Advanced Regression Assignment

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 1:

The main reason for this gap is the model created by Rahul is overfitting. To resolve this, we must use regularization – cross validation is one of the methods to overcome the overfitting.

Question 2

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

Answer 2:

L1 is complex model where L2 is simple model.

L1 is having low bias and high variance where L2 is having high bias and variance

Accuracy will be high for L1 and low for L2

Overfitting chance will be high for L1 but low of L2

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 3:

I would prefer L2 because the model is simple, and as per Occam's razor we need to keep simple as much we can. Also, simple model behaves well in unseen data than the complex one.

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 4:

A model is robust and generalizable when it works well on unseen data. Hence the model should be regularized and keep simple. The accuracy tends to be moderate because of high bias, But it works very well for unseen data as it has low variance hence it is robust and generalizable.

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 5:

For Ridge I will use alpha 25 as it is the optimal value by checking the mean test score and mean train score after doing k folds with Grid search view in the graph which shows the optimal value. The alpha value gives us the best model with good accuracy and no issue with overfitting.

For lasso I will use alpha 250 as it is the optimal value for no overfitting issue and have good r2 score for train and test set. It also gives good accuracy from my best model.