ASSIGNMENT

1 Answer :

Given that the test accuracy is low(48%) and training accuracy is high(97%) ,The gap between the accuracies is because of **overfitting**.

Overfitting is the scenario when the model becomes highly familiar with the data in the training set that the model starts to "remember" the values and as a result of this training accuracy becomes high.

An overfit model result in misleading regression coefficients, p-values , and R-squared statistics.

A good model should fit not just the sample you have, but any new samples you collect from the same population.

The best solution to an overfitting problem is avoidance. Identify the important variables and think about the model that you are likely to specify.

Few methods can be followed:

1. Cleaning the dataset thoroughly
2. Removing outliers .
3. Handling multicollinearity

2 Answer:

Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression.

Ridge Regression : In ridge regression, the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients.

Lasso regression not only helps in reducing over-fitting but it can help us in feature selection.

The main difference between these two is the penalty term.

Ridge regression takes "squared magnitude" of coefficient as the penalty term to the loss function.

Lasso regression takes "absolute value of magnitude" of coefficient as penalty term to the loss function.

Lasso is used in feature selection since it makes the coefficients of the less important features as whereas in Ridge regression, the coefficients only get close to 0 and not 0 and hence cannot be used in feature selection.

Lasso regression is computationally more intensive and ridge regression is computationally less intensive

3 Answer :

Since it is mentioned that both models perform equally well on the test data, we would still prefer the second model since it is the simpler model out of the two. Occam's razor principle states that a model

should be as simple as possible but no simpler and that we need to choose the simpler model of the two, when in doubt. Hence, we pick L2 over L1.

4 Answer :

A model is more generalizable and robust when it is simple.

1. A simpler model is usually more generic than a complex model. Usually,generic models are bound to perform better on unseen datasets.
2. A simpler model requires less training data points,In many cases one needs to work with limited data points.
3. A simple model is more robust and does not change significantly if the training data points undergo small changes.
4. A simple model may make more errors in the training phase but it is bound to outperform complex models when it sees new data

Implications on Accuracy:

A simpler model has HIGH bias and low variance.

Bias quantifies how accurate is the model likely to be on the test data set.

Complex models, assuming you have enough training data available, can do a very accurate job of prediction. Models that are too simple typically have low accuracy. We need to use the technique of regularization to strike a balance between accuracy and model complexity.

5 Answer :

Much like the best subset selection method, **lasso** performs variable selection. The tuning parameter **lambda** is chosen by cross validation. When **lambda** is small, the result is essentially the least squares estimates. As **lambda** increases, shrinkage occurs so that variables that are at zero can be thrown away.

The only difference from Ridge regression is that the regularization term is in absolute value.

Lasso method overcomes the disadvantage of Ridge regression by not only punishing high values of the coefficients β but actually setting them to zero if they are not relevant