Hari Thapliyal
Date: 23-Sep-19
Assignment: ML2-Module3

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
This could be overfitting problem. Model has learned almost everything from the train dataset and in real environment, about which model is completely unaware, is not able to figure out the answer. In ML language we also call this high variance low bias situation. Variance which is an error and measured on the test data is high, but bias which is another kind of error is measured on train data is low.

Second reason this problem can be because dataset has some extremes/ outliers. And model has learned those outliers as well during the training.

Third possible reason is there are columns in the dataset which have multicollinearity and model has learned from those features as well, and now in the test dataset not able to figure out which column contribute how much (in terms of coefficient) and failing to predict.

To overcome this problem we need to do following

1. Ensure outliers are removed from the training dataset. Ensure that every column is scaled either on standard scale or normal scale.
2. Check the relationship between different variables and if multiple columns share high correlation then keep only one out of those many.
3. Finally have some penalty for the overfitting.

How to penalize the over fitting?
Instead of using simple regression algorithms we should advance regression techniques like lasso, ridge or elasticnet for this kind of problem solving. These algorithms put penalty for too much learning, which is not helping in prediction. This penalty put is called lambda. In the cost equation function more the value of lambda means more the penalty for including extra variables.

We need to keep in mind if lambda is zero it means there is no penalty so model can include all the variables and try to remember all the data points. In this situation, the final model will have high variance and low bias. But, if Lambda is high towards infinity it means then there is a heavy penalty for single variable learning. So generalization is penalized. In this situation model will have low variance but high bias. So do a tradeoff of lambda.

Question 2

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

Answer

L1 Regularization is also called Lasso Regression. L2 Regularisation is also called Ridge Regression.

	Lasso (L1 Regularisation)	Ridge (L2 Regularization)
Cost Function Terms before λ is called Error Term and terms after λ including λ is called regularization term. yi is actual, second part of first terms is predicted value. λ is a penalty. β is called vector of coefficients. There are p number of feature/variable in the model There are n number of datapoint or rows in the training dataset	$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \left\|\beta_j\right\|$	$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$
Penalty Term	Absolute value of magnitude of coefficients	Square value of magnitude of coefficients
Feature Selection	If you increase the λ then some coefficient will become insignificant i.e. zero and they will be dropped from the model. So lasso helps in feature selection in automatic	If you increase the λ then some coefficient will insignificant but they never become zero. So if you want to exclude feature you need to manually decide which

	way.	feature to drop and which not.
Memory Consumption/ Performance/ Processing Power	Lasso is slow and takes more time to converge.	Ridge is faster compared to Lasso.
Small λ value	If behaves like OLS (ordinary least square). Lead to over fitting.	If behaves like OLS (ordinary least square). Will lead to over fitting.
Very high λ value	Lead to under fitting	Lead to under fitting.
Number of features can be handled	Thousands of features can be handed smoothly and you get know to important variable very easily.	Thousands of features can be handled smoothly but you need to remove low coefficient variable manually

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

Simplest model should be selected. Simplest is that one which takes the least memory and has least processing requirement among the candidate models.

L2 is simpler in terms of coefficient and intercept, because it need lessor memory space and lessor processing power because lessor post decimal digits.

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
To make sure that the model is robust and generalizable we need to make sure that model is no trying to remember each an every data point from the train dataset. This happen when we set lambda value to 0. It means there is no penalty for including a feature in model so add as much you want.

At the same time we need to remember, if we set lambda to very high then model will not be able to learn anything. Because penalty of including any variable is high therefore only few or almost no feature will be part of the model except intercept.

Therefore we need to choose a right value of lambda. To achieve this we can use GridSearchCV with various values of lambda and plot the results. At whatever level of lambda we see there is no increase of accuracy no matter how many variables we add, we can pick that alpha for the final 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
Optimal Value of Lambda for L1 (Lasso) is 180
Optimal Value of Lambda for L2 (Ridge) is 15
Optimal Value of Lambda for ElasticNet is .01

After cleaning the data, creating dummy variable etc final dataset has 234 features. After I run lasso I get only 81 important features which can predict the sales price. If we are not using lasso then dropping 234-81 = 153 features is too much time consuming. That too when we do not know exactly in what sequence we need to drop them.

Therefore I would choose a model created by lasso regression model.

Finally, even lasso model has many features which has high multicollinearity therefore we need to drop those feature one at a time using RFE, VIF and p-value.