Assignment – Advanced Regression

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-1:

One of the possible reason for this much difference in the training accuracy and test accuracy could be the overfitting. Overfitting is when the accuracy on the test data is very low as compared to the train data. This can be solved by Regularization in which we built a model which as simple as possible but not too naïve. Once our model is regularized, we must not see this much gap in the accuracy of the test and training.

Moreover, Overfitting could be an issue due to Multicollinearity as multicollinearity causes the model to not understand the pattern correctly and hence leading to low accuracy on the test data.

Another possible factor could be outliers, which leads to the similar issue as created by Multicollinearity i.e. the outliers do not allow our model to capture the trend properly.

Thus, Regularization, Handling multicollinearity & removing outliers can solve the problem being faced by Rahul.

Question-2:

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

1. L1 is called the Lasso regularization whereas L2 is called Ridge regularization.
2. L1 gives inbuild feature selection by reducing the coefficients to zero where L2 only reduces the coefficients sharply. Thus, feature selection has to be done again if you're using L2 by some logic i.e. to eliminate value close to 0 but L1 does that internally.
3. L1 is does far more computations than L2 i.e. L2 is computationally efficient than L1
4. L1 adds 'absolute value of magnitude' as the penalty term to the cost function whereas L2 adds 'squared magnitude' of coefficient as penalty term to the cost function.
5. L1 is robust whereas L2 is not so much robust
6. L1 generally produces sparse output whereas L2 does not. Sparsity means that only very few values in a matrix are non-zero which is the case with L1

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-3

Given that both the above models perform equally well on the test dataset, the model L2 seems a bit more simplified than the L1. One thing in terms of complexity to note here is that L1 coefficient is 39.76 where as in L2 the coefficient of x is 43.2 which would computationally efficient as its having less precision. The same is the case with the constant, the constant in L1 would require far more number of bits to represent, also the precision is very high. Thus, we would go with the L2 model.

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

One can make sure that the model is robust and generalisable by checking the accuracy of the model built both on the test as well as the training set. Now, if the model accuracy is low on the test set and high on the train set as compared to the test set.

That clearly means that our model is not generalisable. This happens because our model is so complex that it has made sure to learn the pattern from each and every data point such that when the change occurs in the dataset(just as checking accuracy on the test set), the model performs poorly. Hence, this kind of model is not robust & generalizable.

We can make a generalisable model by using the concept of 'REGULARIZATION' which means that the model must be simple enough but not too naïve. We can use the regularization techniques such as L1 (Lasso) & L2(Ridge) to built a model which performs well on both the test as well as the train set somewhat equally. Thus, the model now can perform well on the test set as well as it's a generalizable model.

However, there are some implications of regularization which results in the dropping of the accuracy on the train set but higher than earlier on the test set. This happens because regularization helps in building a model which learns from the training data but not mug up completely which results in the decreased accuracy but as the model has learned now rather than mugging up, it performs better on the test set.

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-5

From the model that was built in the assignment, the Ridge regression performs better in terms of the accuracy than the Lasso regression. However, I would choose the Lasso regression over the ridge as Lasso provides sparsity i.e. it gives most of the coefficients as 0 whereas in Ridge regression I've some values which are close to 0 but not actually 0. Moreover, the difference between the accuracy of the Ridge regression on the test and train set is 0.3 whereas in the Lasso regression it comes out to be 0.6 but the number of features that Lasso provides after removing the ones with the 0 coefficient are much less in number than provided by the Ridge regression.

Thus, I would go for Lasso regression so that the company could focus upon few parameters for the purchase or sale of the houses rather than providing a whole list of features to be focused upon for just 0.3 accuracy increase.