Statistical Methods in AI	Date:	*10/03/2019*
Instructor: *Dr. Ravi Kiran*	Submitted By:	Tarun Mohandas (2018201008)

Assignment 7: Regulatization and Cross-Validation

1 Question 1

In this assignment, we train a linear regression model on admission dataset. The following is the sample of that dataset.

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	242	317	103	2	2.5	2.0	8.15	0	0.65
1	334	319	108	3	3.0	3.5	8.54	1	0.71
2	4	322	110	3	3.5	2.5	8.67	1	0.80
3	45	326	113	5	4.5	4.0	9.40	1	0.91
4	232	319	106	3	3.5	2.5	8.33	1	0.74

Figure 1: First rows of the admission dataset

1.1 Lasso Regularization

Lasso regularization (L1 regularization) for Linear regression is implemented using the following expression for loss function.

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m}(\hat{y} - y)^2 + \lambda\sum_{i=1}^{n}|\theta|$$

where $\lambda\sum_{i=1}^{n}|\theta|$ is the regularization term and λ is the regularization parameter and θ is the coefficients of the hypothesis.

Since the gradient in gradient descent is derivative of the loss function, the regularization parameter should also be differentiated w.r.t θ. The following is the gradient of the loss function with the regularization term.

$$\frac{\partial L(\theta)}{\partial\theta} = \frac{2}{m}\sum_{i=1}^{m}(\hat{y} - y)x + \lambda\sum_{i=1}^{n}\frac{\theta}{|\theta|}$$

The following graph shows how the regularization parameter λ affects the performance of the model. Here the cost of the model is plotted w.r.t λ.

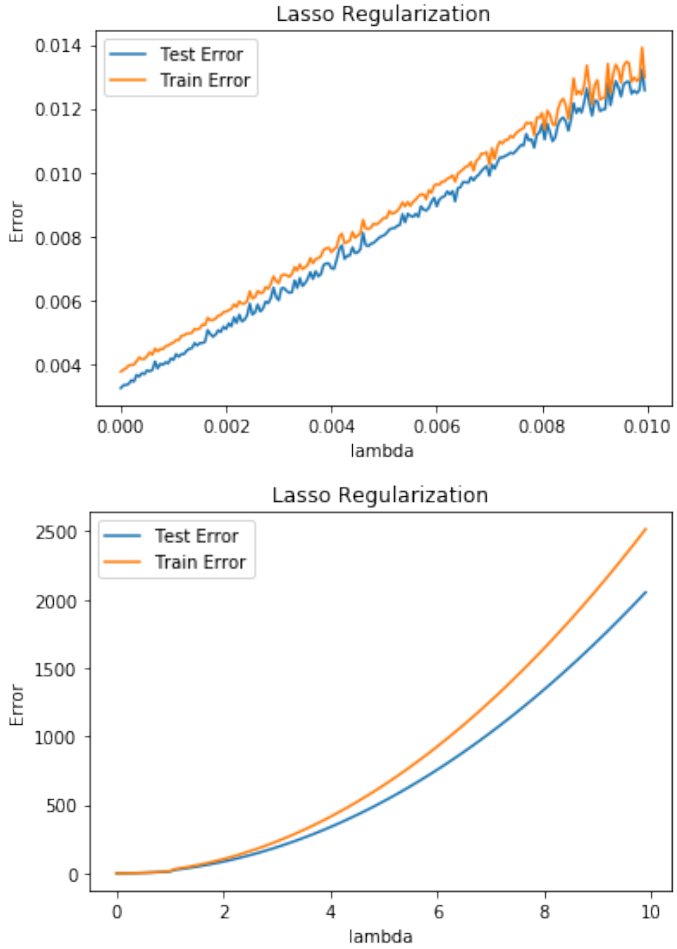

Figure 2: Effect of regularization parameter λ on model performance in case of Lasso Regression

The first figure shows the performance on a very small scale of λ. On very small change in the regularization parameter, the error appears to be steadily increasing linearly as the λ increases. But if the error is compared to λ on a big scale, it appears to be increasing quadratically. The error is increasing because the regularization parameter reduces the overfitting in data and decreases the bias. Therefore, the larger the regularization parameter is, the larger the error.

1.2 Ridge Regularization

Ridge regularization (L2 regularization) for Linear regression is implemented using the following expression for loss function.

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m}(\hat{y} - y)^2 + \lambda \sum_{i=1}^{n} \theta^2$$

Since the gradient in gradient descent is derivative of the loss function, the regularization parameter should also be differentiated w.r.t θ. The following is the gradient

of the loss function with the regularization term.

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{2}{m}\sum_{i=1}^{m}(\hat{y} - y)x + 2\lambda\sum_{i=1}^{n}\theta$$

The following graph shows how the regularization parameter λ affects the performance of the model. Here the cost of the model is plotted w.r.t λ.

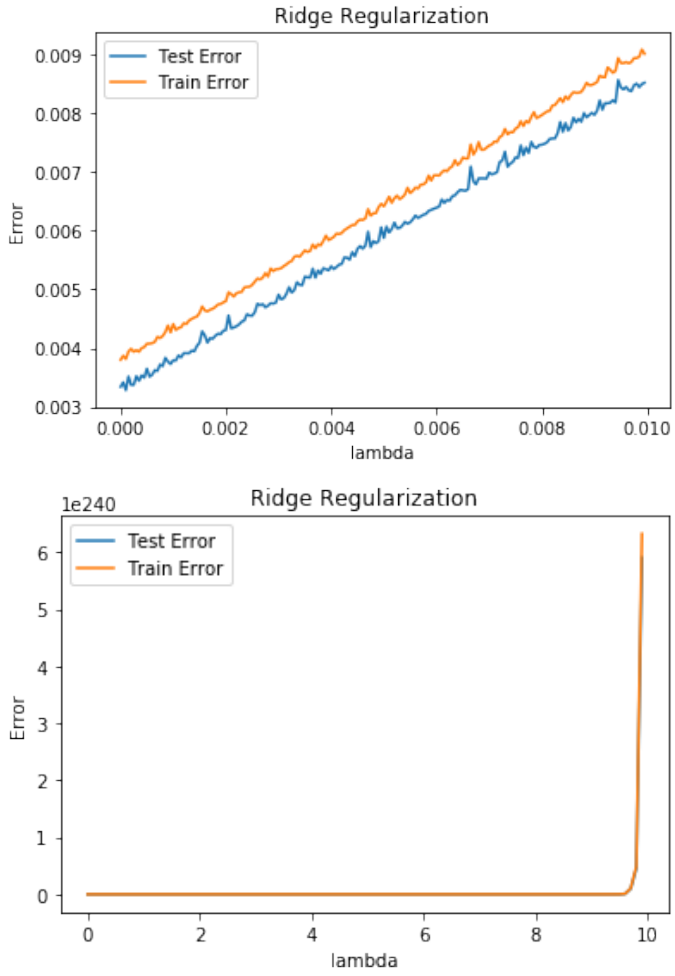

Figure 3: Effect of regularization parameter λ on model performance in case of Ridge Regression

The first figure shows the performance on a very small scale of λ. On very small change in the regularization parameter,just like in the case of lasso regression, the error appears to be steadily increasing linearly as the λ increases. But if the error is compared to λ on a big scale, it appears to be increasing exponentially. The sudden spike in the error is just relative because on a very small scale, they appear to increase exponentially.

1.3 Analyse how the hyper-parameter λ plays a role in deciding between bias and variance.

Usually we use regularization hyper-parameter λ when we have to prevent overfitting. High-variance of a model tends to overfitting of the data. Therefore if the model has high variance it can be rectified by having a relatively high value of λ. Therefore, the higher the value of λ the higher is the bias and lower the value of λ the higher the variance.

In other words, λ is directly proportional to bias and inversely proportional to variance.

1.4 Analyse how the two different regularisation techniques affect regression weights in terms of their values and what are the differences between the two.

L1 regularization tend to shrink some weights to zero. This is because the order of the regularization expression is almost the same as the order of the weights. Since it drives some of the weights to zero, some features associated with those weights will never contribute to the model. Due to this reason, Lasso regularization (L1) is used as a feature selection technique.

L2 regularization will force the weights to be relatively small. The higher the regularization parameter is, the more it will shrink the weights but it wont drive any weights to zero. Due to this reason, L2 (Ridge regularization) is preferred over L1 most of the times.

1.5 Cross validation

1.5.1 K-fold Cross validation

K-fold cross validation is a technique by which we divide the dataset into K parts, keep one part as test set at a time and the rest of the parts as train set and analyze the performance. This is done K times each time having a different part as test set and rest of it as train set.

The following graph shows the average error w.r.t the K value we choose to split the dataset.

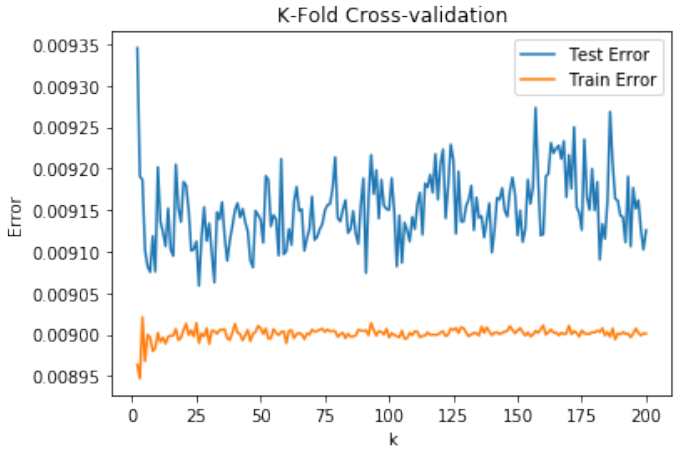

Figure 4: Performance of the model with respect to value of K

As we can see from the figure, the value of K is very important when selecting a model and it changes w.r.t dataset. The training error will remain a constant almost all the time, but the true error will be given away by the test error. We can see that the test error is very unstable w.r.t K. Therefore based on the model, and cross validation, we need to select appropriate K so that we can split the data appropriately to train the model.

1.5.2 Leave One Out Cross Validation

Leave One out cross validation is the special case of K-fold cross validation where the data is split into two parts, one part containing just one data point and the other part containing all the other data points. The testing is done with the one data point and the training is done with the rest of them. This is repeated for every data point taken as test data. The following is the performance of the model for Leave one out cross validation.

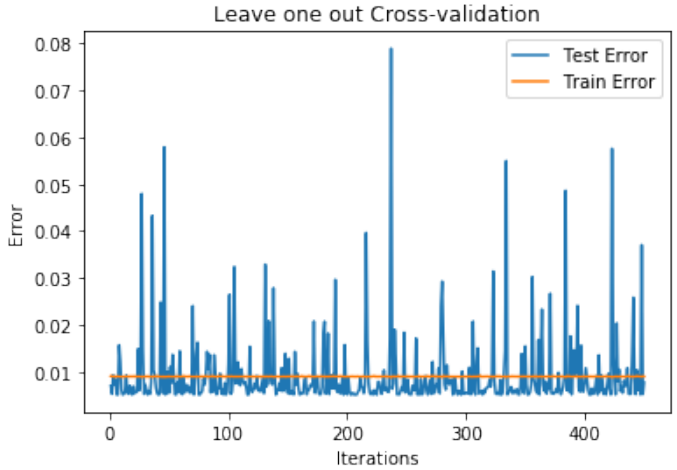

Figure 5: Performance of the model with leave one out cross validation