CSCI-B 555 MACHINE LEARNING

PROGRAMMING PROJECT 2

Experiments with Bayesian Linear Regression

Neha Pai - nrpai@iu.edu

Introduction

The goals of this assignment are:

1. to investigate the effect of the number of examples, the number of features, and the regularization parameter on the performance of the corresponding algorithms
2. to investigate two methods for model selection in linear regression (evidence maximization and cross validation).

In the experiments we will try to evaluate and report the performance in terms of the mean square error given by,

$$MSE = \frac{1}{N} \sum_{i} (\Phi(x_i)^T w - t_i)^2$$

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

The experiments will be carried out on 5 datasets, 2 real data sets and 3 artificially generated datasets. The datasets are labelled train-name.csv, trainR-name.csv for training data and similarly for testing data where train-name contains the training set data matrix and trainR-name contains regression values.

Task 1: Regularization

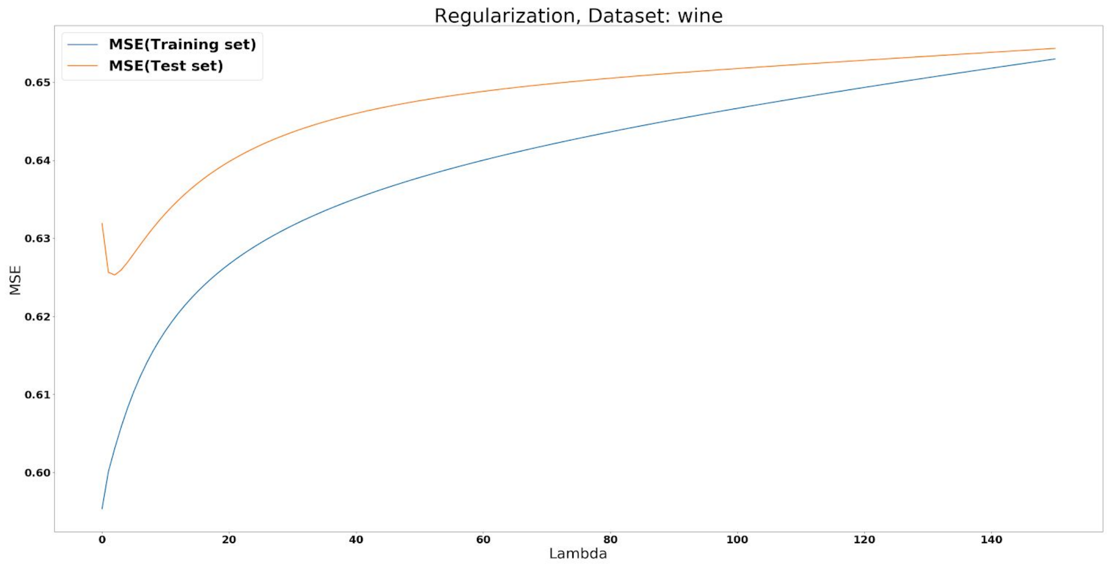
In this task we train the model using Linear Regression by adding a regularization parameter λ in order to control over-fitting. We get,

$$w = (\lambda I + \Phi^T \Phi)^{-1} \Phi^T t$$

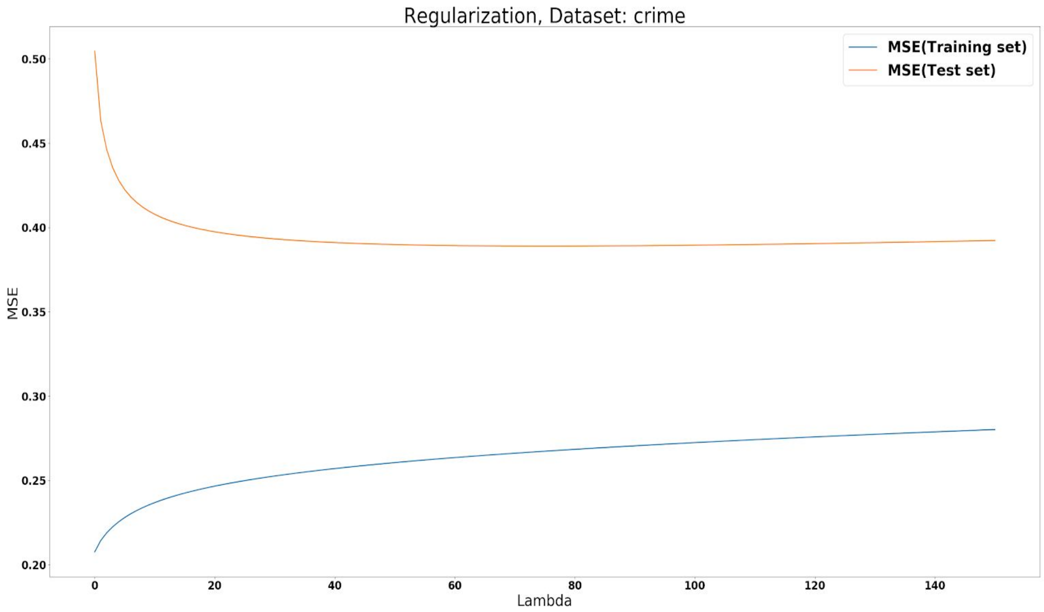
For this experiment we train the model for different values of λ ranging from 0 to 150 and calculate the MSEs on Test Data and analyse the results.

Results

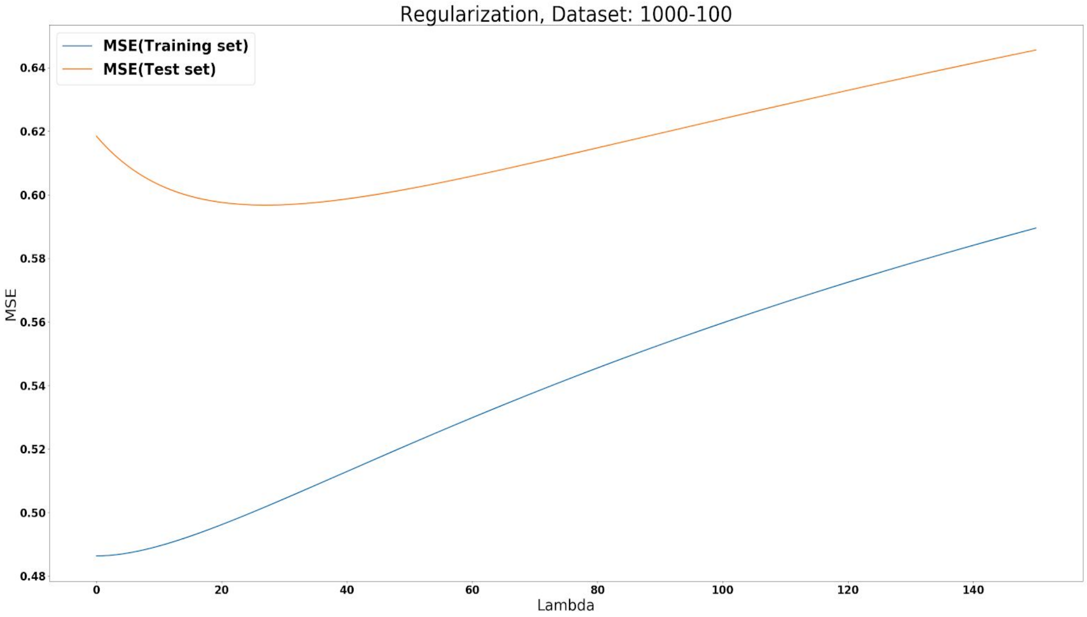
1) Plot of Training and Test Set MSEs for Dataset - Wine (342-11)



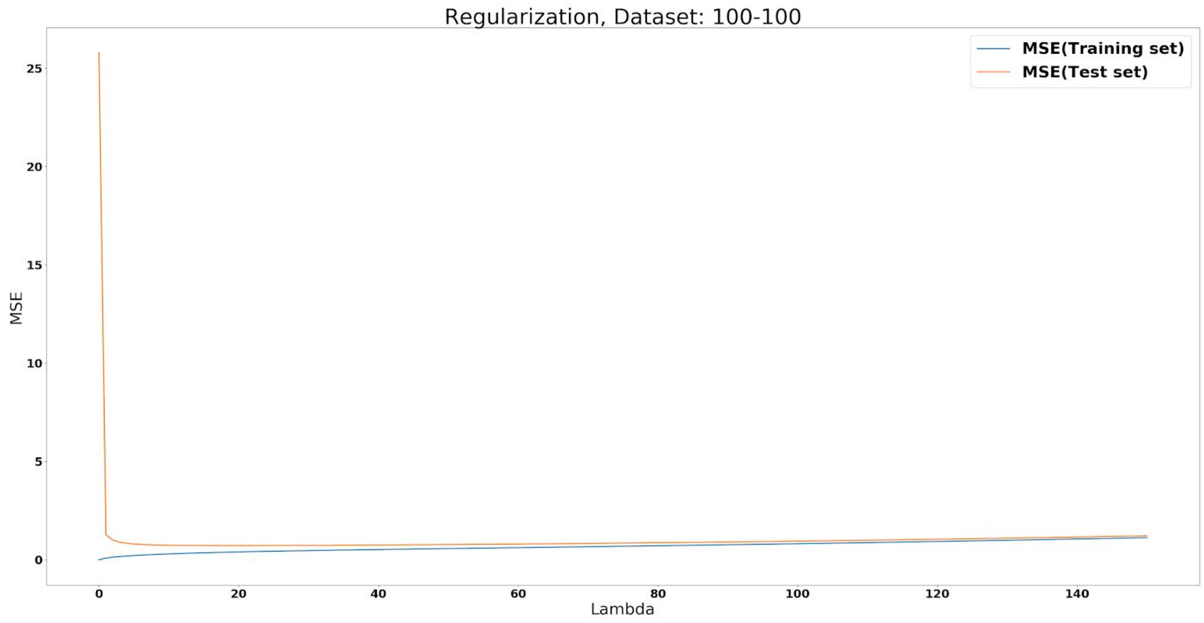
2) Plot of Training and Test Set MSEs for Dataset - Crime (298-100)



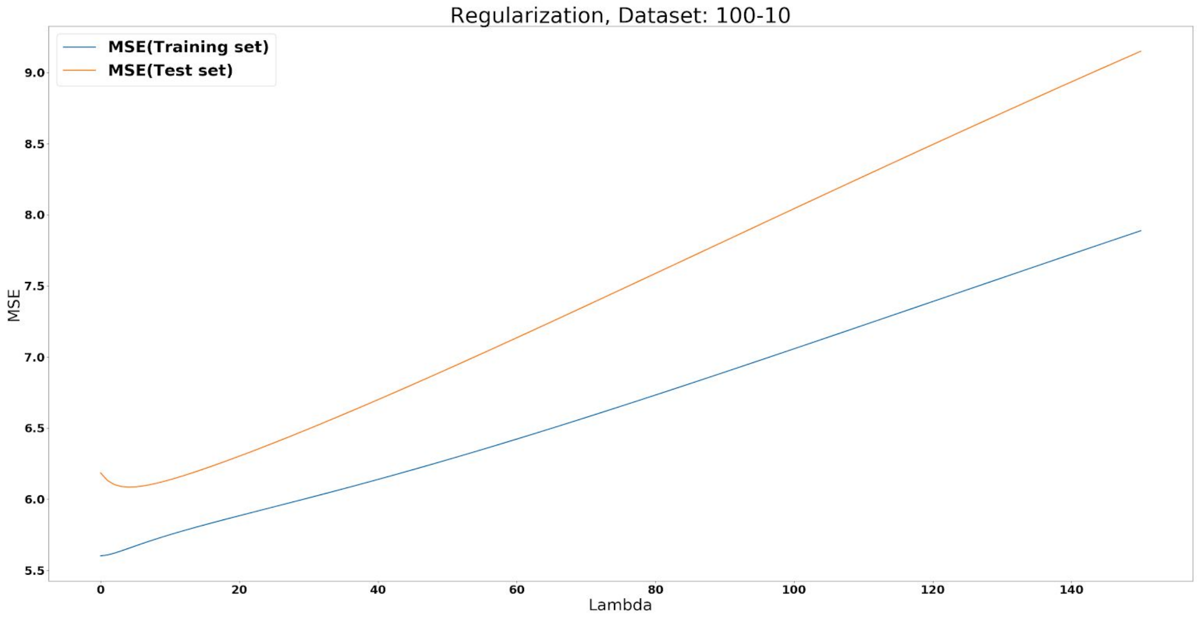
3) Plot of Training and Test Set MSEs for Dataset - 1000-100



4) Plot of Training and Test Set MSEs for Dataset - 100-100



5) Plot of Training and Test Set MSEs for Dataset - 100-10



Questions?

- Why can't the training set MSE be used to select λ?

By observing the plots we can observe that the Training set MSEs are comparatively lower than the Test set MSEs. This may be because we are training the model on train set data so it gives a lower error value, probably because the model fits the data well, but the same gives more error for testing data, which is a case of over-fitting. Hence it is better to choose λ based on the Test set MSE.

- How does λ affect error on the test set? Does this differ for different datasets? How do you explain these variations?

We can summarize the above results as follows:

Dataset (Samples-Features)	**λ**	**Minimum MSE**
Wine (342-11)	2	0.63
Crime (298-100)	75	0.39
1000-100	27	0.59
100-100	18	0.72
100-10	4	6.08

Table 1: Best Test MSE and λ for 5 Datasets

We can observe that if λ is at a ‘good’ value, neither too low nor too high, we can avoid overfitting and get low MSEs. Bigger datasets such as 1000-100 dataset, with more samples or data sets with more features as observed can tolerate larger values of λ to improve accuracy. It is also observed that MSE for Crime and 100-100 dataset remains constant compared to other dataset plots when MSE increases with increasing regularization, which is explained by the increased number of features which the model fits but on a small data sample.

- Comparison of result MSEs to MSEs of the true functions given

The values of MSEs obtained are comparable to the MSE of hidden true functions generating the artificial datasets.

	Result MSE	True Function MSE
1000-100	0.59	0.557
100-100	0.72	0.533
100-10	6.08	5.714

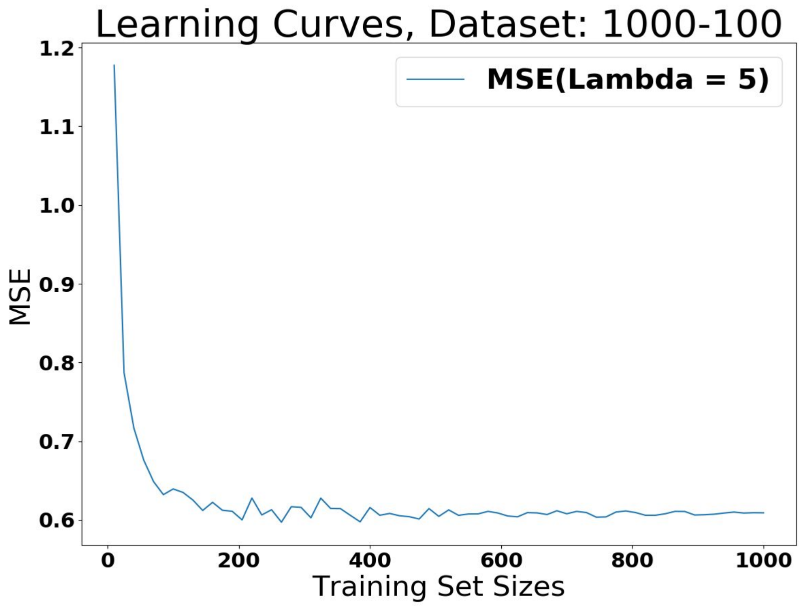
Table 2: Comparison of experiment MSE with MSE of true function

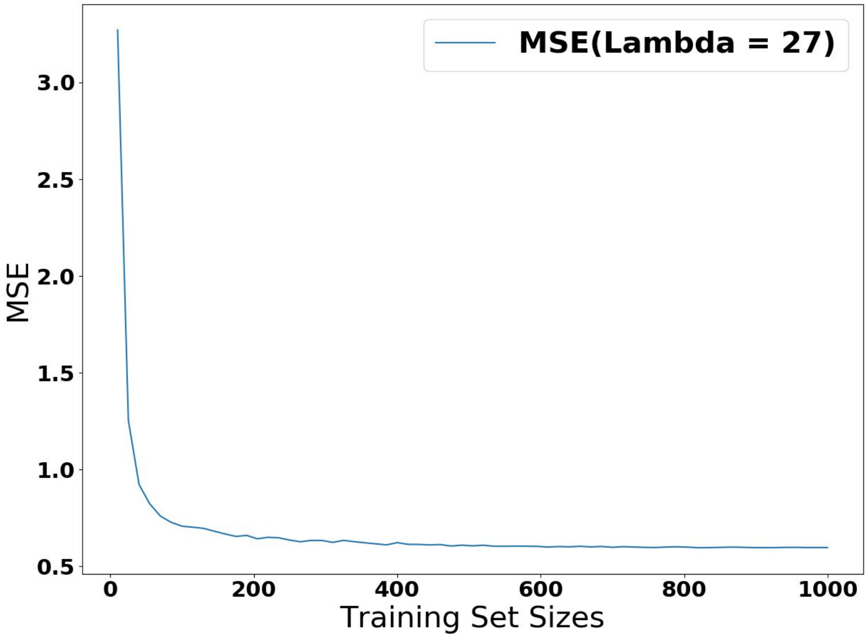
Task 2: Learning Curves

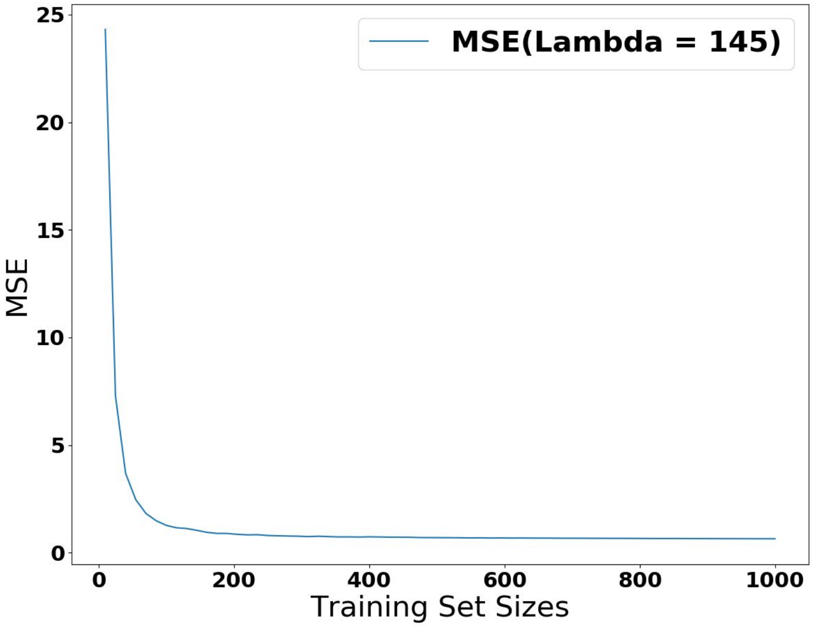
In this experiment we analyse the effect of training set on the performance of the model. We pick three “representative” values of λ from the first part (“too small”, “just right”, and “too large”) for the dataset 1000-100. For each of these values we plot a learning curve for the learned regularized linear regression on this dataset.

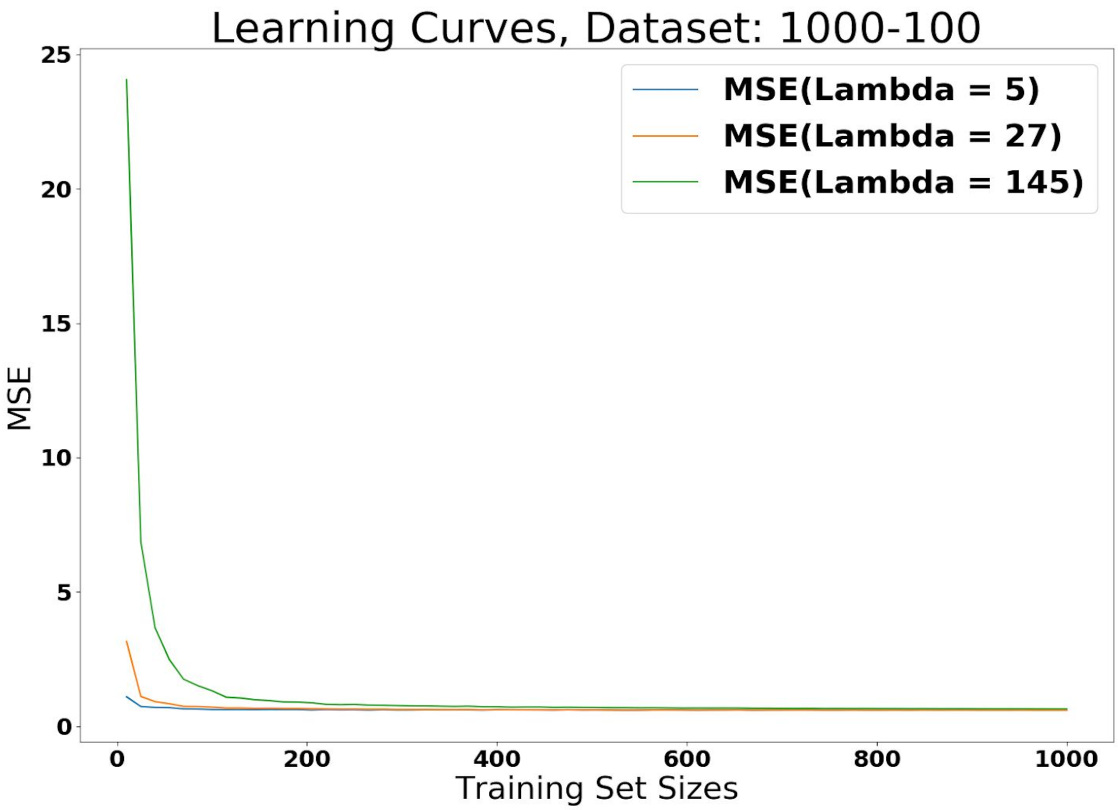
Results

From Task 1, we can observe that a global minima for test set MSE occurs at λ = 27 for dataset 1000-100. To plot our learning curves we pick three λ values - 5, 27 and 145 and consider fractions of training sizes to train the model. The results are as follows:




MSE(Lambda = 27)
MSE
3.0
2.5
2.0
1.5
1.0
0.5
0
200
400
600
800
1000
Training Set Sizes


MSE(Lambda = 145)
MSE
25
20
15
10
5
0
0
200
400
600
800
1000
Training Set Sizes



Questions?

- What can you observe from the plots regarding the dependence of the error on λ and on the number of samples? Consider both the case of small training set sizes and large training set sizes. How do you explain these variations?

  From the above plots we can observe a low performance of the model as expected for smaller samples of training data, which is the result of overfitting. Accuracy increases for larger datasets. For small data samples below 200, the dataset produces large errors for chosen large λ i.e. 150, while smaller values of λ, give better performance. For larger training set sizes, we do not observe distinct variations for different λ values on the model.

Task 3: Model Selection

3.1 Model Selection Using Cross Validation

Results of application of Cross Validation to 5 datasets and report of the values of λ selected, associated MSE and the run time.

	Wine	Crime	1000-100	100-100	100-10
Lambda	2	150	23	18	15
MSE	0.625	0.392	0.597	0.720	6.214
Run Time (in sec)	0.249	0.980	0.746	0.422	0.108

Table 3: Numerical results of Cross Validation for 5 Datasets

- 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?

  Values of λ and test set MSE obtained in task 1 are comparable to our results here. Only the results of Crime and 100-10 datasets do not match the results obtained in Task 1. This may be due to small dataset size. Cross validation is not consistent for small training samples.

3.2 Bayesian Model Selection

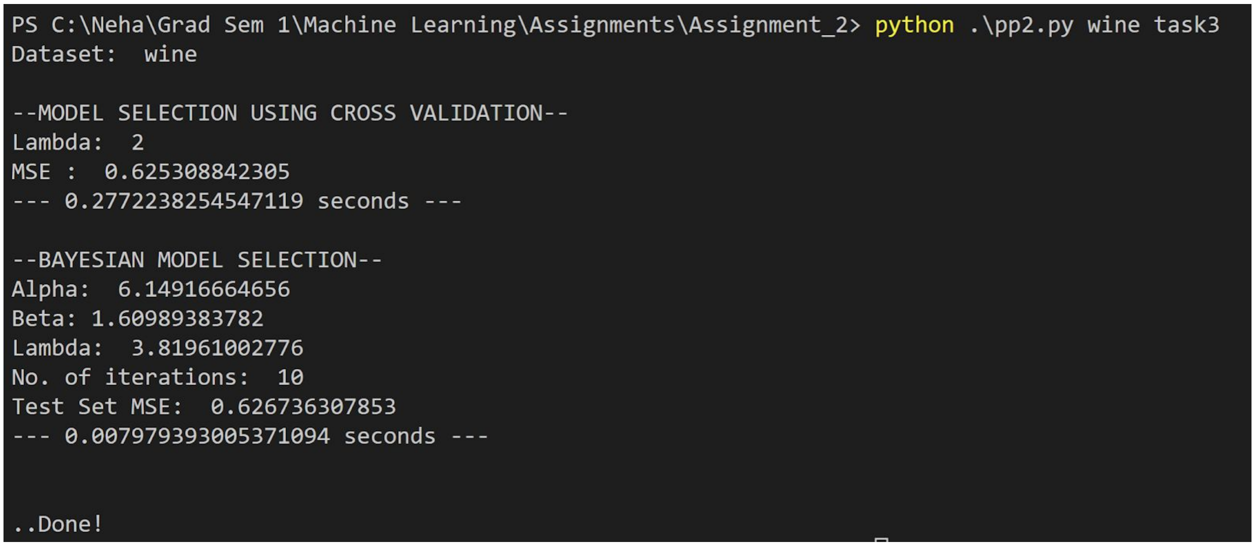
	Wine	Crime	1000-100	100-100	100-10
Alpha	6.149	425.672	10.286	5.179	0.882
Beta	1.609	3.250	1.860	3.127	0.165
Iterations	10	10	12	15	3
Lambda	3.819	130.959	5.529	1.656	5.345
MSE	0.626	0.391	0.608	1.058	6.087
Run Time (in sec)	0.009	0.093	0.082	0.088	0.0019

Table 4: Numerical Results of Bayesian Model Selection on 5 Datasets

- 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?

Values of test set MSE obtained in task 1 are almost comparable, while λ vary from that obtained in Task 1.

Snaps of code outputs for Task 3.1 and Task 3.2:

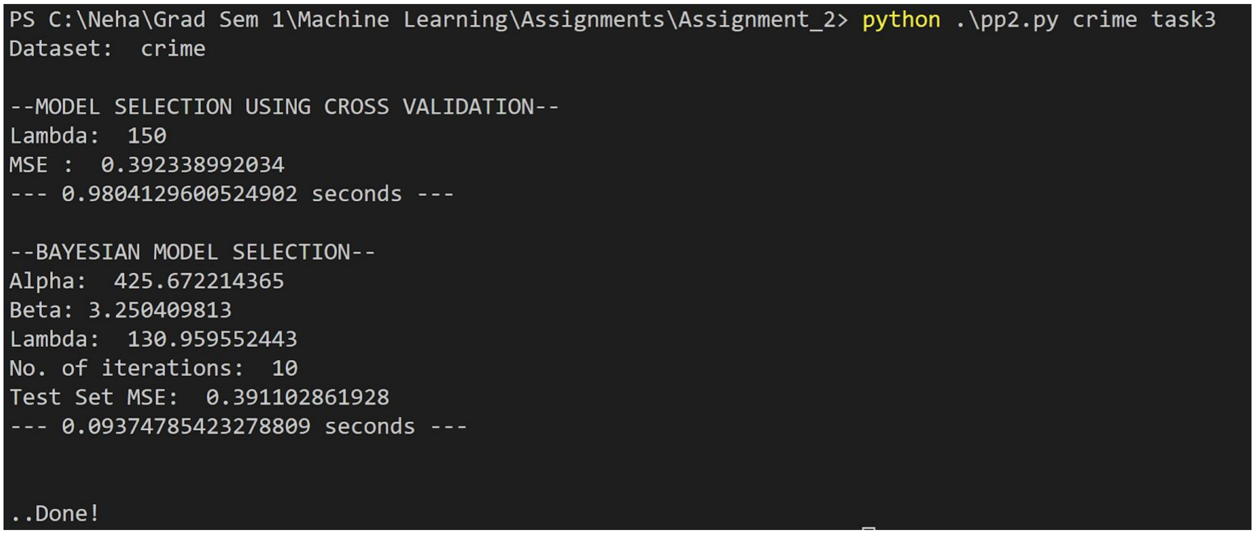


```
PS C:\Neha\Grad Sem 1\Machine Learning\Assignments\Assignment_2> python .\pp2.py wine task3
Dataset:  wine

--MODEL SELECTION USING CROSS VALIDATION--
Lambda:  2
MSE :  0.625308842305
--- 0.2772238254547119 seconds ---

--BAYESIAN MODEL SELECTION--
Alpha:  6.14916664656
Beta: 1.60989383782
Lambda:  3.81961002776
No. of iterations:  10
Test Set MSE:  0.626736307853
--- 0.007979393005371094 seconds ---


..Done!
```

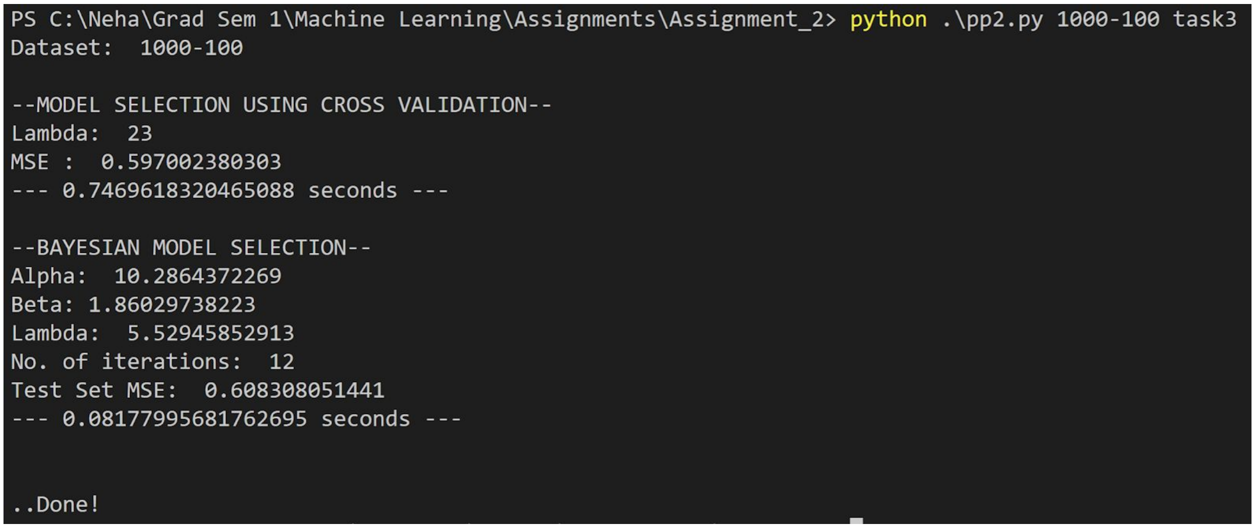


```
PS C:\Neha\Grad Sem 1\Machine Learning\Assignments\Assignment_2> python .\pp2.py crime task3
Dataset:  crime

--MODEL SELECTION USING CROSS VALIDATION--
Lambda:  150
MSE :  0.392338992034
--- 0.9804129600524902 seconds ---

--BAYESIAN MODEL SELECTION--
Alpha:  425.672214365
Beta: 3.250409813
Lambda:  130.959552443
No. of iterations:  10
Test Set MSE:  0.391102861928
--- 0.09374785423278809 seconds ---


..Done!
```

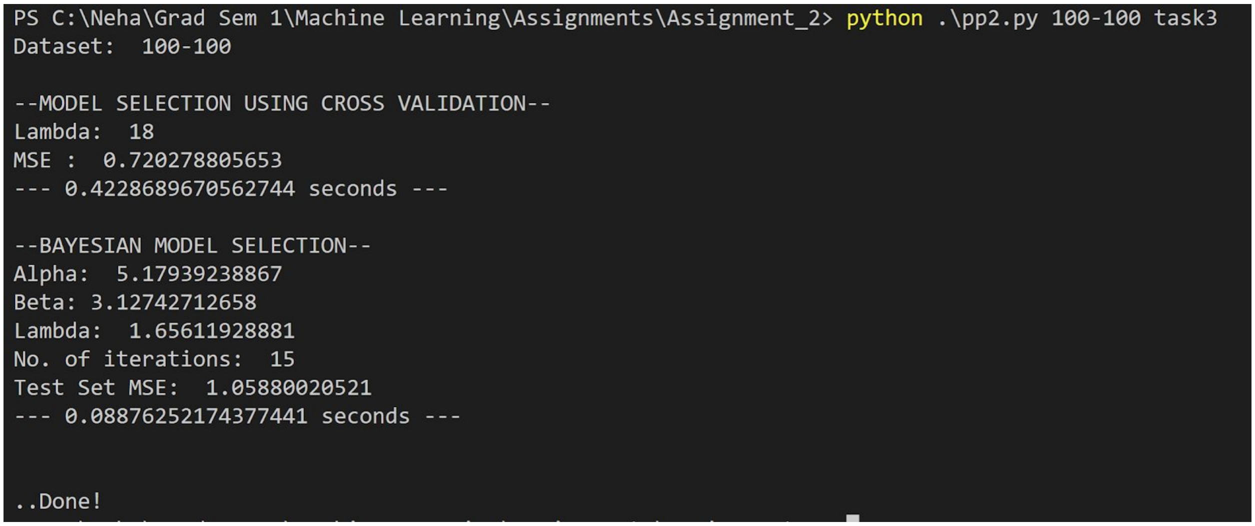


```
PS C:\Neha\Grad Sem 1\Machine Learning\Assignments\Assignment_2> python .\pp2.py 1000-100 task3
Dataset:  1000-100

--MODEL SELECTION USING CROSS VALIDATION--
Lambda:  23
MSE :  0.597002380303
--- 0.7469618320465088 seconds ---

--BAYESIAN MODEL SELECTION--
Alpha:  10.2864372269
Beta: 1.86029738223
Lambda:  5.52945852913
No. of iterations:  12
Test Set MSE:  0.608308051441
--- 0.08177995681762695 seconds ---


..Done!
```

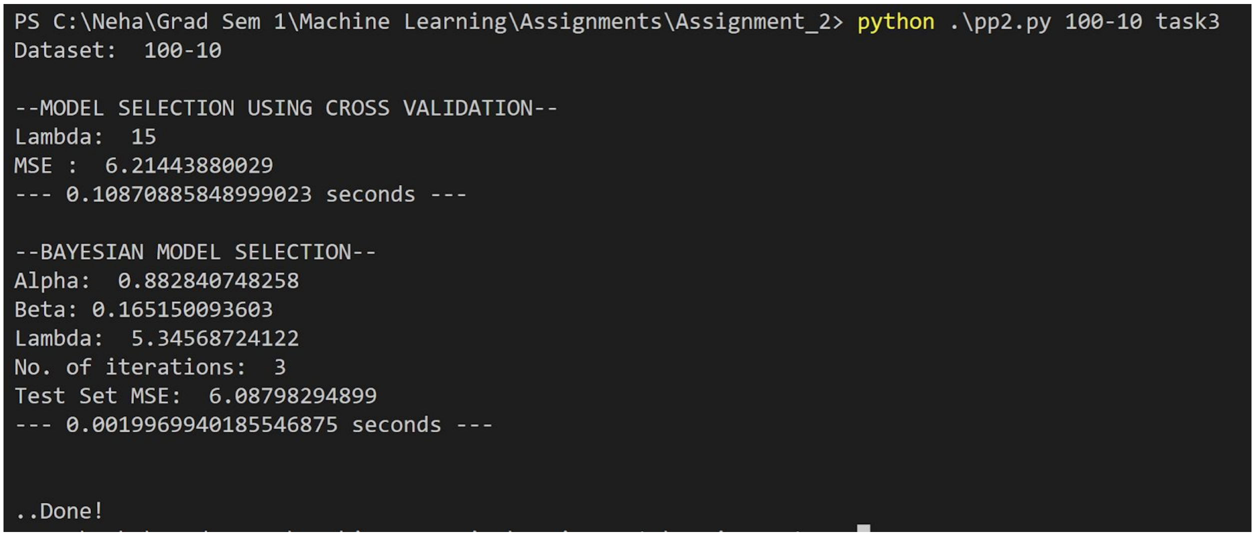


```
PS C:\Neha\Grad Sem 1\Machine Learning\Assignments\Assignment_2> python .\pp2.py 100-100 task3
Dataset:  100-100

--MODEL SELECTION USING CROSS VALIDATION--
Lambda:  18
MSE :  0.720278805653
--- 0.4228689670562744 seconds ---

--BAYESIAN MODEL SELECTION--
Alpha:  5.17939238867
Beta: 3.12742712658
Lambda:  1.65611928881
No. of iterations:  15
Test Set MSE:  1.05880020521
--- 0.08876252174377441 seconds ---


..Done!
```



```
PS C:\Neha\Grad Sem 1\Machine Learning\Assignments\Assignment_2> python .\pp2.py 100-10 task3
Dataset:  100-10

--MODEL SELECTION USING CROSS VALIDATION--
Lambda:  15
MSE :  6.21443880029
--- 0.10870885848999023 seconds ---

--BAYESIAN MODEL SELECTION--
Alpha:  0.882840748258
Beta: 0.165150093603
Lambda:  5.34568724122
No. of iterations:  3
Test Set MSE:  6.08798294899
--- 0.0019969940185546875 seconds ---


..Done!
```

3.3 Comparison

- 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? Please try to think about the results obtained and discuss these questions even if you do not see an obvious trend.

The Cross Validation approach works well for larger datasets and gives a better idea of λ. However, it takes considerable time to run this algorithm compared to Evidence Maximization, which is faster to estimate. We can observe that for cross validation, for the values of λ selected we gained good performance on Test set for all datasets except 100-10, which can be ignored due to low sample size. We get this result even though cross validation is carried out using training set only. The drawback of EM is that we are assuming a prior to carry a bayesian approach. We may choose CV when the algorithm demands accuracy, and choose EM for faster approach.