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

The main idea of the model is to see how well the model is performing on unseen data i.e., test data. These may cause two types of issues underfitting and overfitting.

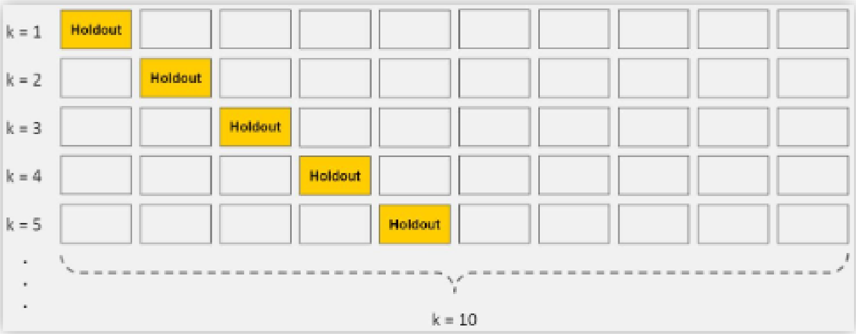
The problem mentioned in the question is the case of overfitting where the model has learnt all about the training set and performing poor on the test set.

The various ways of avoiding overfitting are:

1. Cross-validation
2. Train with more data
3. Remove features
4. Regularization
5. Occam’s Razor

Cross-validation: We divide the train set into training and validation set to tune the model. In k-fold cross validation technique we divide the dataset in k subsets called as folds and iterate on k-1 folds as train set and remaining folds as test set.

This technique is used to tune the hyper parameters and the original test set is unseen to the model.



K-Fold Cross-Validation

Train with more data: If the data is unclean and noisy then this technique will not work out therefore the data fed must clean and relevant to the dataset.

Remove features: Removing irrelevant features from the training data set will make the model simpler and less complex.

Regularization: It is the process to create the model which as simple as possible while performing well on the training set.

There are two type regularized regression:

1. Ridge Regression
2. Lasso Regression

Occam's Razor: A predictive model has to be simple but not simpler. As there are various advantages of using simpler model few of them are:

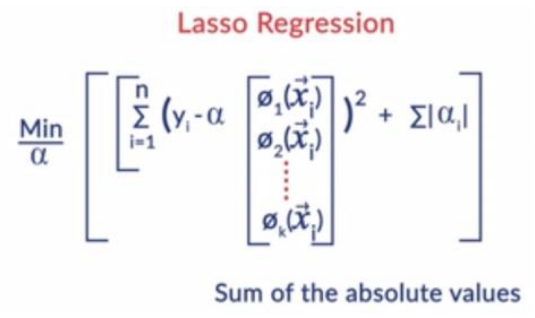
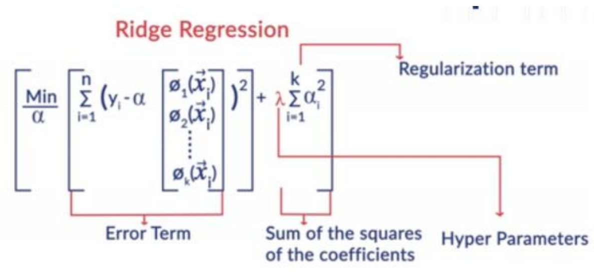
- Simpler model are more generic
- Simpler model requires less training data
- They are more robust

Question 2
List at least four differences in detail between L1 and L2 regularisation in regression.

Answer:
The technique used to deal with overfitting and feature selection is:

1. L1 Regularization or Lasso Regression
2. L2 Regularization or Ridge Regression

In Regularized regression there are two terms: error term and regularization term. Key differences between L1 and L2 regularizations are:

Lasso Regression	Ridge Regression
In lasso, an additional term of "sum of absolute value of the coefficients" is added to the cost function along with the error term  Lasso Regression Min α $\left[\left[\sum_{i=1}^{n}\left(y_i - \alpha \begin{bmatrix}\phi_1(\vec{x}_i)\\ \phi_2(\vec{x}_i)\\ \vdots \\ \phi_k(\vec{x}_i)\end{bmatrix}\right)^2 + \sum\lvert\alpha_i\rvert\right]\right]$ Sum of the absolute values	Here, an additional term of "sum of square of coefficient" is added to the cost function along with the error term.  Ridge Regression Regularization term Min α $\left[\left[\sum_{i=1}^{n}\left(y_i - \alpha \begin{bmatrix}\phi_1(\vec{x}_i)\\ \phi_2(\vec{x}_i)\\ \vdots \\ \phi_k(\vec{x}_i)\end{bmatrix}\right)^2 + \lambda\sum_{i=1}^{k}\alpha_i^2\right]\right]$ Error Term Sum of the squares of the coefficients Hyper Parameters
Lasso regularize the coefficients by reducing them in values, causing shrinkage of the coefficients. It shrinks some of the variables to zero thus helping in feature selection.	Ridge regularize the coefficients but does not shrink them to zero.
When weight of the input features are closer to zero then that leads to sparse solution.	It makes the weight close to zero but not zero therefore the solution is not sparse.
They are robust to outliers	They are not robust to outliers as the square

	terms increase the error differences of the outliers while regularization term tries to penalize the weight.

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:
L1 is much simpler as compared to L2 therefore I would prefer L1.
The reason behind choosing L1 is that bit required by L1 model will be less as compared to L2. Since the model is performing equally well on the test set so I would consider the model which occupies less bit as compared to the other.

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:
Robustness means the property of the model according to which whether the model is tested on the training set or test set, the performance is same.
Generalizable means the model has learnt with few variables but are more likely to predict on the unseen data.
The model can robust and generalizable by following the **Occam's Razor** principle, which says the model should be simple but not simpler.
Advantages of simpler model are:

- Models are more generic: if the model has learnt every feature then that model can answer the question which is similar. But if the question is unfamiliar the complex model is more likely to make error.
  The simpler model just learns the basic principles and when the unfamiliar question arises they are less likely to make error as compared to the complex model.

- Models are robust: Complex models are more sensitive to the data set. As the training set changes they are more likely to swing, which is not in the case of simple models.

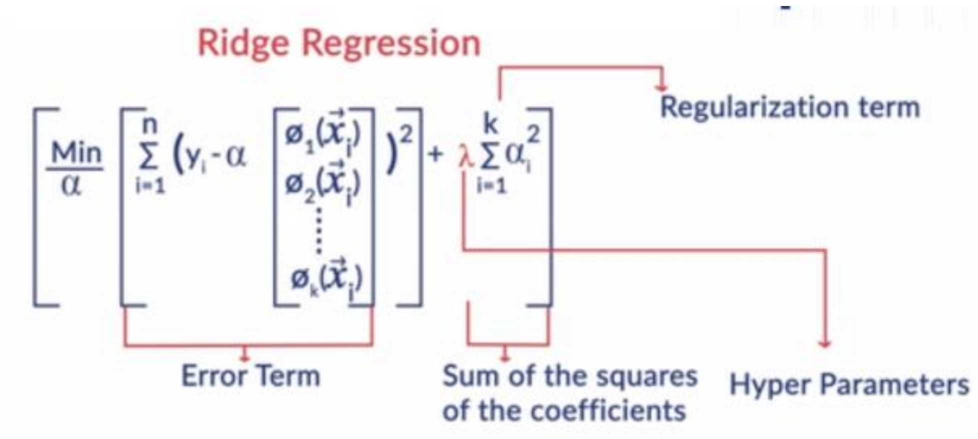
Thus we can say, **Complex models have high variance and low bias and Simple models have high bias and low variance**.
Variance means variance in the model

Bias means deviation from the expected, ideal behavior.

But simpler models make more error in the training set. As the simpler model have high bias they are more likely to make error in predicting the expected output and therefore accuracy is not much compared to the complex model.

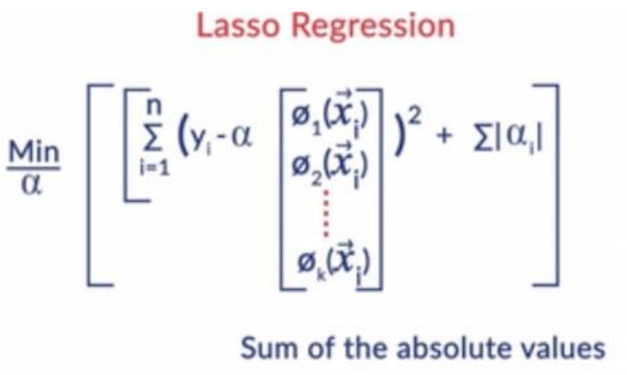
Question 5
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer:
The optimal value of lambda for ridge regression (L2) is 20 and for lasso regression (L1) is 500
The important factor is the penalty term.
The cost function for L2 is:


Ridge Regression
Regularization term
$\underset{\alpha}{\text{Min}} \left[\left[\sum_{i=1}^{n} \left(y_i - \alpha \begin{bmatrix} \phi_1(\vec{x}_i) \\ \phi_2(\vec{x}_i) \\ \vdots \\ \phi_k(\vec{x}_i) \end{bmatrix} \right)^2 \right] + \lambda \sum_{i=1}^{k} \alpha_i^2 \right]$
Error Term
Sum of the squares of the coefficients
Hyper Parameters

L2 has squared magnitude as the penalty term. If the value of lambda is zero then the cost function will have normal error term. Increasing the value of lambda puts too much weight on the model and leads to under-fitting.

The cost function of L1 is:


Lasso Regression
$\underset{\alpha}{\text{Min}} \left[\left[\sum_{i=1}^{n} \left(y_i - \alpha \begin{bmatrix} \phi_1(\vec{x}_i) \\ \phi_2(\vec{x}_i) \\ \vdots \\ \phi_k(\vec{x}_i) \end{bmatrix} \right)^2 + \sum |\alpha_i| \right] \right]$
Sum of the absolute values

Here, the penalty term is absolute value. If the value of lambda is zero then the cost function is just the error term. Increasing the value of lambda makes the model under fit.

Since the number of features in the assignment is more than 200 therefore I would want the lambda to be large therefore I would choose Lasso regression as it does features selection also by shrinking the coefficients equal to zero.