Advanced Regression Assignment 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:

This might have happened because the model that Rahul built is probably a complex one. The complex model suffers from the following drawbacks:

1. A complex model is usually less generic than a simpler model. This becomes important because generic models are bound to perform better on unseen datasets.
2. A complex models bound to underperform compared to simpler models when it sees new data. This happens because of **overfitting**.

The model has become too specific to the data it is trained on. Thus the high training accuracy percentage(97%). But it has failed to generalise to other unseen data points in the larger domain. A model that has actually 'learnt' not just the hidden patterns in the data but also the noise and the inconsistencies in the data. It basically memorizes all the data points. In front of a new dataset, thus it fails to predict successfully.

Question-2:

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

Answer-2:

A regression model that uses L1 regularization technique is called ***Lasso Regression*** and model which uses L2 is called ***Ridge Regression***. Some of the major differences between these two regularization techniques are as follows:

L1 (***Lasso***) Regularization	L2 (Ridge) Regularization
L1 regression adds "***absolute value** of magnitude*" of coefficient as penalty term to the loss function.	**L2 regression** adds "***squared magnitude***" of coefficients as penalty term to the loss function.

L1 regression helps in **feature selection**. Lasso regression results in a sparse solution which means, if there are redundant features, using Lasso, the features might get dropped off. The coefficients of these redundant features become 0.	**L2 regression** includes all of the features in the model. It creates coefficient shrinkage by reducing the feature coefficients to almost 0 but not exactly 0. Thus it doesn't aid in feature selection.
L1 regression is computationally **more intensive** than L2.	**L2 regression** is computationally **less intensive** than L1.
With **L1 regression**, you cannot find a close formed solution of cost function.	With **L2 regression**, you can find a close formed solution of cost function.
Since **L1 regression** helps in feature selection, it is used for modelling cases where the number of features are huge.	**L2 regression** is majorly used to *prevent overfitting*. Since it includes all the features, it is not very useful in case of exorbitantly high number of features.
L1 regression arbitrarily selects any one feature among the highly correlated ones and reduced the coefficients of the rest to zero.	**L2 regression** generally works well in the presence of highly correlated features as it will include all of them in the model but the coefficients will be 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 dataset, which one would you prefer and why?

Answer-3:

The 'complexity' of a model can increase depending on the number of bits required to store the feature coefficients. i.e. a model with coefficient equal to 5.7785 is more complex than one with coefficient equal to 5.

In the above example, the complexity of the model L1 is much higher than that of the model L2. This is because the number of bits required to store the feature and intercept coefficients in the former is more than that in the latter.

Given the fact that both the models perform equally well on the test dataset, it is advisable to select the L2 model because it is a much simpler and generic model and can work well with unseen data.

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:

A robust and generalizable model always performs well on unseen data by keeping the learning algorithm simple yet not making it too naive to be of any use. This results in minimizing overfitting.

One way of creating robust, generalizable models is through Regularized Regression process. In regularized regression, the objective(cost) function has two parts - the error term and the regularization term.

In order to regularize a regression model, we add the Regularization term to the Residual Sum Of Squares term and minimize the entire quantity. This quantity is the cost function for the regression model.

$$\sum(wTxi-yi)2+\boldsymbol{\lambda}\mathbf{R(w)}$$

Here **R(w)** is the regularization term and lambda(**λ**) is the coefficient of regularization or the hyper parameter of the process. Using lambda is a way of penalizing the model for being complex. The optimal value of **λ** is chosen using cross validation.

λ governs the behaviour of the regression algorithm by controlling the complexity of the model. Through **λ,** one tries to strike the delicate balance between Bias and Variance. If λ is sufficiently large, then regularization happens. Complexity is decreased. Even though the bias increases, the variance is lowered making it a generalizable and robust model.

However we have to be careful not to increase λ so much that the model becomes too simple and it under fits the data. In other words, the **accuracy** of the model is affected to a large extent. Ideally Regularization should be used to create an optimally complex models, i.e. models which are as simple as possible while showing acceptable accuracy on the training data.

Question-5:

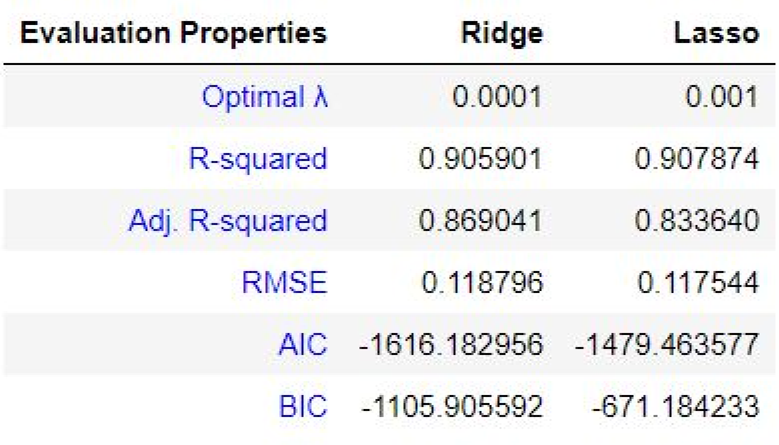
As you have determined the optimal value of lambda for ridge and lasso regression during the assignment, which one would you choose to apply and why?

From the assignment, the optimal values of lambda obtained for ridge and lasso regression are as follows:

For Ridge it was found out to be **0.0001** and for Lasso Regression the value was **0.001.**

Thus, the value of hyper parameter(**λ**) is more for Lasso and less for Ridge.

Now let us consider the evaluation statistics of the two models -- ridge and lasso:



Evaluation Properties	Ridge	Lasso
Optimal λ	0.0001	0.001
R-squared	0.905901	0.907874
Adj. R-squared	0.869041	0.833640
RMSE	0.118796	0.117544
AIC	-1616.182956	-1479.463577
BIC	-1105.905592	-671.184233

As per the table above, the R2 test score of both the models are more or less equal. This means that both the models can successfully explain more than 90% of variance in the data which is a good result.

Since, accuracy of both the models is almost similar and high, choosing a better model from the above two, really boils down to the 2 facts below:

Dimensionality Reduction

Lasso Regression model helps in feature selection by dropping the redundant features. We have seen during the assignment, that lasso was successful in dropping close to 37% of the feature which were redundant.

So in case we are looking for dimensionality reduction, we can surely go with the lasso model of regularized regression.

Dimensionality Reduction

Ridge regression generally works well in the presence of highly correlated features as it will include all of them in the model but the coefficients will be distributed among them depending on the correlation.

Lasso on the other hand arbitrarily selects any one feature among the highly correlated ones and reduced the coefficients of the rest to zero. This may result in some important feature(as per business) getting dropped. There is no control over feature dropping.

Generally, Ridge Regression is the go to technique for analysing multiple regression data that suffer from multicollinearity.

From the figure above, we see that although the R2 score for both the models are a 90%(approx value), the adjusted R2 value for ridge model is 87% and for lasso model it is 83%. A bigger gap between R2 and adjusted R2 scores in lasso suggests the presence of higher multicollinearity among the model features.

Thus in case we are looking for multicollinearity reduction, we can surely go with the ridge model of regularized regression.

Combining Lasso and Ridge

Thus we are faced with a trade off between dimensionality reduction and multicollinearity reduction while choosing between Lasso and Ridge models.

The best course of action in such a case is to combine both the models.

We can use Lasso regression with optimal value of lambda to select the most significant features that can predict the dependent variable successfully. And then we can finally build the ridge model using only these select features.

This will result in both dimensionality reduction and at the same time take care of multicollinearity as well.