ASSIGNMENT PART-II

(Submitted by- Darshna)

Question-1:

Rahul built a logistic regression model having a training accuracy of 97% while the test accuracy was 48%. What could be the reason for the seeming gulf between test and train accuracy and how can this problem be solved.

Answer:

The reason for the seeming gulf between test and train accuracy is overfitting where a model becomes too specific to the data it is trained on and fails to generalise to other unseen data points in the larger domain. A model that has become too specific to a training dataset has actually not just the hidden patterns in the data but also the noise and the inconsistencies in the data. In a typical case of overfitting, the model performs very well on the training data but fails miserably on the test data.

This problem can be solved using below methods:-

1. Cross validation:- Cross-validation is a powerful preventative measure against overfitting. Idea used in cross validation is to use your initial training data to generate multiple small train-test splits. Use these splits to tune your model. Cross-validation allows us to tune hyperparameters with only our original training set. This allows us to keep our test set as a truly unseen dataset for selecting our final model.

2. Regularization:-Regularization is a technique for penalizing large coefficients in order to avoid overfitting, and the strength of the penalty should be tuned. Regularization is a process used to create an optimally complex model, i.e. a model which is as simple as possible while performing well on the training data. Through regularization, one tries to strike the delicate balance between keeping the model simple, yet not making it too naive to be of any use. Ridge regression and Lasso regression. Both these methods are used to make the regression model simpler while balancing the 'bias-variance' tradeoff. In ridge regression, an additional term of "sum of the squares of the coefficients" is added to the cost function along with the error term, whereas in case of lasso regression, a regularisation term of "sum of the absolute value of the coefficients" is added.

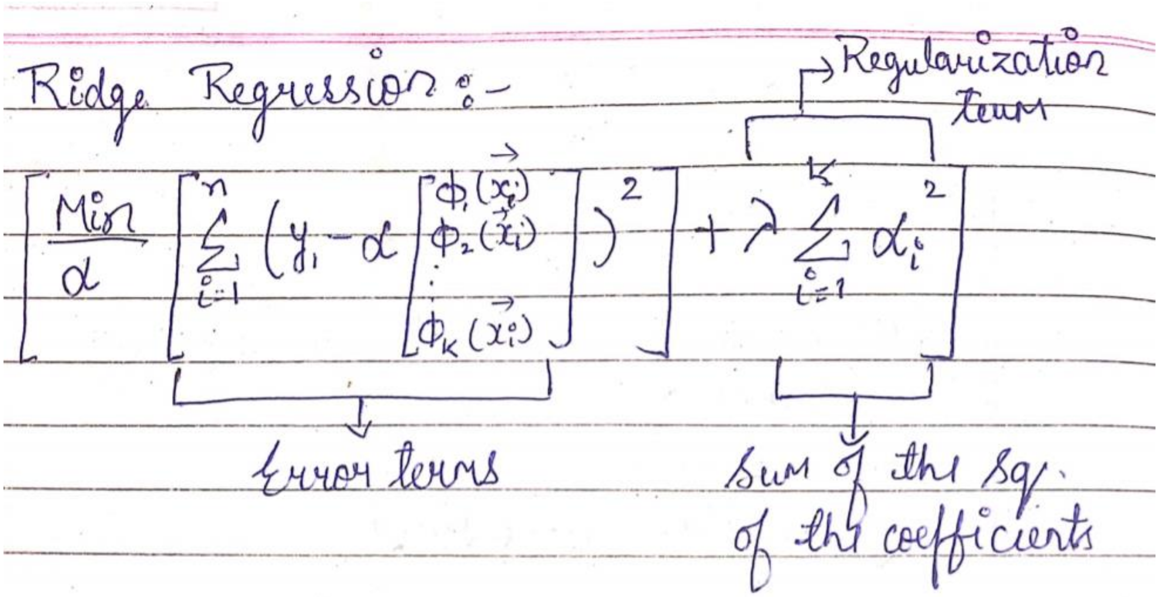
Question-2:

List at least 4 differences in detail between L1 and L2 regularization in regression.

Answer:

i. A regression model that uses L1 regularization technique is called Lasso Regression and model which uses L2 is called Ridge Regression.

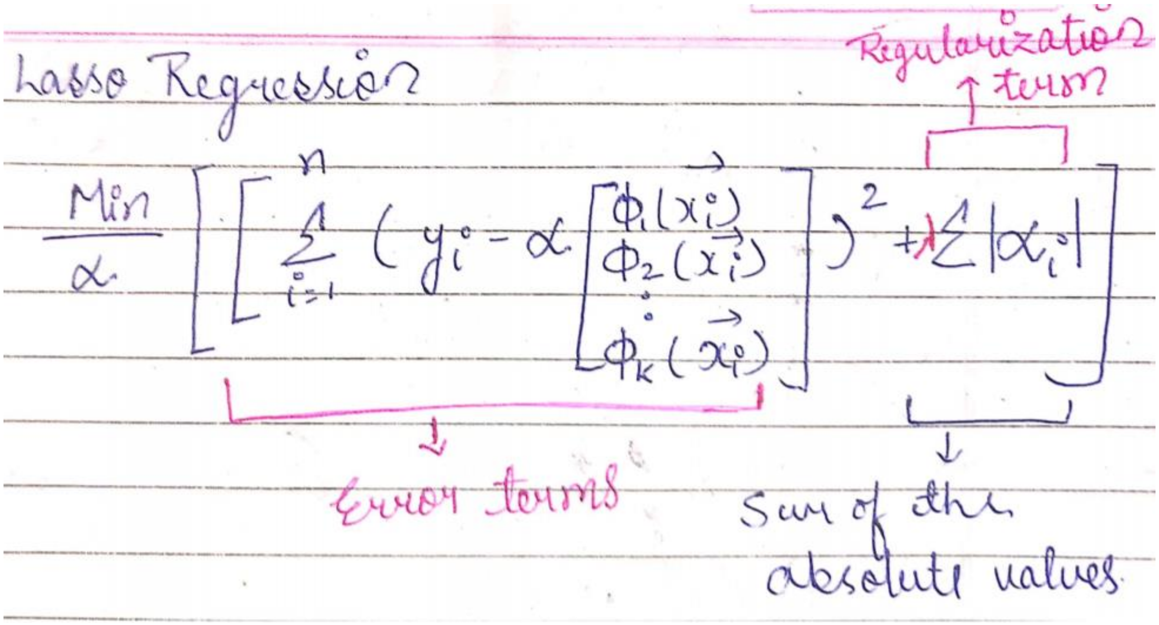
ii. In ridge regression, an additional term of "sum of the squares of the coefficients" is added to the cost function along with the error term, whereas in case of lasso regression, a regularisation term of "sum of the absolute value of the coefficients" is added.



Ridge Regression :-

$$\frac{\text{Min}}{\alpha}\left[\underbrace{\sum_{i=1}^{n}\left(y_1-\alpha\begin{bmatrix}\phi_1(\vec{x_i})\\ \phi_2(\vec{x_i})\\ \vdots\\ \phi_k(\vec{x_i})\end{bmatrix}\right)^2}_{\text{Error terms}}+\overbrace{\lambda\underbrace{\sum_{i=1}^{k}\alpha_i^2}_{\text{Sum of the sq. of the coefficients}}}^{\text{Regularization term}}\right]$$

And



Lasso Regression

$$\frac{\text{Min}}{\alpha}\left[\underbrace{\left[\sum_{i=1}^{n}\left(y_i-\alpha\begin{bmatrix}\phi_1(\vec{x_i})\\ \phi_2(\vec{x_i})\\ \vdots\\ \phi_k(\vec{x_i})\end{bmatrix}\right]\right)^2}_{\text{Error terms}}+\lambda\overbrace{\underbrace{\sum|\alpha_i|}_{\text{Sum of the absolute values.}}}^{\text{Regularization term}}\right]$$

iii. Ridge is computationally less intensive than Lasso. Ridge regression almost always has a matrix representation for the solution while Lasso requires iterations to get to the final solution.

iv. Lasso trims down the coefficients of redundant variables to zero, and thus indirectly performs variable selection also. Ridge, on the other hand, reduces the coefficients to arbitrarily low values, though not zero.

v. Lasso has reduced complexity by variable selection and the error is also comparable to Ridge regression. Thus, it has lesser complexity and will perform better than Ridge Regression.

vi. Lasso regression is that it results in model parameters such that the lesser important features' coefficients become zero.

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 dataset, which one would you prefer and why?

Answer:

The two given linear models performs equally well on the test dataset.

L1: y = 39.76x + 32.648628

L2: y = 43.2x + 19.8

I would prefer **L2: y = 43.2x + 19.8** because L2 model is simpler than L1. Here L2 consumes less number of bits as compared to other model or for representing L2 we need less bits and for L1 we need more bits for representing.

L1: y = 39.76x + 32.648628, has coefficient "39.76", i.e. 3976/100, which is not simple integer as compared to L2: y = 43.2x + 19.8, has coefficient "43.2", i.e. 432/10, which is simplpe integer.

y = 43.2x + 19.8, has simple integer coefficients. Hence, it will give the simplest model.

Below is the reason for selecting simple model:-

i. A simpler model is usually more generic than a complex model. This becomes important because generic models are bound to perform better on unseen datasets.

ii. A simpler model requires less training data points. This becomes extremely important because in many cases one has to work with limited data points.

iii. A simple model is more robust and does not change significantly if the training data points undergo small changes.

iv. A simple model may make more errors in the training phase but it is bound to outperform complex models when it sees new data. This happens because of overfitting.

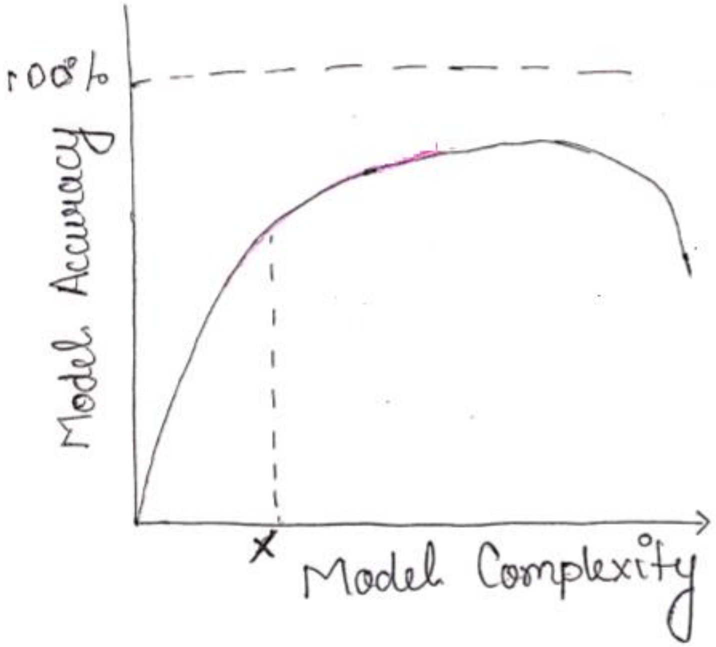
<u>Question-4:</u>

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?

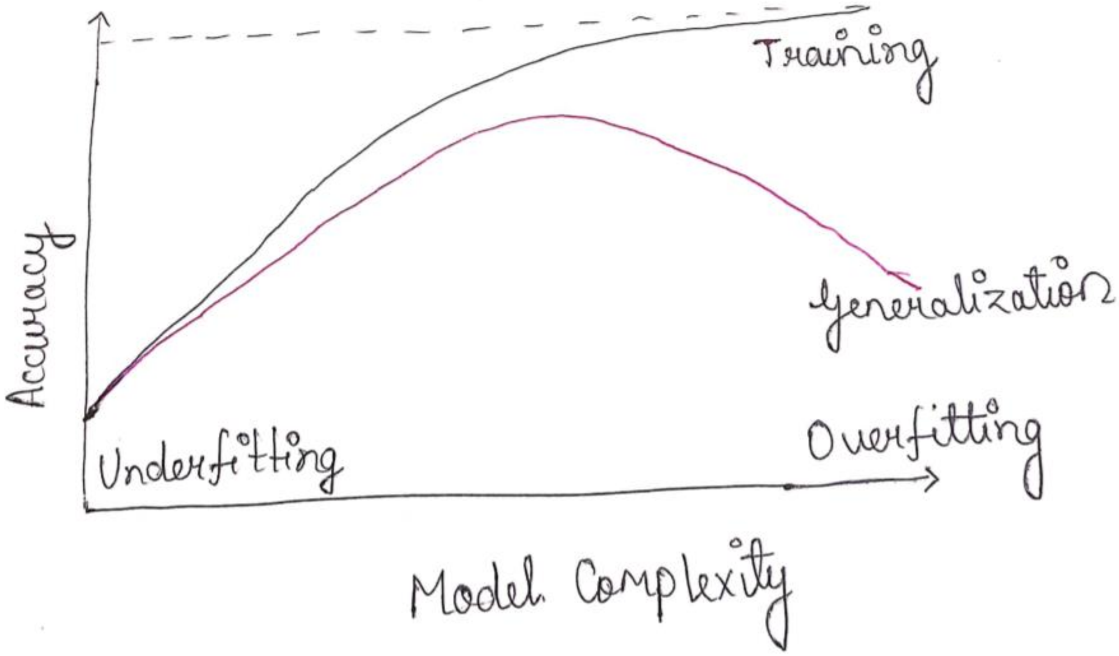
<u>Answer:</u>

We can say a model is robust and generalisable when it is simple. Such models works well even if there are small changes in training data. They are immune to the specifics of training data provided . It rather picks essential characteristics that is invariant across any training datasets. Generic models are bound to perform better on unseen datasets. Such models may make more errors in the training phase but it is bound to outperform complex models when it sees new data. This happens because of overfitting.

There should be a trade off between Accuracy and generalisation (robustness) of the model.



It shows that with increasing levels of complexity we obtaining diminishing returns in terms of accuracy, which never reached 100% accuracy.



The model accuracy on the training data is quite high for the complex model compared to the simple linear model. However, if we extrapolate the fitted lines for both the models, the complex model fails miserably while predicting test data.

Question-5:

As you have determined the optimal value of lambda for ridge and lasso regression during the assignment, which one would you choose to apply and why?

Answer:

Determining the optimal value of lambda for ridge and lasso regression during the assignment

Comparing each of the cross-validation scores allows me to identify the best model for predicting house prices. The ridge regression produces the highest score which implies that shrinking the coefficients for select features improves the overall explanatory power of the model. However, because the scores are so similar, using lasso eliminates coefficients and allows for a more interpretable model, therefore I will make predictions using **lasso regression.**

For selecting which variables are significant in predicting the price of a house we use lasso regression which results in model parameters such that the lesser important features' coefficients become zero and those who has non zero values are the features which are important for predicting price of a house.

In Lasso regression we see from the plot that the r2 score of testing and training keeps on decreasing as we increase the value of hyperparameter. So i select the value of lambda where testing value is high for lambda value and model is simple.

In Ridge regression we see from the plot that the test r2 value first increases and then decreases forming a bell curve. But the training r2 value keeps on decreasing as we increase the value of hyperparameter, which is in accordance with the bias-variance trade-off. So i select the value of lambda where testing value is high for lambda value.