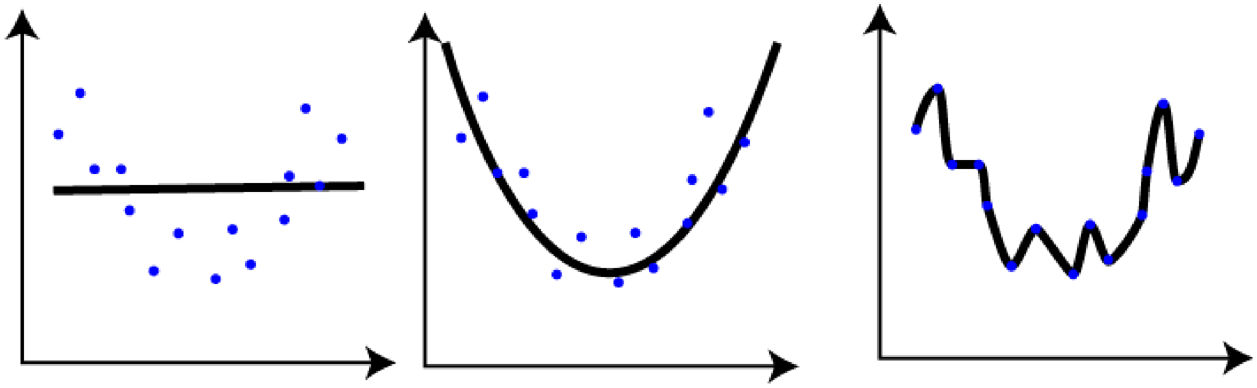
Question 1:
Rahul built a logistic regression model having a training accuracy of 97% while the test accuracy was 48%. What could be the reason for the seeming gulf between test and train accuracy and how can this problem be solved.
Answer:

To find accuracy by implementing a design model on first train set, then on test set. If the accuracy is satisfactory, we tend to increase accuracy of data-sets prediction either by increasing or decreasing data feature or features selection or applying feature engineering in our machine learning model. But sometime our model maybe giving poor result. The poor performance of our model maybe because, the model is too simple to describe the target, or may be model is too complex to express the target. So here comes the concept of overfitting and underfitting. Overfitting and underfitting can be explained using below graph. By looking at the graph on the left side we can predict that the line does not cover all the points shown in the graph. Such model tend to cause underfitting of data .It also called High Bias. By looking at the graph on the left side we can predict that the line does not cover all the points shown in the graph. Such model tend to cause underfitting of data .It also called High Bias. Where as the graph on right side, shows the predicted line covers all the points in graph. In such condition you can also think that it’s a good graph which cover all the points. But that’s not actually true, the predicted line into the graph covers all points which are noise and outlier. Such model are also responsible to predict poor result due to its complexity.It is also called High Variance.Now, Looking at the middle graph it shows a pretty good predicted line. It covers majority of the point in graph and also maintains the balance between bias and variance.



from the picture (from left to right)

(i) under fitting (ii) good fitting (iii) over fitting

to solve our problem, we can use cross-validation to select the best model by creating models with a range of different degrees, and evaluate each one using k-fold cross-validation. The model with the lowest cross-validation score will perform best on the testing data and will achieve a balance between underfitting and overfitting.

1.A simpler model is usually more generic than a complex model. This becomes important because generic models are bound to perform better on unseen datasets.

2.A simpler model requires less training data points. This becomes extremely important because in many cases one has 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. This happens because of overfitting.

Question 2:
List at least 4 differences in detail between L1 and L2 regularization in regression.

A regression model that uses L1 regularization technique is called *Lasso Regression* and model which uses L2 is called *Ridge Regression*.
Ridge regression adds "*squared magnitude*" of coefficient as penalty term to the loss function. Here the *highlighted* part represents L2 regularization element.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

Here, if *lambda* is zero then you can imagine we get back OLS. However, if *lambda* is very large then it will add too much weight and it will lead to underfitting. Having said that it’s important how *lambda* is chosen. This technique works very well to avoid over-fitting issue.

Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds “*absolute value of magnitude*” of coefficient as penalty term to the loss function.

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Again, if *lambda* is zero then we will get back OLS whereas very large value will make coefficients zero hence it will under-fit.

The key difference between these techniques is that Lasso shrinks the less important feature’s coefficient to zero thus, removing some feature altogether. So, this works well for feature selection in case we have a huge number of features.

Traditional methods like cross-validation, stepwise regression to handle overfitting and perform feature selection work well with a small set of features but these techniques are a great alternative when we are dealing with a large set of features.

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?

1) The second model **43.2x + 19.8** can be considered because the equation shows the complexity of curve that is going to be plotted
2) The complexity leads to high storage in memory(bits)

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?

1) Hyper parameters help us to reduce or increase the complexity of the model

2) Regularized models put penalty for having complex algorithms or complex coefficients.

3) So to make Robust and Generalized model we should build a model which is not too complex but complex enough to produce less error

Implications of regularization on accuracy:

1) Regularized model will be having medium accuracy.
2) This is because regularized model generalizes the data and does not over fit.

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?

1) I'll choose Lasso because of the below reasons
2) Ridge – Lambda- 70, Lasso- Lambda- 0.007(Evaluated from my regression analysis)
3) Lambda of lasso is very less when compared to ridge. So at fewer lambdas I'm attaining good model and as lambda is small error will also be very small.