Assignment Solution (Part - 2)

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.

The model built by Rahul has succumbed to the perils of overfitting. It has tried to memorise the training data making it less generalisable and with high variance leading to poor test performance. The problem can be solved using regularisation. Both L1 and L2, that is, Lasso and Ridge can be used to fight the problems of overfitting.

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

L1 (Lasso)	L2 (Ridge)
This can be used for feature selection.	This can not be used for feature selection.
The regualrization term contains the sum of absolute values of the coefficients.	The regualrization term contains the sum of squares of the coefficients.
Computataionally intensive compared to Ridge.	Ridge is computationally less intensive than Lasso.
The solution can not be determined using a simple matrix function.	The solution can be determined using a simple matrix function.

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?

L2. Because it is simpler and hence easy to generalise and more robust.

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?

The model should be as simple as possible. This might lead to a decrease in training accuracy but this will be make the model more robust and easy to generalise. This can also be understood using the Bias variance tradeoff.

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?

From the output of both models, it is evident that Lasso does feature selection and its output is much simpler than Ridge's, without any compromise in the accuracy over the test data. So, I will go with the lasso regression.