Advance Regression Assignment (Part-2)

Question-1:

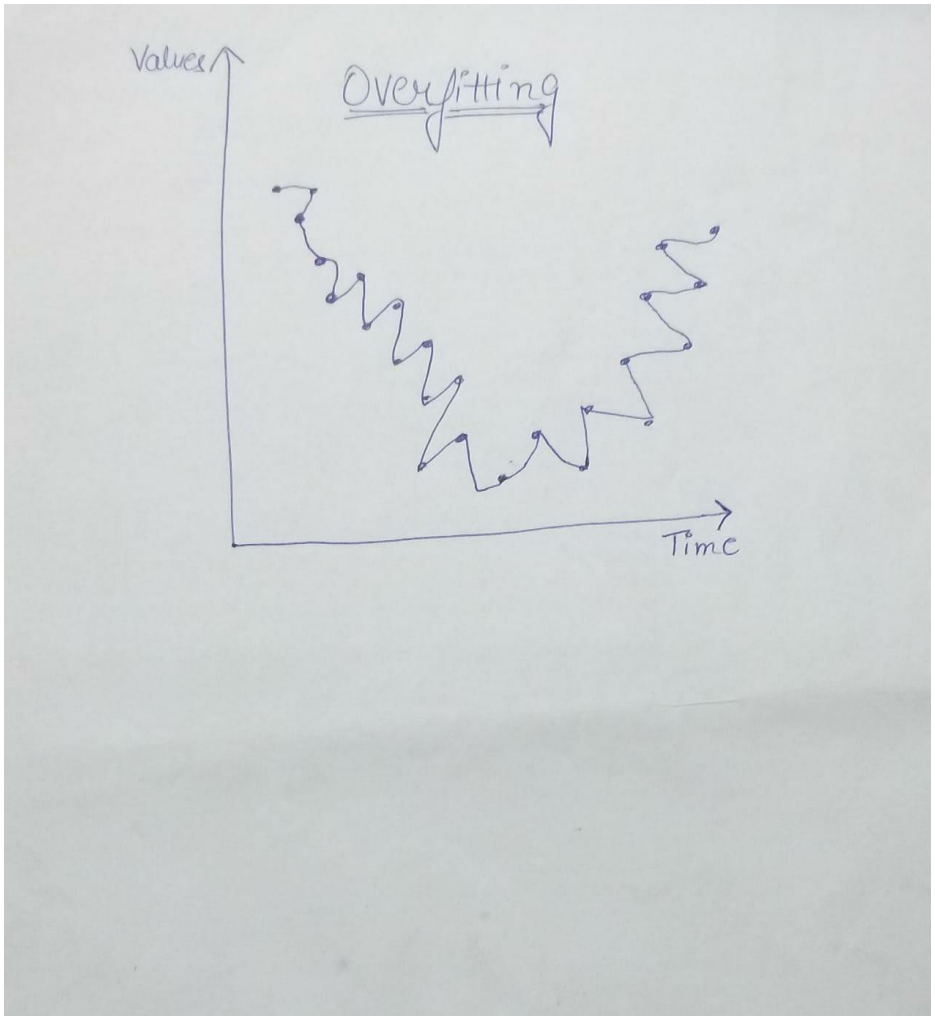
Rahul built a logistic regression model having a training accuracy of 97% while the test accuracy was 48%. What could be the reason for the seeming gulf between test and train accuracy and how can this problem be solved.

Solution-1:

The reason for this is OVERFITTING.

This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. It makes it less generalizable with high variance leading to poor test performance.

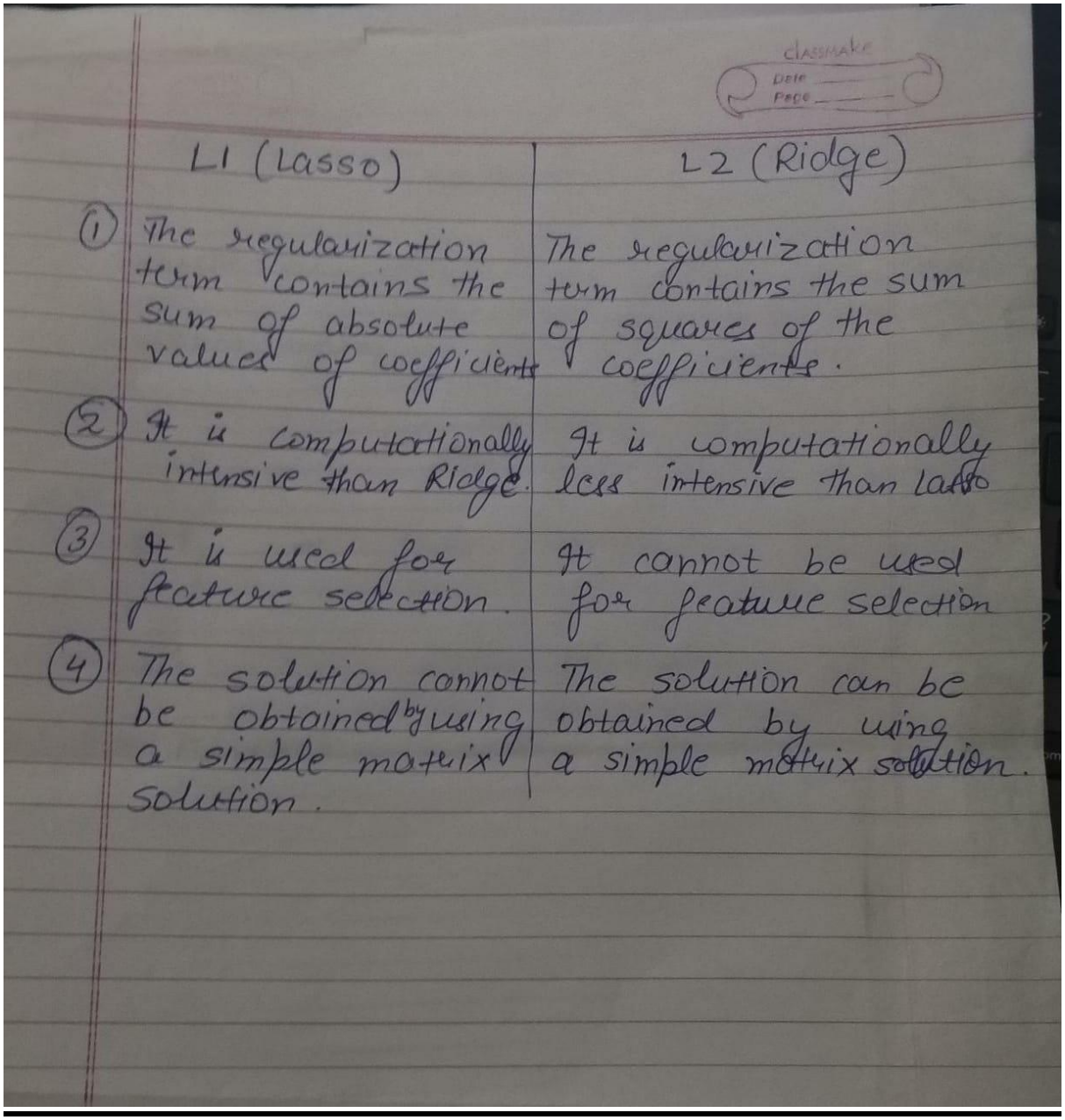
This can be resolved by REGULARIZATION. Both RIDGE and LASSO can be used for this.



Question-2:

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

Solution-2:



	L1 (Lasso)	L2 (Ridge)
(1)	The regularization term contains the sum of absolute values of coefficients	The regularization term contains the sum of squares of the coefficients.
(2)	It is computationally intensive than Ridge.	It is computationally less intensive than Lasso
(3)	It is used for feature selection.	It cannot be used for feature selection
(4)	The solution cannot be obtained by using a simple matrix solution.	The solution can be obtained by using a simple matrix solution.

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?

Solution-3:

I would prefer L2.

Reasons for this –

1. It is simpler
2. It is robust
3. It is easy to generalize

Question-4:

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Solution-4:

To keep the model robust and generalizable, we need to keep it as simple as possible.

More complex models are not good.

Simpler model leads to decrease in training accuracy but it maintains the robustness of the model and keeps it generalizable**.**

Let's understand it using BIAS-Variance Trade off

If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without over fitting and under fitting the data.

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?

Solution-5:

I will prefer Lasso regression as it helps in feature selection by making most of the coefficients zero.

There is no compromise in accuracy and also has simpler output.