ASSIGNMENT- SUBJECTIVE QUESTIONS

Q1 - 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:-

Training accuracy is the accuracy by which model predict the dependent variable of the data point on which it has been built upon or trained. On the other hand test accuracy is the accuracy by which model predict on the data it has never seen. The model having such a high difference in the train and test accuracy is more likely have completely memorised the training data, but when it comes to test data which he had not seen, it failed to deliver the results. This is know as Overfitting. This problem can be solved by regularisation (e.g. Ridge or Lasso) these are additional criterion that are traded against the complexity of the model built. They try to make the model simple so that it is less likely to get overfitted.

Q2 - List at least four differences in detail between L1 and L2 regularisation in regression.

Answer:-

The regression model that uses L1 regularisation technique is called Lasso Regression and the model which uses L2 regularisation technique is called Ridge regression.

The differences between them are:

1- The regularisation term in L1 is the sum of the absolute value of the weight, while L2 is the sum of the square of the weight.

2- The L1 regularisation in the regression gives us the sparse output, whereas L2 regularisation in the regression gives us non sparse output.

3- The L1 regularisation is not computational efficient i.e. it does not have analytical solution but L2 regularisation have analytical solution which allowed it to be computational efficient.

4- L1 regularisation has feature selection properties i.e. it shrinks the less important feature’s coefficient to zero, thus removing the feature altogether, whereas L2 regularisation does not have that properties.

5- Regularisation contours for L1 regularisation is of diamond shape/square shape whereas the contours for the L2 regularisation is of circular shape.

Q3 - 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:-

The model L2 is better given both performed well because it is more simpler i.e. it is more round off, than L1 where the coefficients have more decimal places. So why would we need more complex coefficients and increase the computation power. And also as the model complexity increases variance also increases. And simpler models are usually more generic.

Q4 - 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:-

The robustness is the characteristic describing a model's ability to effectively perform while it's variables or assumptions are altered. The generalisability is the characteristic describing a model's ability to find abstract pattern in data and which can be applied to a large variety of data it can encounter. To make a model more robust and generalizable it should be simpler, remove the outliers etc., so that the model can predict well for the unseen data. By making model more robust and generalizable the accuracy of the model gets reduced. Since it is measured by how much the model is able to predict the training data. And since the model is trying to be more robust, it will try to avoid the noise in the data and try to be more generic this leads to lesser accuracy.

Q5 - 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:-

It depends on the data we have Ridge regression is bit easier to implement and faster to compute and on the other hand Lasso regression have feature selection property. So if we have less computation power then we would go for Ridge regression and if we want to reduce the number of feature we would go for the Lasso regression. But generally we go for Lasso regression because feature selection is most important in most cases.