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:

This clearly seems to be a problem of overfitting. The model is remembering the data instead of learning from it. The model has memorized the training data and is not able to differentiate between the actual pattern and the noise in the dataset.

We can use following techniques to solve the overfitting problem:

1. Bias-Variance tradeoff: We should select a model which is simple enough to not over-fit but complex enough to have low variance. Model should be able to have both low bias and complexity.
2. Regularization: Model should be penalized for the increase in complexity. We can use different regularization techniques like Lasso, Ridge regression which penalizes for increasing the complexity, hence finding a model which learns from data instead of just overfitting.
3. K-Fold Cross Validation: We can use initial training dataset to generate multiple mini-test train splits and using these splits we can train the model. In K-Fold cross validation we partition the training set into k subsets (folds). We train the model on k-1 folds and test it on the remaining fold. We do this iteratively and hence reduce overfitting.

Question 2:

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

Answer:

Regression Model using L1 regularization technique is also known as Lasso Regression and Model using L2 regularization is Ridge Regression.

Main differences between these regression techniques:

1. The penalty term which is applied in the cost function.
   Ridge Regression adds square magnitude of coefficient as penalty term to the loss function.

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

Cost Function for Ridge Regression
with p features

Lasso adds absolute value of magnitude of coefficient as penalty term to the loss function.

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

Cost function for Lasso Regression
with p features.

2. Both Lasso and Ridge regularize the coefficients by reducing them in value, essentially causing shrinkage of coefficients. Both perform different measures of shrinkage which depends on the value of hyperparameter, λ. In this process Lasso shrinks the less important features' coefficients to zero, thus performing feature selection as well. This is not the case with Ridge regression. Hence, Lasso is primarily used when we have huge numbers of features and we want to do feature selection.
3. L2 regularization's shape is like a sphere, and value of the weights try to be as low as possible to minimize the function though the values will not be zero. Whereas in L1 regularization's shape is diamond like and the weights are lower in the corners of the diamond. These corners have one of the features as 0, thus leading to sparse matrices. So, in other words if there are multi-colinear features than Ridge tries to distribute the weights across by keeping it minimum. Whereas Lasso, chooses 1 feature randomly and makes other co-linear features zero.
4. Ridge is computationally efficient as it has analytical solutions. Lasso can be computationally inefficient in case of non-sparse cases.

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:

Both of these models have same number of features = 1.

Based on Occam's Razor, we should pick the simpler model if two model shows similar performance on finite test or training dataset.

Here L1 is more complex than L2 based on the size of best possible representation of the model. L1 uses more number of bits in binary encoding of the model (it has too many bits for precision). Both constant term and coefficient will take less space and easier to compute for different values of x.

Hence it is more complex and we would prefer L2.

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?

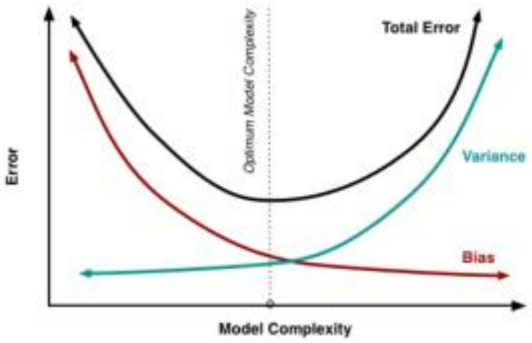
Answer:

We should try to make the model as simple as possible, so that it is robust and generalizable. Simpler models require fewer training sample for effective training. They are more robust as they are not sensitive to the specifics of the training dataset as their more complex counterparts. They try to learn from data rather than learning the dataset itself.

Simple models have low bias and high variance. Bias is the deviation from expected and variance refers to the variance in the model.

Simple model makes more errors in the training dataset. Complex models lead to overfitting as they work well with training dataset but fails when applied to test samples. We should choose a model which has similar performance across both test and training sets.

This is called bias-variance tradeoff. We would try to build a model which is simple enough to have both bias-variance to be minimum. Based on the following curve optimum model complexity should lie at the intersection point of the bias variance tradeoff. This is the model which we should choose.



We also use regularized regression (Ridge/ Lasso etc.), which imposes penalty when we add mode features to the model. So, by getting an optimum value of hyperparameter for regularized regression we can take care of Bias-Variance tradeoff, which helps us to make model more robust and generalizable.

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:

Here in this dataset we can see that both the Ridge regression and the Lasso regression have almost same value of R-Square.

Our dataset size is also small, so computation time isn't relevant here.

We can see that Lasso produces similar accuracy with lesser features (116 non zero for alpha = 55) vs 211 (non-zero features for alpha = 3) features present in the Ridge regression. Hence Lasso, produces a simpler model giving same level of accuracy.