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

Problem / Reason

1. The prime reason could be **Overfitting**. The model has learnt the training dataset too well and it will yield high accuracy results for this dataset (train). However, for the dataset's available in the practical world (test), the unseen datasets, it would perform poorly. The model captures all the noise and the training dataset data pattern. The model thus is a **Complex Model**. It would have **high variance** and **low bias**.
2. The **sample size is small** leading to overfitting and not a generalized model trained on various patterns of the dataset.
3. The features selected in the model may have **high collinearity** or may have **high p-values**.

Solution

1. Make Model simpler with low variance and high bias. Utilize regularization techniques to achieve this.
2. Perform Cross Validation by validating the training model on a validation set. Techniques like Kfold can be adopted to achieve this.
3. While building models ensure high p-value and high VIF (multicollinearity) features are removed while modelling.
4. Increase the sample size. There are techniques using which the sample size can be increased.

Question 2

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

Answer 2

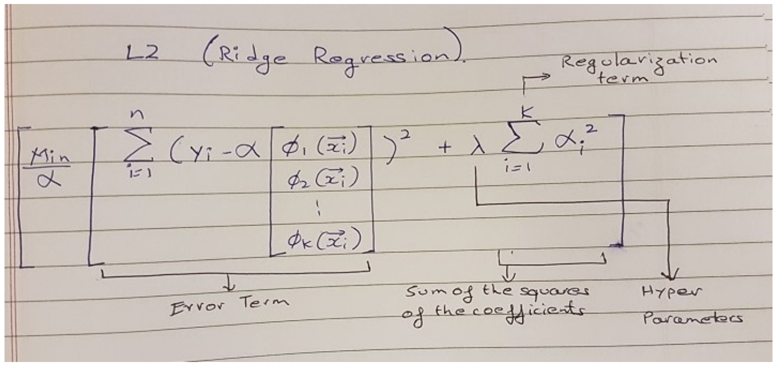
L1 Regularization is referred as Lasso and L2 Regularization is referred as Ridge.

1. In L1 regression the sum of the absolute value of the coefficients is added to the cost function.

L1 (Lasso Regression)

$$\underset{\alpha}{\mathrm{Min}} \left[\sum_{i=1}^{n} \left(y_i - \alpha \begin{bmatrix} \phi_1(\vec{x}_i) \\ \phi_2(\vec{x}_i) \\ \vdots \\ \phi_K(\vec{x}_i) \end{bmatrix} \right)^2 + \lambda \underbrace{\sum_{i=1}^{k} |\alpha_i|}_{\text{Sum of the absolute values}} \right]$$

In L2 regression the sum of the squares of the coefficients is added to the cost function along with the error term



2. L1 Regularization trims down the coefficients of redundant variables to zero. Hence it indirectly performs variable selection as well.
   L2 Regularization reduces the coefficients to quite low values but h not zero.

3. L1 Regularization is computationally inefficient algorithm as it does not have an analytical solution while L2 Regularization is computationally efficient.

4. L1 Regularization gives sparse outputs, that is only few entries are non-zero whereas L2 regularization gives non-sparse outputs.

5. L2 Regularization almost always has a matrix representation for the solution, while L1 Regularization requires iterations to get to the final solution.

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

L2, since the complexity is lower in comparison with L1. L1 requires more storage space than L2.

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

1. Simple Models have low variance and high bias and are thus more robust and generalizable in comparison to Complex Models which have high variance and low bias.
2. Implications – A complex model works well on the train dataset and learns the noise along with the data patterns of the train dataset. When it encounters real-world datasets then it fails to function and the accuracy of the model reduces considerably.
3. We need to ensure the Bias-Variance trade off to ensure that the model is neither overfitting nor underfitting.

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?

Lasso Regression as it clamps down on the number of features by reducing the coefficients of more than 50% features to zero and hence they can be dropped from the model. Ridge Regression on the other hand reduces the coefficients value close to zero but not exactly zero. Hence, Lasso Regression helps in dimensionality reduction.