Coursework
Machine Learning (B555 - 17011)
Department of Computer Science, IU Bloomington

Fall 2021
Prof. Roni Khardon

Topic : Programming project 2 – Linear Regression Model Selection & Implementation Report
Submitted by : Pratap Roy Choudhury, MS in Data Science, Fall2021

1. Overview: Experiments with Bayesian Linear Regression

The goals in this assignment are to explore the role of regularization in linear regression and to investigate two methods for model selection (evidence maximization and cross validation). In all the experiments we should report the performance in terms of the mean square error

$$\mathrm{MSE} = \frac{1}{N}\sum_i (\phi(x_i)^T w - t_i)^2$$

where the number of examples in the corresponding dataset is N.

2. Data

Data for this assignment is provided in a zip file pp2data.zip. We have 4 datasets and each dataset comes in 4 les with the training set in train-name.csv the corresponding labels (regression values) in trainR-name.csv and similarly for test set. We have two real datasets **crime** and **wine** and two artificial datasets **artsmall** and **artlarge**.

Note that the train/test splits are fixed and we will not change them in the assignment (in order to save work and run time).

For the artificial data we can compare the MSE results to the MSE of the hidden true functions generating the data that give 0.533 (artsmall), and 0.557 (artlarge).

3. Implementation

3.1. Task 1: Regularization

In this part we use regularized linear regression, i.e., given a dataset, the solution vector ***w*** is given by equation (3.28) of Bishop's text.

$$\mathbf{w} = \left(\lambda \mathbf{I} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^{\mathrm{T}}\mathbf{t}.$$

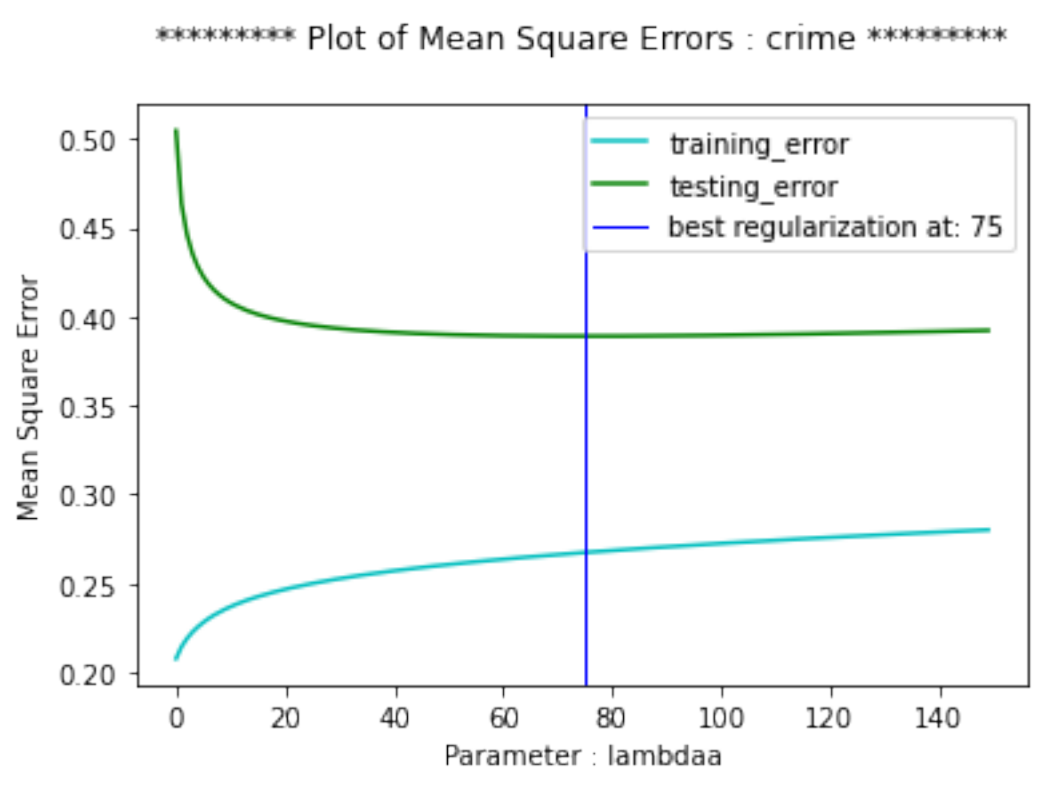
For each of the 4 datasets, we need to plot the training set MSE and the test set MSE as a function of the regularization parameter (use integer values in the range 0 to 150). For each dataset it is useful to put both curves on the same plot. In addition, compare these to the MSE of the true functions given above.

Qs. Provide the results/plots and discuss them: Why can't the training set MSE be used to select 'λ'? How does 'λ' affect error on the test set? Does this differ for different datasets? How do you explain these variations?

The results and plots of regularization process applied on different datasets are listed below –

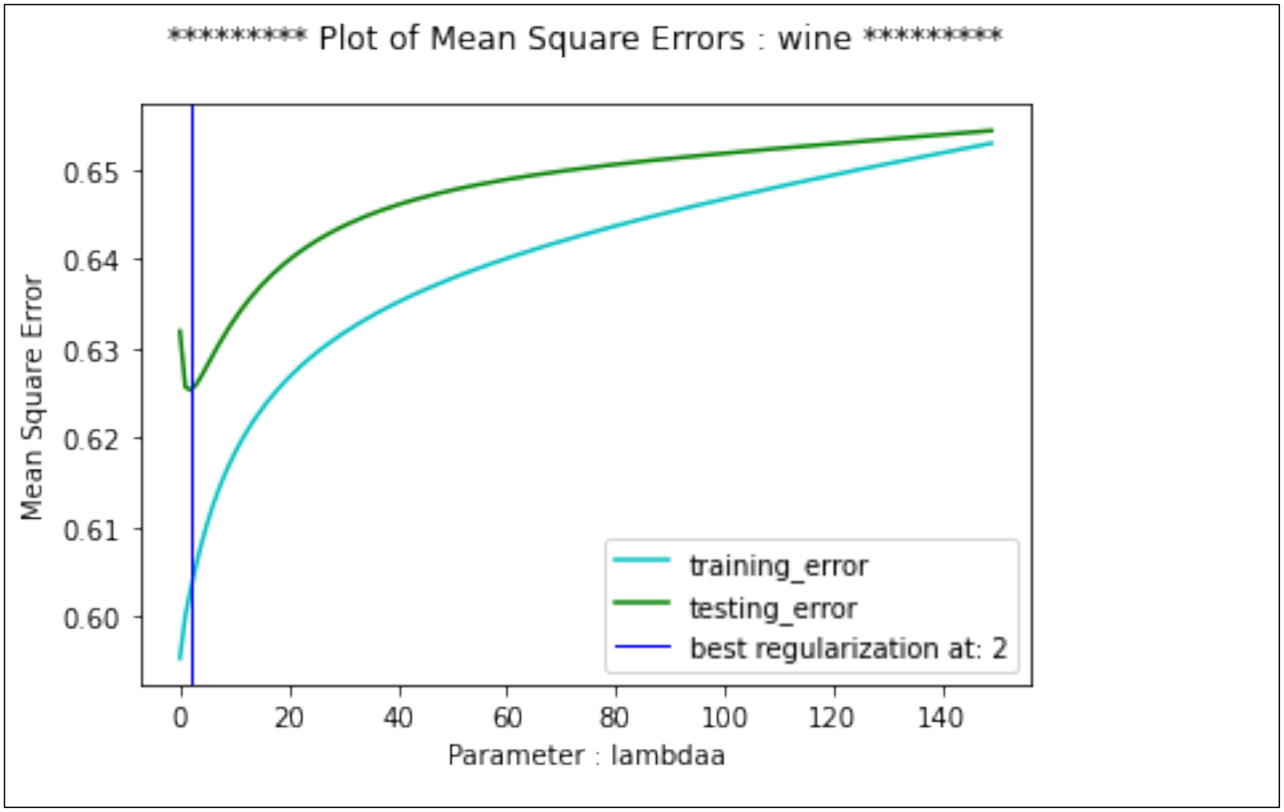
Data Set: ***crime***

```
        ============ Processing the data file : crime ============

TRAINING SET :  (298, 100) (298, 1)
TESTING SET :  (1695, 100) (1695, 1)

Training data size : 298
Training feature size : 100

>>>>> Enter the Upper limit of Lambdaa : 150

        ################ Execution Segment of Task-1 : Regularization ################

        !!!!!!!!!!!!!!!! RESULTS Task 1 - Regularization !!!!!!!!!!!!!!!!

Minimum MSE of complete test data :  0.389
Lambdaa for Minimum MSE of complete testing data :  75

 Execution time for Task 1 - Regularization : 1.032 seconds
```



Data Set: ***wine***

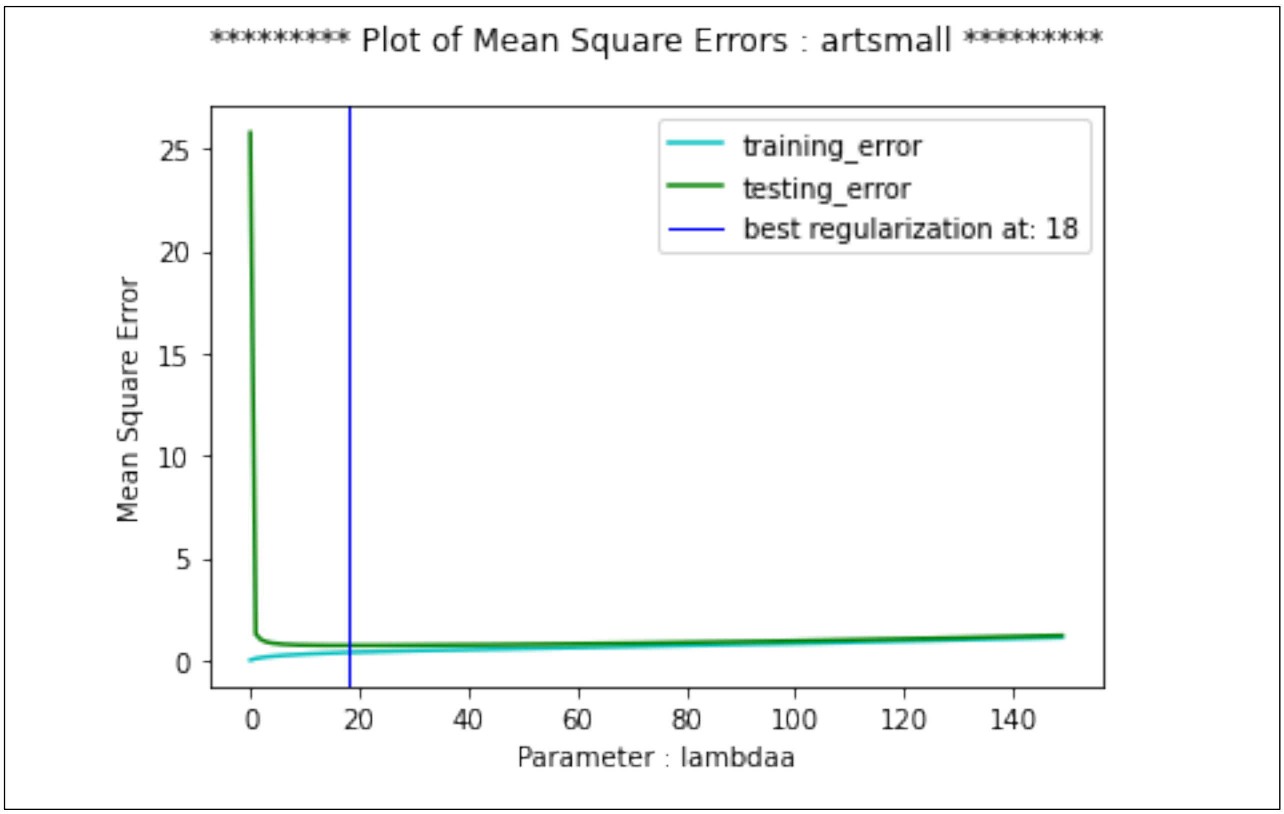
```
        ============ Processing the data file : wine ============

TRAINING SET :  (342, 11) (342, 1)
TESTING SET :  (4556, 11) (4556, 1)

Training data size : 342
Training feature size : 11

>>>>> Enter the Upper limit of Lambdaa : 150

        ################ Execution Segment of Task-1 : Regularization ################

        !!!!!!!!!!!!!!!! RESULTS Task 1 - Regularization !!!!!!!!!!!!!!!!

Minimum MSE of complete test data :  0.625
Lambdaa for Minimum MSE of complete testing data :  2

 Execution time for Task 1 - Regularization : 2.159 seconds
```



The test set MSE has a steep fall at $\lambda = 2$ and then again increased gradually as λ increased.

Data Set: ***artsmall***

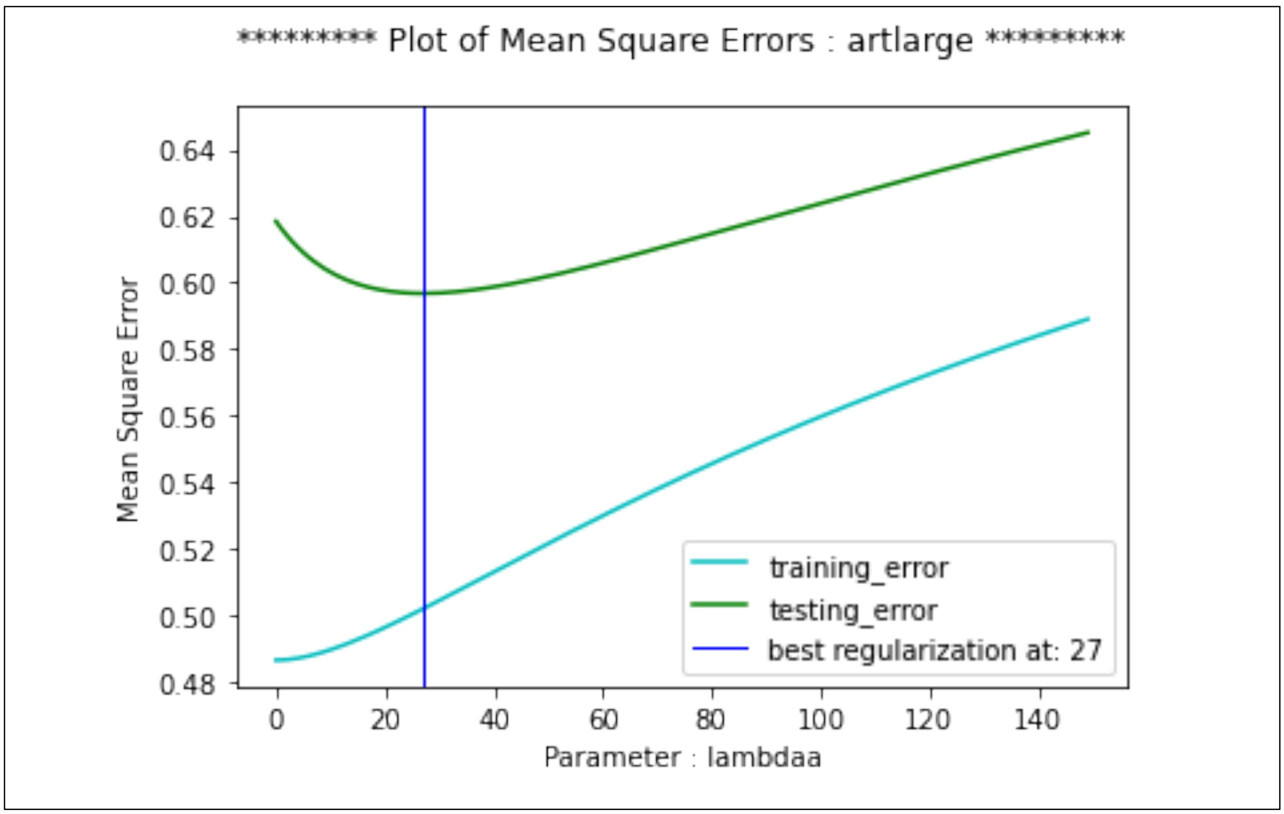
```
        ============ Processing the data file : artsmall ============

TRAINING SET :  (100, 100) (100, 1)
TESTING SET :  (1000, 100) (1000, 1)

Training data size : 100
Training feature size : 100

>>>>> Enter the Upper limit of Lambdaa : 150

        ################ Execution Segment of Task-1 : Regularization ################

        !!!!!!!!!!!!!!!! RESULTS Task 1 - Regularization !!!!!!!!!!!!!!!!

Minimum MSE of complete test data :  0.72
Lambdaa for Minimum MSE of complete testing data :  18

 Execution time for Task 1 - Regularization : 0.592 seconds
```



The test set MSE has a steep initial fall and get its minimum at λ = 18 and thereafter maintains almost similar MSE as λ increased.

Data Set: ***artlarge***

```
        ============ Processing the data file : artlarge ============

TRAINING SET :  (1000, 100) (1000, 1)
TESTING SET :  (1000, 100) (1000, 1)

Training data size : 1000
Training feature size : 100

>>>>> Enter the Upper limit of Lambdaa : 150

        ################ Execution Segment of Task-1 : Regularization ################

        !!!!!!!!!!!!!!!! RESULTS Task 1 - Regularization !!!!!!!!!!!!!!!!

Minimum MSE of complete test data :  0.597
Lambdaa for Minimum MSE of complete testing data :  27

 Execution time for Task 1 - Regularization : 1.135 seconds
```



The test set MSE decreases gradually and gets its minimum at $\lambda = 27$ and then again increased gradually as λ increased.

From the above results and plots of training and testing errors, it's clearly visible that the training error is increasing gradually as the regularization parameter increases and minimum always at $\lambda = 0$. Hence the training set MSE cannot be used to select λ**.**

The MSE of test set decreases initially from $\lambda = 0$ till a certain limit and reaches its global minima. From that point either the MSE increases very slightly as λ increases or maintain similar MSE values with minute variation. The variation in change in MSE values differ for different dataset but all of them follow the similar pattern of initial decrease and then increase gradually from the global minima.

Note: The experiments in this task tell us which value of 'λ' is best in every case in *hindsight*. That is, we need to see the test data and its labels to choose 'λ'. This is clearly not a realistic setting, and it does not give reliable error estimates. The next two tasks investigate methods for choosing 'λ' automatically without using the test set.

3.2. Task 2 : Model Selection using Cross Validation

In this part we use 10 fold cross validation on the training set to pick the value of 'λ' in the same range as above, then retrain on the entire train set and evaluate on the test set.

To select parameter **a** of algorithm A(a) over an enumerated range **a** $\in V_1. . . . V_K$ using dataset D

we do the following:

I. Split the data D into 10 disjoint portions.
II. For each value of a in $V_1; : : : ; V_K$:
    a. For each i in 1 : : : 10
        i. Train A(a) on all portions but i and test on i recording the error on portion i
    b. Record the average performance of a on the 10 folds.
III. Pick the value of a with the best average performance.

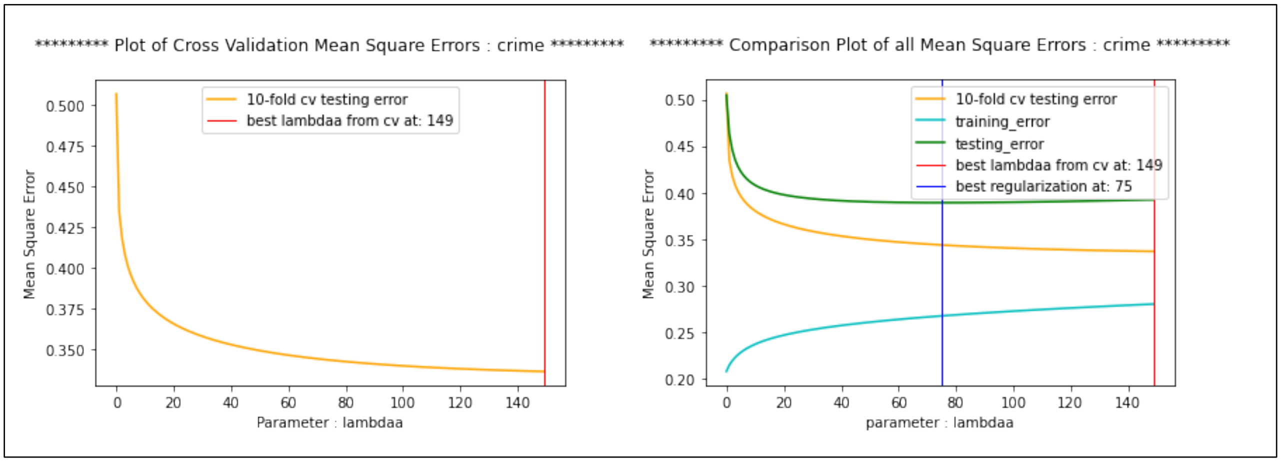
Now, in the above, D only includes the training set and the parameter is chosen without knowledge of the test data. We then retrain on the entire train set D using the chosen value and evaluate the result on the test set.

Qs. Implement this scheme, apply it to the 4 datasets and report the values of 'λ' selected, associated MSE and the run time. How do the results compare to the best test-set results from part 1 both in terms of the choice of 'λ' and test set MSE?

The 10-fold cross validation is applied for 150 'λ' for all the datasets and then the test set MSE is fetched from the Task-1 test set MSE for best chosen 'λ' where the MSE of all the validation sets is minimum.

Data Set: ***crime***

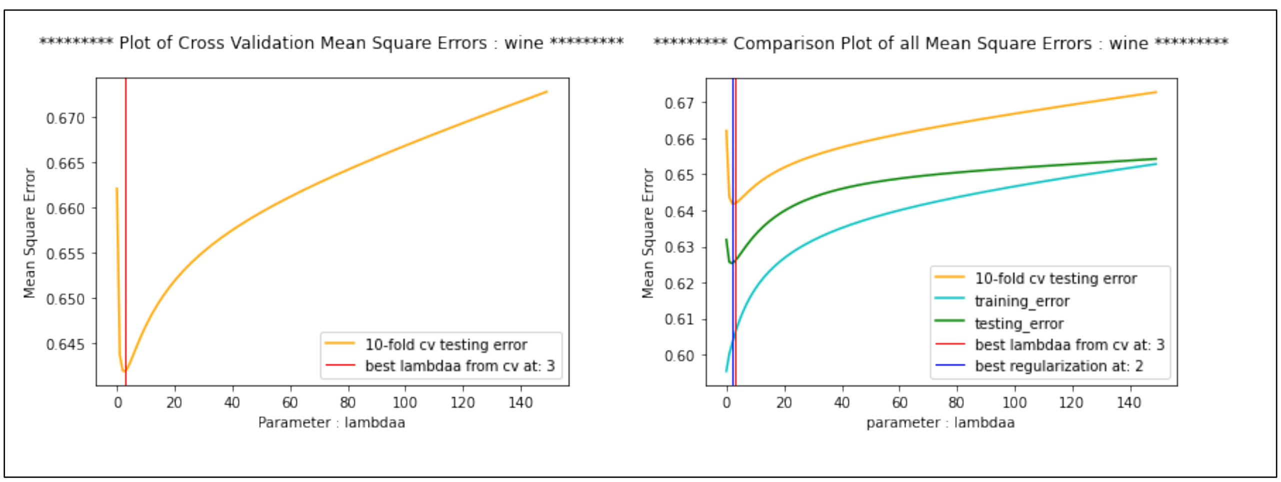
```
        ################ Execution Segment of Task-2 : Model Selection using Cross Validation ################

  Enter the number of Cross validation folds : 10

        !!!!!!!!!!!!!!!! RESULTS Task 2 : Model Selection using Cross Validation !!!!!!!!!!!!!!!!

  Minimum MSE of 10 fold cross validation data :  0.337
  Lambdaa for Minimum MSE of 10 fold cross validation data :  149
  MSE of test data for best chosen lambdaa = 149 :  0.392

   Execution time for Task 2 - Model Selection using Cross Validation : 2.658 seconds
```



The λ chosen from cross validation in **crime** dataset is at the extreme limit which is 149 here compared to the best test-set result from Task-1 where λ was 75 and MSE was 0.389, and from Task-2, at λ = 149, MSE is 0.392.

Data Set: ***wine***

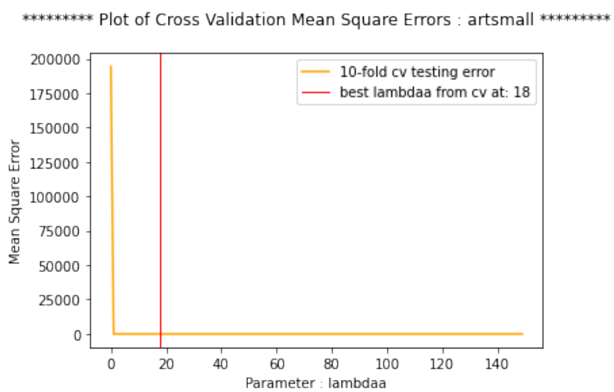
```
        ################ Execution Segment of Task-2 : Model Selection using Cross Validation ################

Enter the number of Cross validation folds : 10

        !!!!!!!!!!!!!!! RESULTS Task 2 : Model Selection using Cross Validation !!!!!!!!!!!!!!!

Minimum MSE of 10 fold cross validation data :  0.642
Lambdaa for Minimum MSE of 10 fold cross validation data :  3
MSE of test data for best chosen lambdaa = 3 :  0.626

 Execution time for Task 2 - Model Selection using Cross Validation : 1.67 seconds
```

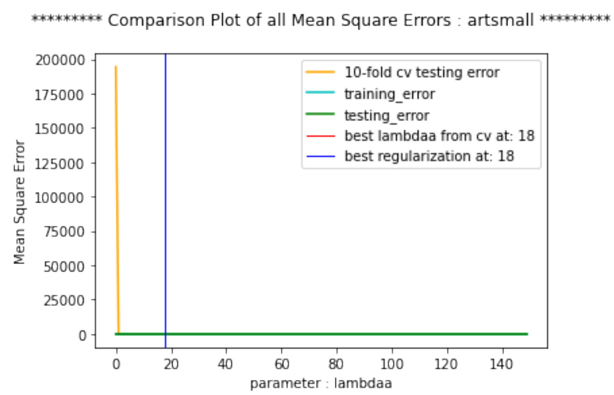


The λ chosen from cross validation in **wine** dataset is almost same compared to the best test-set result from Task-1 where λ was 2 and MSE was 0.625, and from Task-2, at λ = 3, MSE is 0.626. Thus the model selected performs almost similar.

Data Set: ***artsmall***

```
################ Execution Segment of Task-2 : Model Selection using Cross Validation ################

Enter the number of Cross validation folds : 10

        !!!!!!!!!!!!!!!! RESULTS Task 2 : Model Selection using Cross Validation !!!!!!!!!!!!!!!!

Minimum MSE of 10 fold cross validation data :  0.705
Lambdaa for Minimum MSE of 10 fold cross validation data :  18
MSE of test data for best chosen lambdaa = 18 :  0.72

 Execution time for Task 2 - Model Selection using Cross Validation : 2.06 seconds
```

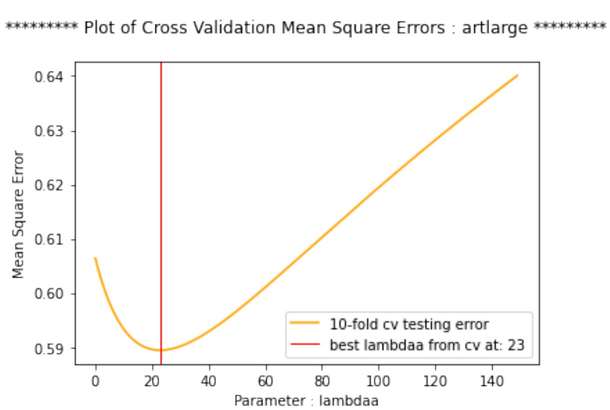


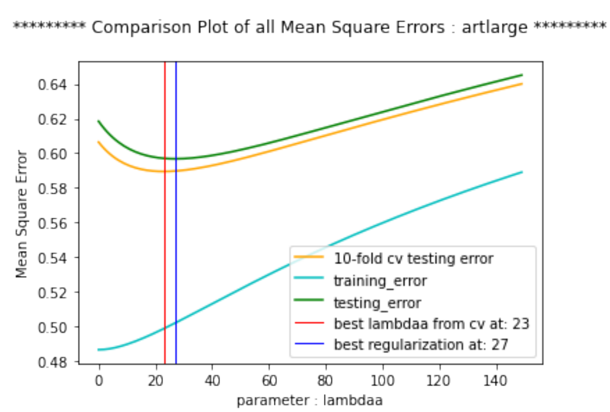


Here both the model behaves exactly the same and the best λ = 18 where MSE is minimum 0.72.

Data Set: ***artlarge***

```
################ Execution Segment of Task-2 : Model Selection using Cross Validation ################

Enter the number of Cross validation folds : 10

        !!!!!!!!!!!!!!!! RESULTS Task 2 : Model Selection using Cross Validation !!!!!!!!!!!!!!!!

Minimum MSE of 10 fold cross validation data :  0.589
Lambdaa for Minimum MSE of 10 fold cross validation data :  23
MSE of test data for best chosen lambdaa = 23 :  0.597

 Execution time for Task 2 - Model Selection using Cross Validation : 4.484 seconds
```





Here the selected model parameter λ = 23 gives the minimum MSE 0.597 which almost same from the Task -1 best test-set result which came for λ = 27.

3.3. Task 3 : Bayesian Model Selection

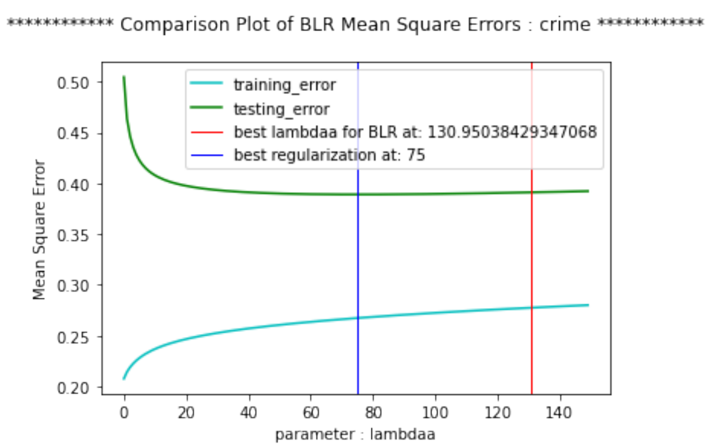
In this part we consider the formulation of Bayesian linear regression with the simple prior w~N(0, $\frac{1}{\alpha}I$). Recall that the evidence function (and evidence approximation) gives a method to pick the parameters α and β. Referring to Bishop's book, the solution is given in equations (3.91), (3.92), (3.95), where m_N and S_N are given in (3.53) and (3.54). As discussed in class these yield an iterative algorithm for selecting α and β using the training set. We can then calculate the MSE on the test set using the MAP (m_N) for prediction.

This scheme is pretty stable and converges in a reasonable number of iterations. You can initialize α and β to random values in the range [1, 10] and stop the algorithm when the difference in α and β values is < 0.0001.

Qs. Implement this scheme, apply it to the 4 datasets and report the values of α and β, the effective $\lambda = \frac{\alpha}{\beta}$, the associated MSE and the run time. How do the results compare to the best test-set results from part 1 both in terms of the choice of λ and test set MSE?

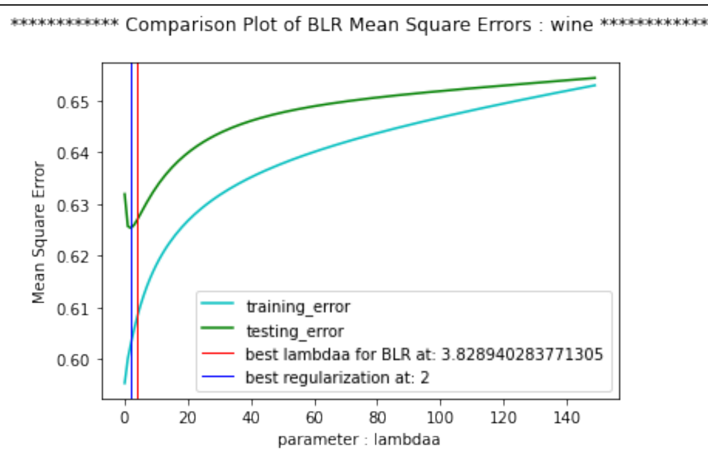
Data Set: ***crime***

```
        ################ Execution Segment of Task-3 : Bayesian Model Selection ################

 Initialized Hyper-parameter : Alpha = 3.8822618748065874  Beta = 9.22846340011492

        !!!!!!!!!!!!!!!! RESULTS Task 3 : Bayesian Model Selection !!!!!!!!!!!!!!!!

 Finalized Alpha parameter :  425.6453317083991
 Finalized Beta parameter :  3.2504320930780364

 Lambdaa for Bayesian Hyper-parameters :  130.95
 MSE of test data set for Bayesian lambdaa 130.95038429347068 is :  0.391

  Execution time for Task 3 : Bayesian Model Selectio : 0.078 seconds
```



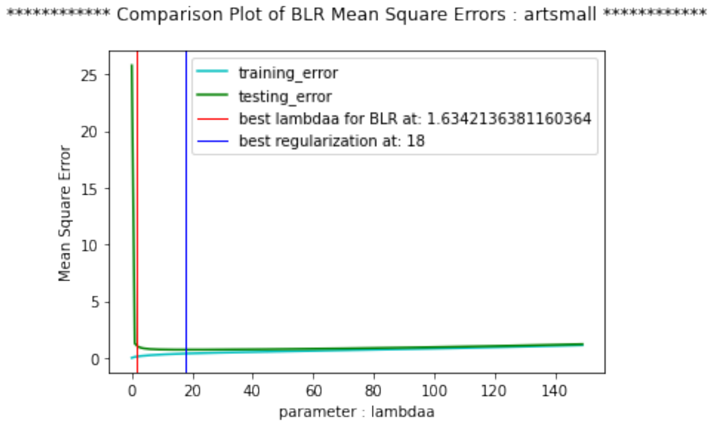
Data Set: ***wine***

```
        ################ Execution Segment of Task-3 : Bayesian Model Selection ################

Initialized Hyper-parameter : Alpha = 6.3714604338577665  Beta = 8.38606960401361

        !!!!!!!!!!!!!!!! RESULTS Task 3 : Bayesian Model Selection !!!!!!!!!!!!!!!!

Finalized Alpha parameter :  6.1638640340577675
Finalized Beta parameter :  1.609809393001733

Lambdaa for Bayesian Hyper-parameters :  3.829
MSE of test data set for Bayesian lambdaa 3.828940283771305 is :  0.627

 Execution time for Task 3 : Bayesian Model Selectio : 0.036 seconds
```



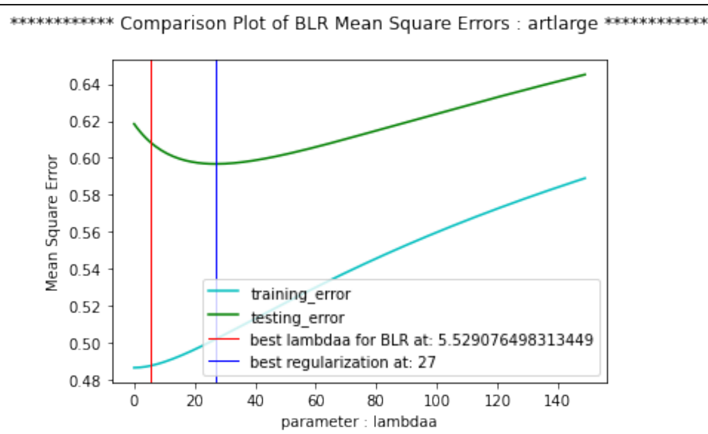
Data Set: ***artsmall***

```
        ################ Execution Segment of Task-3 : Bayesian Model Selection ################

Initialized Hyper-parameter : Alpha = 8.691986416323857  Beta = 1.9383238162897802

        !!!!!!!!!!!!!!!! RESULTS Task 3 : Bayesian Model Selection !!!!!!!!!!!!!!!!

Finalized Alpha parameter :  5.154695560110029
Finalized Beta parameter :  3.1542360434909202

Lambdaa for Bayesian Hyper-parameters :  1.634
MSE of test data set for Bayesian lambdaa 1.6342136381160364 is :  1.063

 Execution time for Task 3 : Bayesian Model Selectio : 0.039 seconds
```



Data Set: ***artlarge***

```
################ Execution Segment of Task-3 : Bayesian Model Selection ################

Initialized Hyper-parameter : Alpha = 7.096032597026071  Beta = 7.308875335670949

!!!!!!!!!!!!!!! RESULTS Task 3 : Bayesian Model Selection !!!!!!!!!!!!!!!

Finalized Alpha parameter :  10.28579276918194
Finalized Beta parameter :  1.8603093613046313

Lambdaa for Bayesian Hyper-parameters :  5.529
MSE of test data set for Bayesian lambdaa 5.529076498313449 is :  0.608

 Execution time for Task 3 : Bayesian Model Selectio : 0.062 seconds
```



3.4. Task 3 : Discussion of Results

Qs. Tabulate together the values obtained in tasks 1-3 and use this to discuss the following questions- 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?

The results from all the above 3 tasks are tabulated below.

Dataset	Task 1 - Regularization			Task 2 - 10 fold Cross Validation			Task 3 - Bayesian Model Selection		
	Model Parameter Lambda	Test Set MSE	Run time (seconds)	Model Parameter Lambda	Test Set MSE	Run time	Model Parameter Lambda	Test Set MSE	Run time
Crime	75	0.389	1.032	149	0.392	2.658	130.95	0.391	0.078
Wine	2	0.625	2.159	3	0.626	1.67	3.829	0.627	0.036
ArtSmall	18	0.72	0.592	18	0.72	2.06	1.634	1.063	0.039
ArtLarge	27	0.597	1.135	23	0.597	4.484	5.529	0.608	0.062

Model Selection method comparison :

- For crime data, Bayesian method selects λ less than that of 10-fold cross validation with similar MSE and its almost 30 times faster.
- For wine dataset, both the method selects similar λ with similar MSE but the Bayesian method runs much faster.

Preferred method for model selection :

- Bayesian Model Selection method is preferred over 10-fold cross validation for Crime and Wine dataset as it gives almost similar model parameter λ with similar MSE in much faster run time.
- 10-fold Cross Validation is preferred for artificial datasets as it selects better λ for which test set MSE is much lesser, ignoring the run time which is indeed few seconds higher.