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

The model built by Rahul has a high training accuracy (97%) and low test accuracy (48%). This means that the model fits well on the training data, but doesn't fit that well on test data or unseen data. This is known as Overfitting. Instead of learning from the test data, the model has memorized the data points and that is why the train accuracy is high. But since the test data is unseen data, the model doesn't generalize well for test data and the accuracy is low. This also happens when the model becomes too complex. The complex model thus build will included all the noise and coincidences (or outliers) present in the test data, and the accuracy will be low when predictions are made on unseen data.

This problem can be solved by using regularized regression techniques. Regularization is a process used to create a model that is as simple as possible while performing well on the training data. Such a model is known as an optimally complex model. It isn't any more complex than it should be. It follows the principle which says "Everything should be made as simple as possible, but not simpler. Since linear regression doesn't account for model complexity and only tries to minimize the error, it may result in coefficients that are arbitrarily complex. Regularization introduces a regularization term along with the error term. This regularization term introduces a penalty for overly complex coefficients. This term is added to the cost function. The more complex the coefficients become, the more is the imposed penalty. These methods are used to make the model simpler, while also balancing the variance-bias trade-off.

Two such techniques are Ridge Regression and Lasso Regression. Ridge regression introduces a term which is equal to the sum of the squares of the coefficients. In Lasso regression, the term is equal to the sum of absolute value of the coefficients. Also, Lasso reduces the coefficients of the lesser important features and makes them zero. This indirectly performs feature selection.

Question 2

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

Answer

Lasso Regression is L1 regularization technique and Ridge Regression is L2 regression technique. Key differences are as below:

1. L1 regression adds the sum of the absolute value of the coefficients as the regularization term to the error term whereas L2 regression adds the sum of the squares of the coefficients as the regularization term to the error term.
2. L1 regression is computationally more intensive since a simple matrix function cannot yield the solution unlike L2 regression where the solution can be obtained by a simple matrix function.
3. L1 regression indirectly performs feature selection since it reduces the coefficients of insignificant features to zero. L2 regression merely reduces the coefficients of insignificant features to a low value and no feature selection happens.

4. L1 regression is preferred more for scenarios where the number of features is very high. This is due to its capability of generating sparse solutions. In such cases, getting sparse solution is of a great computational advantage. L2 regression is preferred more for scenarios where prevention of overfitting is required for a relatively less number of features.
5. L1 regression and L2 regression work differently when there are features which are highly correlated. L1 regression arbitrarily selects any one feature from the highly-correlated ones, and sets the coefficients of the other features to zero. This doesn't work that well as compared to L2 regression where all the highly-correlated features are included and the coefficients are distributed among them depending on the correlation.

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

Given that both the models perform equally well on the test data set, model L2 is preferred over model L1. The reason being model L2 is simpler that model L1. It can be said by looking at the coefficients of the features. Model L1 has complex coefficients as compared to model L2. A simple model is more robust and more generalizable for unseen data. Complex coefficients might lead to wild swings in values even for a small variation in data. Owing to its simplicity and generalizability, model L2 will be preferred.

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer

A simple model is generalizable and is more robust since it has low variance. To make a model robust and generalizable, the principle followed is that a model should be as simple as possible, but not simpler. A model shouldn't be too simple to be considered naïve. Also, it shouldn't be too complex to be considered over-fitted. Such a model is known as optimally complex model. It is a model which is as complex as it needs to be, and not more.

A simple model has low accuracy and hence high bias. The accuracy is low since a simple model doesn't try to fit the data perfectly and aims to be generalizable. There is a bias variance trade-off involved in the process of obtaining a model which is optimally complex.

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

The alpha value chosen for Ridge Regression is 20 since the negative mean absolute error peaks at this value and then falls. The alpha value chosen for Lasso is 0.001 since it is giving optimal results for the train-test error comparison.
I will choose to apply Lasso Regression. Even though RMSE as obtained on the test data are similar for both the regression techniques, the train error and the test error are closer for Lasso in this case as compared to Ridge. This suggests that the model obtained by Lasso is a more generalizable model. Also, Lasso is providing a model which feature selection applied on it.