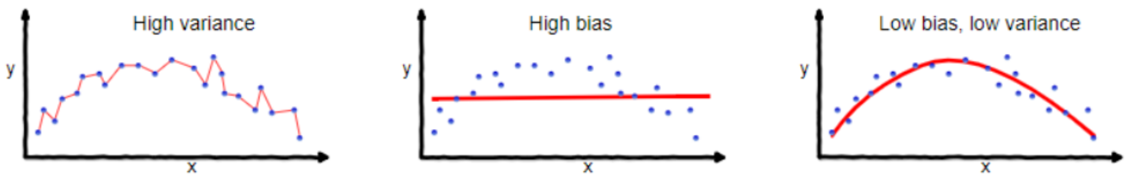
Question 1

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?

Ans:

- This happens because of overfitting, the model tries to fit each and every point on the training data set therefore it fails to generalize the data because of it fails to predict on test data set,
- There is nice analogy related to overfitting suppose there are 2 students A and B, A tries to mug up all the answers for the questions given by teacher to study instead of understanding the concepts while student B studies the concepts nicely and also mugs up little where ever necessary and concentrates also on concepts.
- Out of these 2 student A will perform better than B if you give same questions as in training but in test student B will perform better this same scenario is true for the above situation where model has mugged up almost every thing during training so it performs obviously well on training data set but when test data set is given it fails to predict accurately.

Therefore we need to find the model which maintains the good balance between underfit and overfit as shown in the figure:



overfitting underfitting Good balance

Good Balance helps the model to generalise better as well as fit the data points better so it performs better on test data set also.

The main is to make the model simpler not to simple also, There are many methods used to achieve this

1.cross validation

2.using models which gives penalty on increasing the complexity eg - lasso, ridge

3.using accuracy measures of models which gives penalty for using complex models eg-adjusted r2

4.Training with more data

5. Removing correlated features

6.Removing insignificant features which does not contribute much to model

Question 2

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

Ans2:

1. Lasso regression(L1):

    a. Adds absolute value of coefficients as penalty to the cost function or loss function

    Formula:

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

Cost function

    b. Lasso shrinks less important features to 0's thus helping in feature selection.

    c. Lasso is more unstable i.e. slight change in lambda leads to noticeable change in model accuracies(r2,mse,etc).

    d. Lasso is more computational expensive since there is no closed form solution available for it

2. Ridge Regression(L2):

    a. Adds square of the absolute value of coefficients as penalty to the cost function or loss function

Formula:

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

Cost function

    b. Shrinking of less important features to 0's does not happen.

    c. Ridge is stable i.e. slight change in lambda doesn't lead to noticeable change in model accuracies(r2,mse,etc).

d. Ridge is more simpler when compared to lasso since there is closed form solution available

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?

Ans3.

2 Linear models are developed L1 and L2 which performs equally well on testing data set. Looking at the coefficients of L1 and L2, L1 model coefficients are bit complex compared to L2 model coefficients. The representation complexity of L1 is higher than that of L2 since we need greater number of bits to represent the value *32.648628* and to retain its precision.

In L2 model we have 1 digit after the decimal point, Hence number of bits required to represent single decimal digit :it ranges from 0 to 9 Hence maximum of 4 bits required to represent the value 0 to 9

In L1 model we have maximum of 6 digit after decimal it ranges from 000000 to 999999 which requires 20 bits to represent the number

So L1 model is more complex than L2 .

L2 is preffered because its much simpler but gives same performance as L1

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?

Ans4:

Robustness: The robustness is the property that characterizes how effective your algorithm is while being tested on the new independent (but similar) dataset. In the other words, the robust algorithm is the one, the testing error of which is close to the training error.

•Generalizable: How well learning algorithm generalizes on the training data(its like learning the concepts of the subject not mugging up) so that it can apply well on the test data.

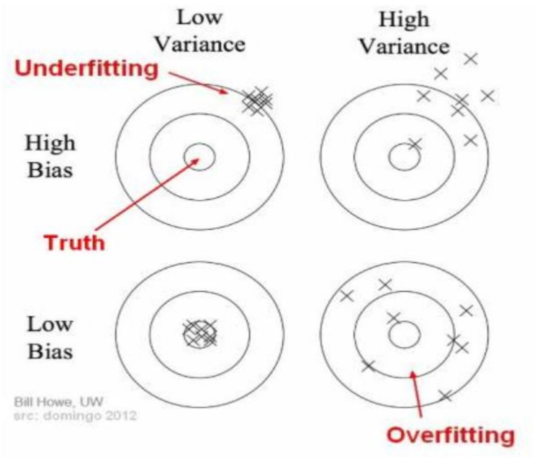
•Bias in Machine Learning is defined as the phenomena of observing results that are systematically prejudiced due to faulty assumptions. However, without assumptions, an algorithm would have no better performance on a task.

•Variance: is the sensitivity of the model to the small change in training dataset.

•Here if the model makes approximations of the values to be predicted i.e. it generalizes the training dataset therefore bias will be high but variance will be low ,the model will be more robust.

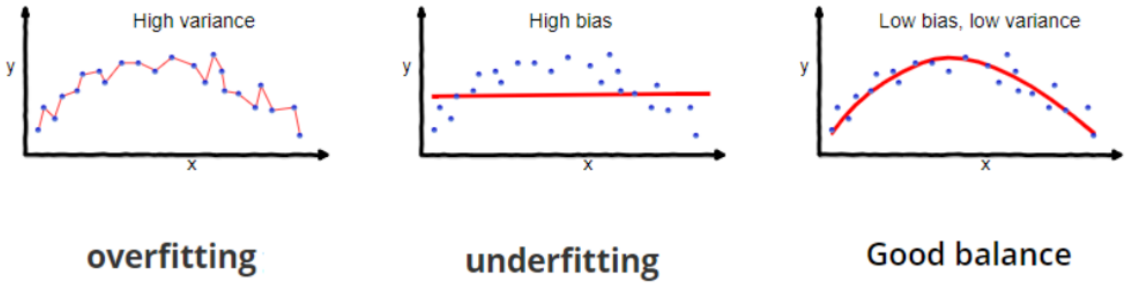
•Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data

•Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.



•There is always trade off between bias and variance so we need to chose the optimum cutoff between the two. The accuracy of the model also increases on test data.

•If the model is overfit then its going to have high variance but low bias, if model is underfit then its going to have high bias but low variance,we need to choose the middle of these 2.



There are many methods used for finding optimum cutoff between bias and variance

1.cross validation

2.using models which gives penalty on increasing the complexity eg - lasso, ridge

3.using accuracy measures of models which gives penalty for using complex models eg-adjusted r2

4.Training with more data

5. Removing correlated data

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?

Ans5:In my case I would be choosing lasso because lasso model shrined many coefficients of variables to 0 making the model simpler which is one of the feature of lasso but retaining the same or even higher accuracy than ridge but lasso takes more time to learn since it is computational expensive.