **Cross Validation**: We can use cross-validation to prevent overfitting. Cross-Validation uses initial training dataset to create multiple folds of smaller train-test dataset splits. It allows us for hyper parameter tuning using training set only and test set remain unseen for final model selection.

2. **Bigger Training Dataset:** Sometimes giving more data for training the model helps algorithm understand the pattern clearly.

3. **Regularization:** By adding a penalty to the cost function.

4. **Removing Features:** This can be achieved manually or selecting model with built-in feature selection.

Question 2

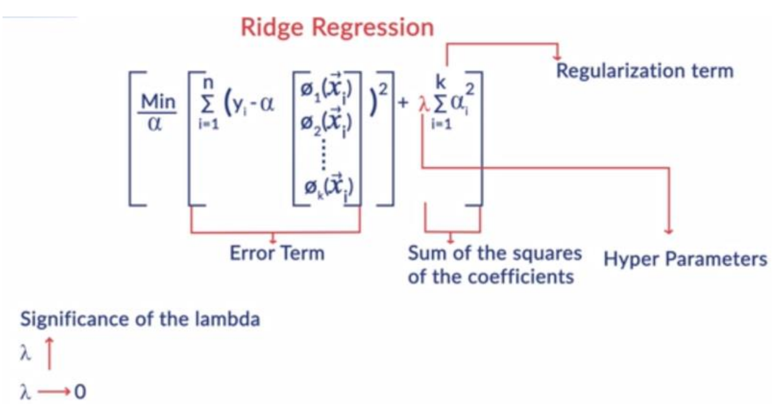
List at least four differences in detail between L1 and L2 regularization in regression.

Answer:

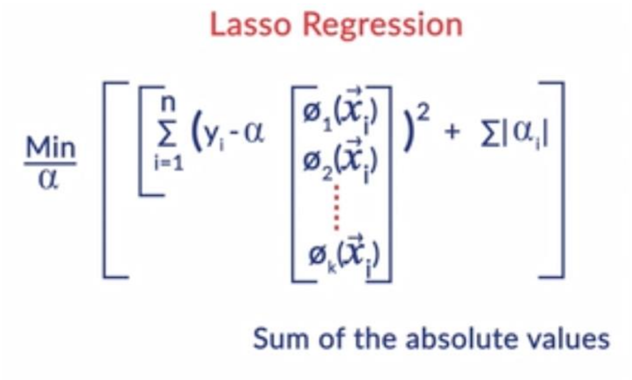
The model using L1 regularization is Lasso Regression and it adds "absolute value of magnitude" of coefficients as penalty term to the loss function. Also, the model using L2 regularization is Ridge Regression and it adds squared magnitude of the coefficients as penalty term to the loss function.

Both Ridge and Lasso methods basically regularizes the coefficients by reducing them in value, causing shrinkage of the coefficients.

<u>Loss Function Using Ridge Regression:</u>



Loss Function Using Lasso Regression:



Both perform different measures of shrinkage which depends on the value of hyper-parameter λ

1. **Feature Selection**: Lasso shrinks the coefficients for less important features to zero hence removes some features altogether. Due to this property Lasso regression (L1 regularization) can be used for feature selection in case of large number of feature. This is not possible with Ridge Regression (L2 regularization).
2. **Sparsity**: L1 regularization can produce very small feature with large coefficient and most of the feature with zero coefficient. Hence it creates sparse matrix compared to L2 regularization.
3. **Efficiency**: L2 Regularization computation is more efficient compared to L1 regularization.
4. **Robustness:** Since in L2 regression squares the error, there is much larger impact of error on the model than in L1 regression. This L2 regression is more sensitive to the outliers compared to L1 regression.

Question 3

Consider two linear models:

L1: y = 39.76x + 32.648628

And

L2: y = 43.2x + 19.8

Given the fact that both the models perform equally well on the test data set, which one would you prefer and why?

Answer:

We would consider model L2 considering the fact that the model is simple and since the coefficients uses less bits of data it less complex compared to L1.

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?

Answer:

We can make the model more robust and generalizable by making it simpler with less complexities. To reduce the complexity of the model we must ensure that the model is neither under-fitted not over fitted.

Generic strategy to make the model more robust and generalizable by Hold Out Strategy or Cross Validation. the basic idea is to keep aside some data that will not in any way influence the model building

While creating the model robust & generalizable we choose the model which gives the least test error score.

To achieve this following strategy can be applied:

1. Using metrics like Mallows Cp, BIC, AIC or adjusted R Squared which take into account both simplicity and model fit by penalizing the model for being too complex(overfit), hence more representative of unseen 'test error'.
2. Estimating the error via a validation set or cross-validation approach.

Question 5

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:

It depends on the number of features used in the model. In case of Lasso Regression does both parameter shrinkage and variable selection automatically. Since, Lasso does the feature elimination also in case of large number of features, LASSO would be the preferred choice. However, Ridge is faster to implement hence, in case of small size features we will prefer Ridge regression.