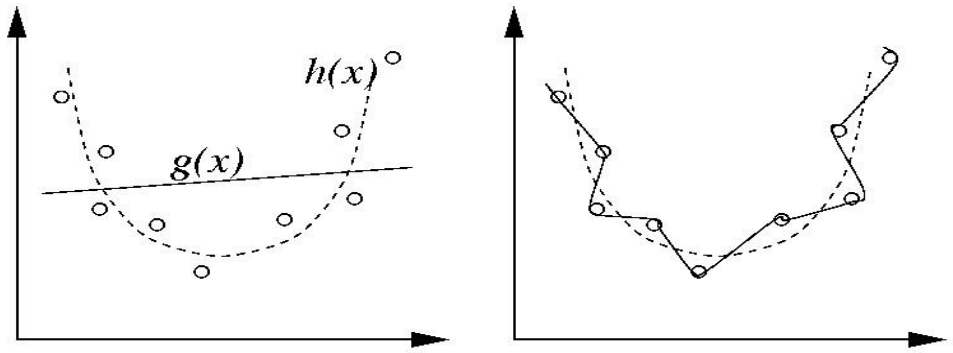
Question 1

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)

The training accuracy is high compared to test accuracy. It is cause of overfitting. It fits the training data too well and generalizes bad.



Overfitting can have many causes and usually is a combination of the following:

- **Too powerful model**: e.g. you allow polynomials to degree 100. With polynomials to degree 5 you would have a much less powerful model which is much less prone to overfitting
- **Not enough data**: Getting more data can sometimes fix overfitting problems
- **Too many features**: Your model can identify single data points by single features and build a special case just for a single data point. For example, think of a classification problem and a decision tree. If you have feature vectors (x1, x2, ..., xn) with binary features and n points, and each feature vector has exactly one 1, then the tree can simply use this as an identifier.

We can overcome the problem of overfitting by

- **Generalisation:** A generalized model can help reduce the overfitting as we can see in the figure above.
- **Regularization:** Using Ridge and lasso regression as a way to regularize the model. Hence reduce the problem of overfitting. Lasso makes the coefficients to zero and in ridge regression it makes the coefficients to weigh less in the final model.

Question 2

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

Answer 2)

L1 Regularization

$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

L2 Regularization

$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda \sum_{j=0}^{M} W_j^2$$

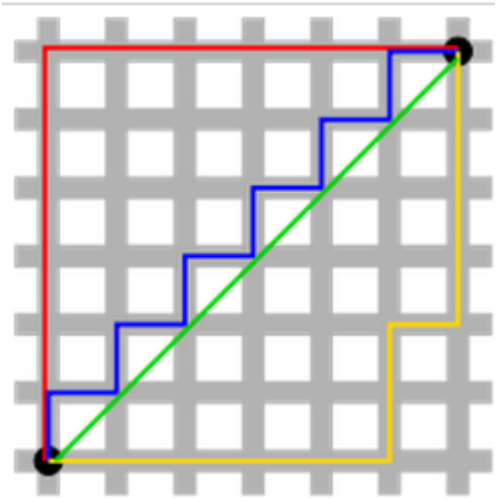
Loss function

Regularization Term

L1 Regularization is also termed as Lasso regression.

L2 Regression is also termed as Ridge regression

Solution uniqueness: For this picture below:



The green line (L2-norm) is the unique shortest path, while the red, blue, yellow (L1-norm) are all same length (=12) for the same route. Generalizing this to n-dimensions. This is why L2-norm has unique solutions while L1-norm does not.

Built-in feature selection: is frequently mentioned as a useful property of the L1-norm, which the L2-norm does not. This is actually a result of the L1-norm, which tends to produces sparse coefficients (explained below). Suppose the model have 100 coefficients but only 10 of them have non-zero coefficients, this is effectively saying that "the other 90 predictors are useless in predicting the target values". L2-norm produces non-sparse coefficients, so does not have this property.

Sparsity refers to that only very few entries in a matrix (or vector) is non-zero. L1-norm has the property of producing many coefficients with zero values or very small values with few large coefficients.

Computational efficiency. L1-norm does not have an analytical solution, but L2-norm does. This allows the L2-norm solutions to be calculated computationally efficiently. However, L1-norm solutions does have the sparsity properties which allows it to be used along with sparse algorithms, which makes the calculation more computationally efficient.

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)

We will prefer L1 model. The reason for selecting L1 model is that beta 1 value of L1 model is 39.76, however for L2 it is 43.2. The impact made by this coefficient in predicting the output is less for L1 compared to L2. And hence the model is well generalized.

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)

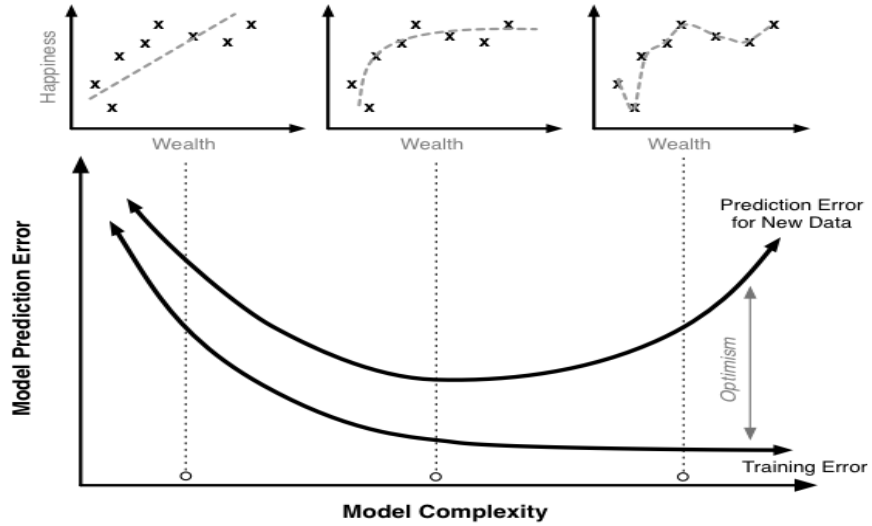
You should follow these two steps while building any model:

1. Carry out exploratory data analysis by examining scatter plots of explanatory and dependent variables.

2. Choose an appropriate set of functions which seem to fit the plot well, build models using them, and compare the results

3. If we have less data in our dataset then we can use techniques like cross validation to generalize model better.

4. We can use regularization techniques to generalize model better.



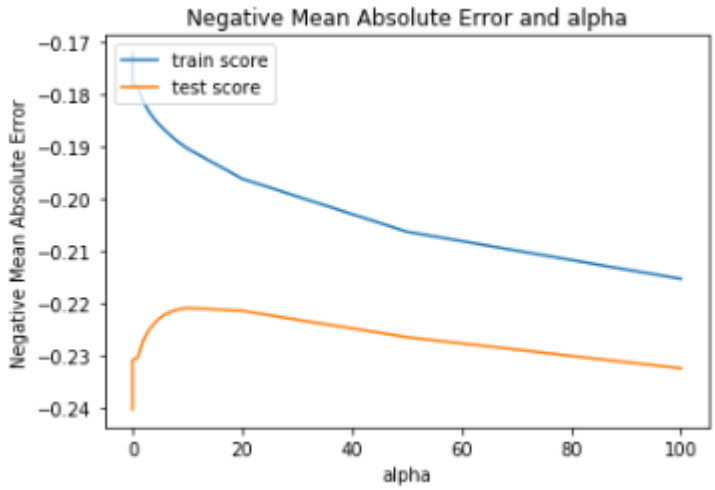
The implication that it causes on accuracy of the model can be understood from the picture in above diagram. When you increase complexity of your model, it is more likely to overfit, meaning it will adapt to training data very well, but will not figure out general relationships in the data. In such case, performance on a test set is going to be poor. Such model is great at remembering, but when it encounters data it has not seen before, it gets 'confused', if you like The error term always increases when make the model 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 5)

Based on our regression we created the graphs as below



For this above graph we chose a value of alpha=10 for ridge and alpha=0.001 for lasso.

We took ‘mean_test_score’ as a parameter. And chose alpha for the highest mean_test_score.

The reason behind this was that we had to reduce the test error in the dataset.