Ques1. 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?

Solution-The difference between the test and train accuracy is basically due the problem of Over fitting.

Over fitting is basically a phenomenon where a model becomes way too complex than what is warranted for the task at hand and as a result suffers from bad generalization properties.

The test and train accuracy can be explained as:-

The train accuracy is the accuracy of a model on examples it was constructed on.

The test accuracy is the accuracy of a model on examples it hasn't seen.

For example a 'model' that 'learns' perfectly from any given dataset is one that just memorizes the entire dataset. Clearly the error committed by such a model on the dataset which it was 'trained' on would be zero. However it is clear that such a model can do nothing other than answer questions (though perfectly) about the dataset it was trained on. The 'model' will effectively be incapable of doing anything better than a random guess on test data point that is outside the training dataset. This would be an extreme example of the Over fitting phenomenon — perfect on the training data but unacceptably large error on test data and hence causing the difference between test and train accuracy.Hence If our model does much better on the training set than on the test set, then we're likely overfitting.

Here are a few of the most popular solutions for overfitting which we have studied till now :

1.Cross-validation-Here we use our initial training data to generate multiple mini train-test splits and use these splits to tune our model

2.Train with more data

3.Regularization-Regularization is the process used in machine learning to deliberately simplify models. Through regularization the algorithm designer tries to strike the delicate balance between keeping the model simple yet not making it too näive to be of any use.

Ques2. List at least 4 differences in detail between L1 and L2 regularization in regression.

Solution-The major difference between the L1 and L2 is that, L2 is the sum of the square of the weights, while L1 is just the sum of the weights.

L1 is basically know as Lasso regression and L2 is known as Ridge regression.

L2 regularization is computational more efficient because of its analytical solutions where as L1 is computational in-efficient on non-sparse cases.

L1 is basically a in-built feature selection technique where as L2 is not .

L1 is also computational more intense than L2.

Ques3. 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?

Solution-I would prefer L2 as it is computationally less complex than L1 ,as L1 requires more number of bites to store the data in comparison to L2.Hence as a result of equal test performance also , we prefer L2.

Ques4. 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?

Solution- A predictive model has to be as simple as possible, but no simpler. There is an important relationship between the complexity of a model and its usefulness in a learning context because :

• Simpler models are usually more generic and are more widely applicable (are generalizable)

• Simpler models require fewer training samples for effective training than the more complex ones .

Hence we generalize our model and our accuracy is also maximum on a unseen data set when they are simple.

Ques5. 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?

Solution- Ridge and lasso are two different forms of regularized linear regressions.In the ridge, the coefficients of the linear transformation are normal distributed and in the lasso they are Laplace distributed. In the lasso, this makes it easier for the coefficients to be zero and therefore easier to eliminate some of your input variable as not contributing to the output and hence a feature selection technique.

Whereas ridge is a bit easier to implement and faster to compute, which may matter depending on the type of data you have and generally , when you have many small/medium sized effects you should go with ridge. If you have only a few variables with a medium/large effect, go with lasso.

In our housing assignment analysis , we have alpha values as approximately 50 in ridge having 90 and 95 r squared value for test and train in ridge regression and 500 alpha value in lasso with 90 and 94 r squared values for train and test set respectively.

Here I will choose lasso as the model size is less.
