

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:

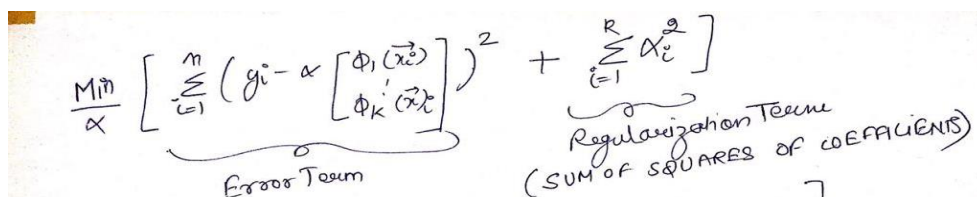
The reason behind the difference of 49% in train and test accuracy is because the model best fits only on training data set and it is unable to generalize beyond the data on which it is trained. This occurs because model is overfitting the training data set.

Overfitting may result in the case where model used to predict is too complex (due to the involvement of maximum features) and it is made in such a way that it has learnt the training data. To avoid overfitting of data a model should be optimally complex i.e., maintaining the complexity along with simplicity. **This can be achieved by performing regularized regression.**

In linear regression complexity is not taken into account it only focuses on reduction of error term.

Whereas in regularized regression focus in on striking the balance between error term and regularization term. Regularized regression can be done using two techniques as below for large data sets.

1. **Ridge regression** : Along with error term a generalized term (sum of squares of coefficients of raw attributes) is added.



The image shows a handwritten mathematical formula for the Ridge regression loss function. The formula is:
$$\frac{\text{Min}}{\alpha} \left[\underbrace{\sum_{i=1}^n \left(y_i - \alpha \begin{bmatrix} \phi_1(\vec{x}_i) \\ \vdots \\ \phi_k(\vec{x}_i) \end{bmatrix} \right)^2}_{\text{Error Term}} + \underbrace{\sum_{i=1}^R \alpha_i^2}_{\text{Regularization Term (SUM OF SQUARES OF COEFFICIENTS)}} \right]$$

2. **Lasso regression** : Along with error term a generalized term (sum of absolute values of coefficients of raw attributes) is added.

$$\frac{1}{2} \min_{\alpha} \left[\underbrace{\sum_{i=1}^n (y_i - \alpha [\phi_1(\vec{x}_i) \dots \phi_k(\vec{x}_i)])^2}_{\text{Error Term}} + \underbrace{\sum_{i=1}^k |\alpha_i|}_{\substack{\text{Regularization Term} \\ (\text{SUM OF ABSOLUTE VALUES OF COEFFICIENTS})}} \right]$$

Lasso is preferred over ridge as it performs feature selection as well by reducing the coefficients of irrelevant features to 0. For small data sets it's advised to use cross validation and step wise regression techniques.

Question-2:

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

Answer-2:

L1 and L2 regularization regression techniques refers to Lasso and Ridge regression respectively. These are used to regularized the model to maintain a correct balance between the model complexity and simplicity thereby reducing the chances of overfitting and helps in performing optimum feature selection.

Lasso Regression (L1)	Ridge Regression (L2)
This has two components error term and regularization term. Regularization term is explained as sum of absolute values of coefficients of attributes. This term is added to the cost function.(Above picture in question 1)	This has two components error term and regularization term. Regularization term is explained as sum of squared values of coefficients of attributes . This term is added to the cost function.(Above picture in question 1)
It helps is performing feature selection as well by reducing coefficients of certain irrelevant features to 0.	It doesn't reduce the coefficients to 0 for any of the features.
It is computationally expensive.	It is computationally less expensive.

It is slower in execution process as compared to Ridge but data can be interpreted in a more efficient and simple manner as it eliminates irrelevant features completely.	It's execution time is faster but the interpretation of data is complex since while representing it includes unimportant features as well.
---	--

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:

If both L1 and L2 performs equally well on test dataset , then we should go for simple model indeed.

1. Firstly simplicity factor can be judged in terms of representational power. By looking at the equations we can clearly determine that L2 is simpler than L1 as total memory bits and precision required to represent this equation are far less than what needed in L1.

L1(32.648628 upto 6 decimal places) needs to have more bits in terms of handle decimal precision values as compared to L2(Only upto 2 decimal places).

2. Secondly L2 can generalize better than L1 because of reduced complexity.

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-4:

A model if built with regularization techniques (Ridge/Lasso) can be made more generalizable. Along with minimizing error term, additionally we try to minimize regularization term as well (Sum of squares of coefficients/Sum of absolute values of coefficients). Regularization term makes sure that model doesn't become too complex, it ensures that model is not too naïve at the same time not too complex. This can be monitored as the coefficients start becoming more complex than altogether error term + regularization term will tend towards maximization.

Reducing the complexity of model may impact accuracy of the model in a negative way, to ensure acceptable accuracy of the model we need to have a balanced value of lambda thereby maintaining the bias variance trade off. Therefore lambda is the regulatory hyperparameter here which should be optimally chosen to strike a balance between complexity and generalization.

As lambda tends to increase then any sort of irregularity in coefficient values can be controlled. Also as it tends to 0 then we are ultimately doing an unregularized regression.

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?

Answer-5:

As predicted in the assignment that both Ridge(lambda = 6) and Lasso(lambda = 80) performs well on training data maintaining the complexity and generalizability trade off of the model.

Ridge training vs test score.(R2 score)

Train R2 Score : 0.9214310084979984
Test R2 Score : 0.8946591534509657

Lasso training vs test score.(R2 score)

Train R2 Score: 0.9257155417611125
Test R2 Score : 0.9021989679633343

Lasso regression has an edge above Ridge since it performs optimal feature selection as well by reducing the coefficients of irrelevant/unimportant features to 0. This gives us the reduced set of features which is optimal and makes the representation easy.

Also Lasso provides us with the sparse solutions , i.e the touch points between error term contours and lasso regularization contours are more likely to be on one or more of the axes which directly implies that coefficients of certain features are reduces to 0 , thereby performing optimal feature selection.

In comparison to Lasso , ridge doesn't perform any kind of feature selection as all coefficient values are greater or less than 0 or tending towards 0 but not exactly 0.