MACHINE LEARNING B-555
PROGRAMMING PROJECT -2
REPORT

Task 1 : Regularization

Regularized Linear regression was performed on the 5 datasets and their plots are given here for reference. The MSE's given for the hidden functions match the MSE found in the graphs for each dataset.

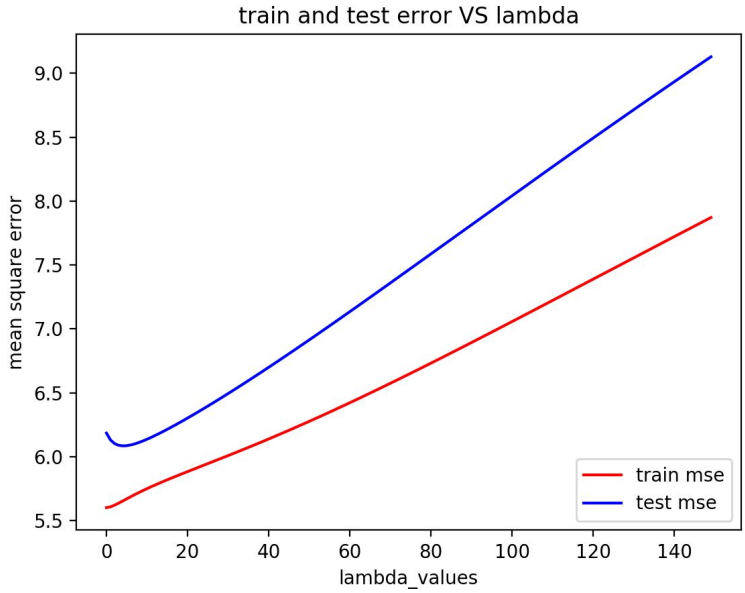


Fig : 100-10 data set

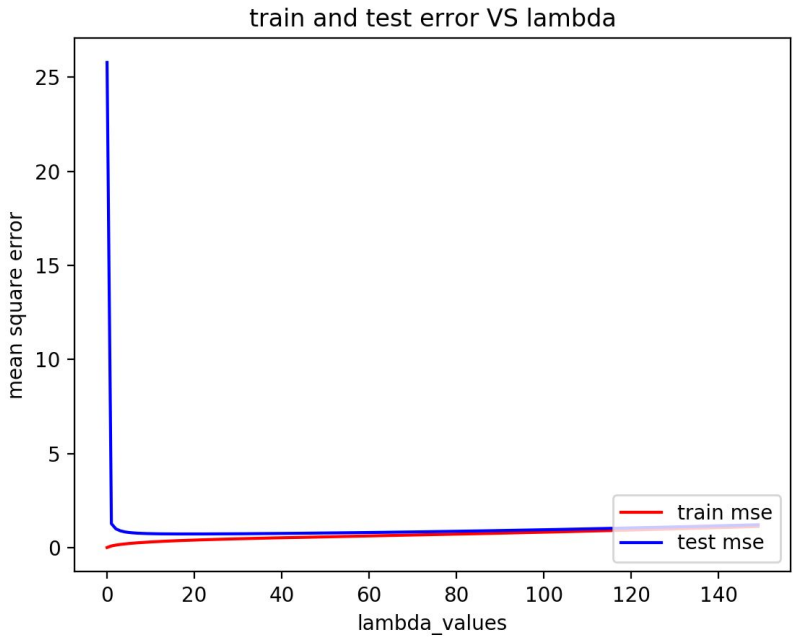


Fig : 100-100 dataset

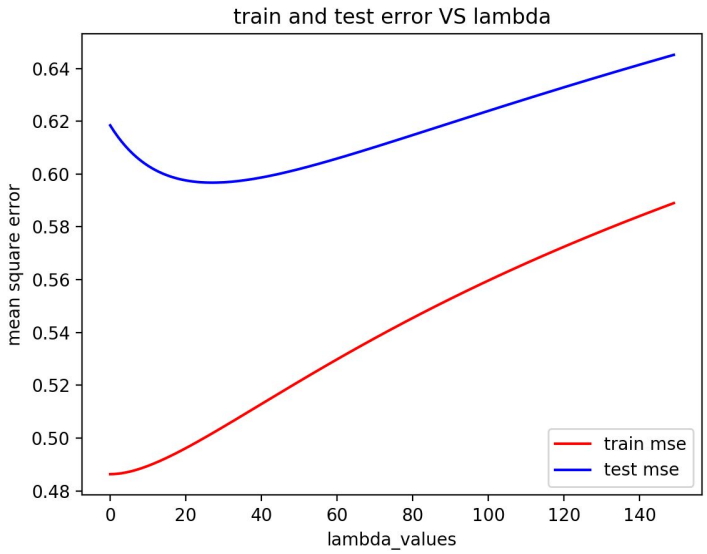


Fig : 1000-100 data set

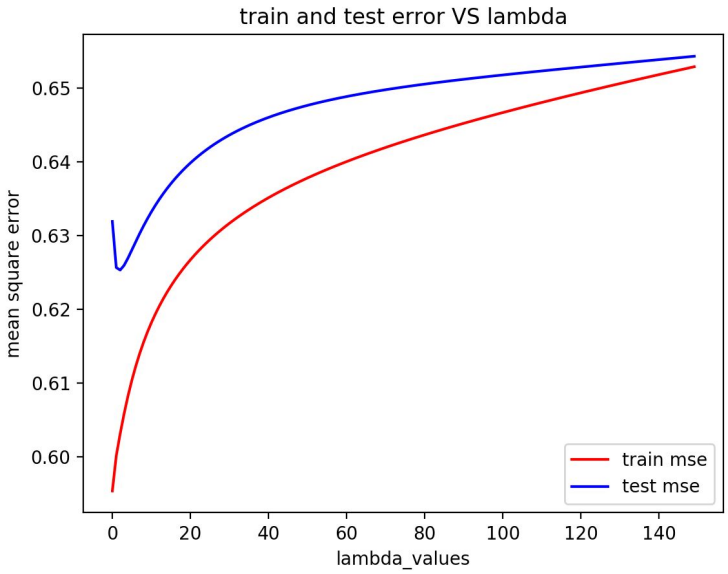


Fig: wine dataset

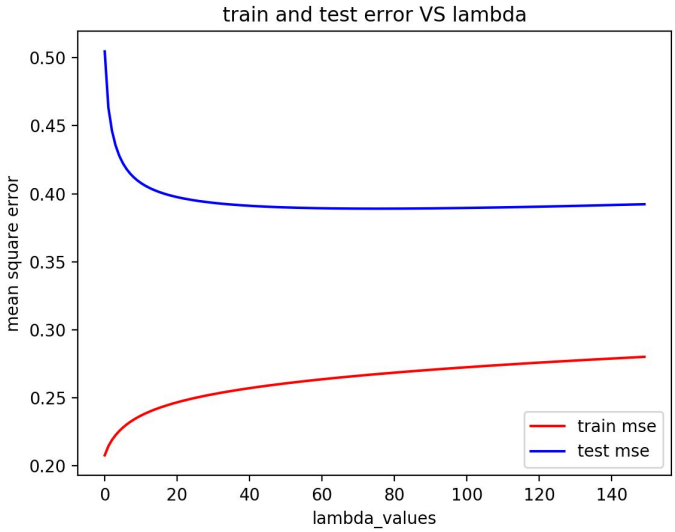


Fig : crime dataset

Q) Why can't the training set MSE be used to select 'lambda'?

A) Training set MSE cannot be used to set 'lambda' as it would not depict correctly on unseen data that is test data. The lambda we get from the training data causes overfitting in the test data which is shown clearly. MSE for train data cames at lambda= 0 but at the same lambda MSE of test set is very high. So, it is not a good choice to select lambda based on MSE of training data as it would not be good enough for predictions of test data.
To solve this issue of overfitting, regularization is used.

Q) How does lambda affect error in test set? Does this differ for different datasets?How do you explain these variations?

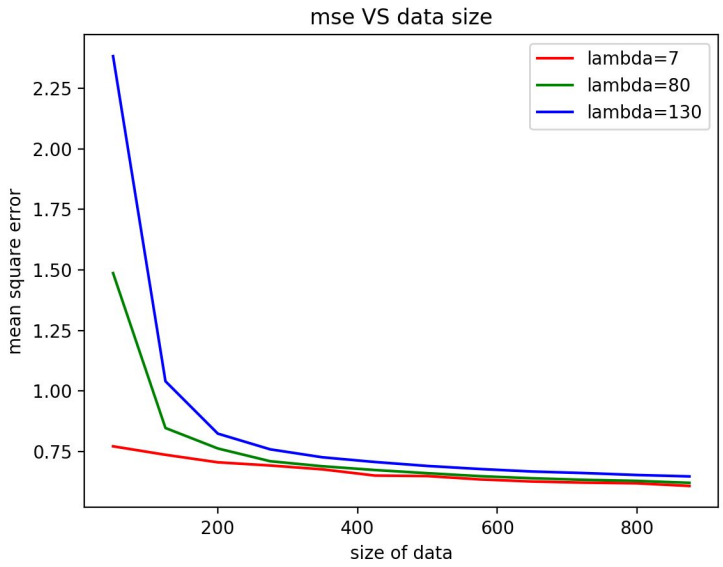
A) We can see that as lambda increases the MSE first decreases and then increases or stays the same. At a particular lambda the MSE is low, this is where the regularization parameter helps in minimizing our error and avoids overfitting of our model.

Yes, for different datasets lambda affects differently on them. But when lambda is very close to 0 the pattern is the same for all the data sets. I.e high MSE at low lambda value.

From the plots we can see that for the best lambda(where MSE is low) is higher when the number of features of the dataset increases. This basically means that having a lot of features may not help us in getting a better model (bias becomes low and variance becomes high). When this happens, our model will lead to overfitting and to counter this affect regularization (lambda) comes into play and may take a high value if a lot of features exist for a dataset.

Task 2 : Learning curves

Took the lambda values as 7(too low) , 80(just right) and 130(too high) and plotted the graphs of each of them with size of data.



As the lambda increases, there is higher drop in value of MSE as data set size increases, and there is no much difference when the data set size increases.
We know that when the data size is low (there is high bias and low variance) so it is more of a underfitting case than overfitting so high lambda values are not desirable(MSEs become high). Small lambda values can take care of regularization better.

At large data sets it doesn't matter as MSE come down drastically for any lambda.

Task 3.1 : Model selection using cross Validation
The results of model selection using cross validation are as follows:

1. **100-10 dataset**

```
min error is 6.196955677575395  and its lambda value is  15
time 0.12529128199999995
```

2. **100-100 dataset**

```
min error is 0.7051380071501065  and its lambda value is  18
time 1.5585891929999998
```

3. **1000-100 dataset**

```
min error is 0.5894285086233152  and its lambda value is  23
time 1.927794049
```

4. **Crime dataset**

```
min error is 0.33666636564609564  and its lambda value is  149
time 1.6603620930000003
```

5. **Wine dataset**

```
min error is 0.642288756624437  and its lambda value is  2
time 0.13792865899999995
```

Q) How do the results compare to the best test-set results from part 1 both in terms of the choice of lambda and test set MSE?

A) The lambda and mse values are almost comparable in both part 1 and task 3.1.

Although the run time for task 3.1 is high since we use cross validation as it iterates a lot of times before giving results.

Task 3.2: Bayesian model selection

We initialize alpha and beta, and through an iterative process calculate Sn and Mn and again find alpha and beta which will converge eventually.

The results for different data sets are :

1. **100-10 dataset**

```
the final alpha, beta values are  [[0.8824704]]   [[0.16515479]]
the corrresponding lambda value is [[5.34329273]]
similarly, the gamma value  [[6.25437698]]
number of iterations 2
6.087838984763466
```

Run time = 0.011144118000000036

2. **100-100 dataset**

```
the final alpha, beta values are  [[5.26274712]]   [[3.02701184]]
the corrresponding lambda value is [[1.73859482]]
similarly, the gamma value  [[61.59630065]]
number of iterations 1
1.004248034114374
```

Run time: 0.12367758299999998

3. **1000-10 dataset**

```
the final alpha, beta values are  [[10.28602139]]   [[1.86030511]]
the corrresponding lambda value is [[5.52921203]]
similarly, the gamma value  [[93.19945644]]
number of iterations 2
0.6083148639537531
```

Run time: 0.04174119700000001

4. **Crime dataset**

```
the final alpha, beta values are  [[425.64536384]]   [[3.25043207]]
the corrresponding lambda value is [[130.95039525]]
similarly, the gamma value  [[29.17432011]]
number of iterations 19
0.3911023074715292
```

Run time: 0.10661346299999996

5. **Wine dataset**

```
the final alpha, beta values are  [[6.16397917]]   [[1.60980873]]
the corrresponding lambda value is [[3.82901338]]
similarly, the gamma value  [[7.36011856]]
number of iterations 30
0.626746238542359
```

Run time: 0.024137770000000003

From the output of Bayesian model selection it can be said that the results are similar to findings of lambda from task 1. Here, we have only used training set to find our optimal lambda whereas in task 1 we have also used our test data. In this case we don't use test set MSE and still get good results (finding optimal lambda) compared to task 1.

Task 3.3 : Comparison

Q)How do the two model selection methods compare in terms of effective λ, test set MSE and run time? Do the results suggest conditions where one method is preferable to the other?

A) Based on run time: Bayesian model takes much less time(from the results above) then cross validation as it does not contain as many loops.

Based on MSE: MSE values are rather similar in both the cases of model selection. Except for the 100-100 data case, where there is a little bit of disparity.

The conclusions that we can draw from these results are bayesian model selection can be used if low latency system is of utmost importance.