Assignment Part-II

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 problem faced by Rahul is a case of "overfitting". When our model does very good in train data, it tries to memorize it and its complexity would be high. Therefore, the variance would be high for such model. It would perform very well on the data it was trained on but fails miserably with new data.
The problem can be solved by introducing some bias into the model. This can be done by regularization. It aims to make the model simpler. It does a simplification of the training algorithm to control model complexity. Ridge and Lasso are two methods of regularization in regression.

Question-2:
List at least 4 differences in detail between L1 and L2 regularization in regression.

Answer:

L1 Regularization(Lasso)	L2 Regularization(Ridge)
The regularization term is the sum of the absolute values of the coefficients.	The regularization term is the sum of the squares of the coefficients
Lasso shrinks some of the variable coefficients to 0 if they are irrelevant.	Ridge enforces the coefficients to be lower but it does not enforce them to be zero.
As it shrinks some of the variable coefficients to 0, thus it performs variable selection.	Ridge regression can't zero coefficients. Therefore, does not perform variable selection.
Lasso is computationally more intensive	Ridge is computationally less intensive

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:

I would prefer L2(Ridge). As both of the models perform equally well on the test dataset, we would choose the more computationally effective one i.e. Ridge. Also only one feature is used, therefore we do not need any feature selection which is done by Lasso method.

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:

When building a model, we need to make sure that it is robust and generalizable. It should perform well in train data as well on test data. It should not have high variance such that it memorizes the train data and fails with low accuracy on test data. The model should not be too complex. We need to ensure that the model has a good balance between its bias-variance. This can be done by regularizing the model using lasso and ridge which tries to perform model simplification by trying to shrink variable coefficients. Lasso can further help in feature section making the model more robust.

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:

From our model we get lambda for Lasso regularization as 500 and for Ridge regularization we get optimal lambda as 8.
I would choose Lasso with an optimal lambda value of 500.
This ensures that I perform a feature selection as we have a very large number of features for the dataset. Also, a higher lambda ensures I am giving a stricter penalty on features and also not losing on model performance. Though it is computationally more intensive, it does meet our needs for the required feature selection