CSCI-B555 MACHINE LEARNING

PROGRAMMING PROJECT 2

Experiments with Bayesian Linear Regression

Experiment 1:

In this experiment we are investigating the effect of the number of examples, the number of features, and the regularization parameter on the performance of the corresponding algorithms.

Task 1: Regularization

In this task regularized linear regression is used, i.e., given a data set we find the solution vector ***W*** use it to calculate Mean Squared Error (MSE), given by:

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

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n)\}^2$$

Thus, for each **λ** from 0 to 150 we calculate **W** and calculate training set MSE and test set MSE as function of **λ.** This same process is done for the all the 5 given data sets and the following results are obtained:

Results:

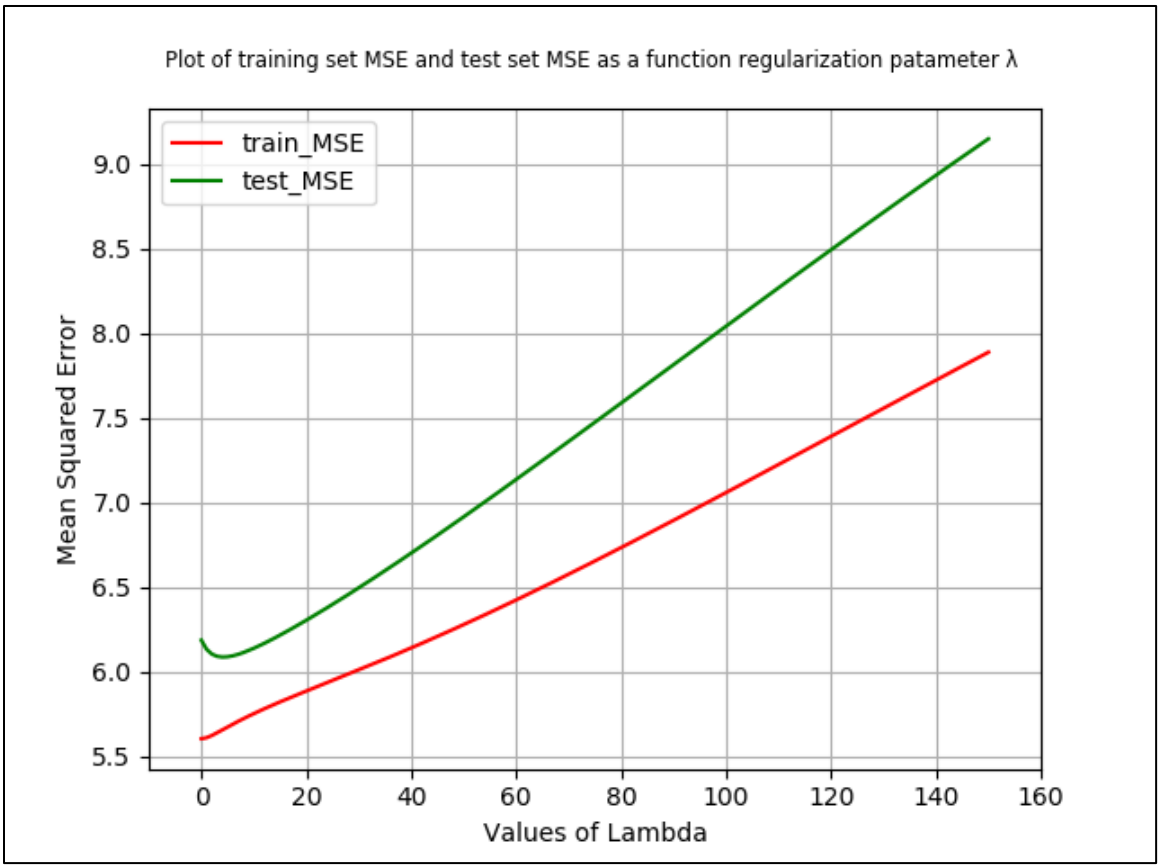


Figure 1: Plot for 100_10 data set

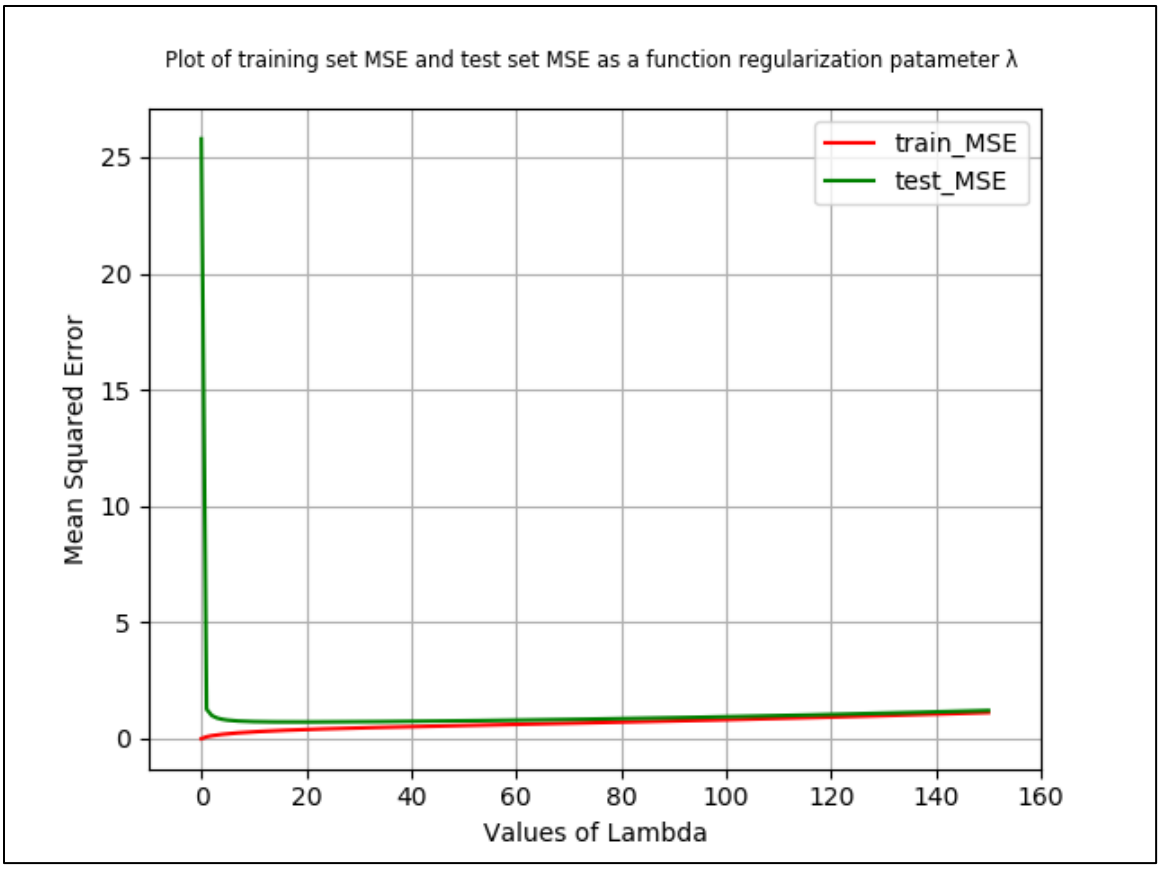


Figure 2: Plot for 100_100 data set

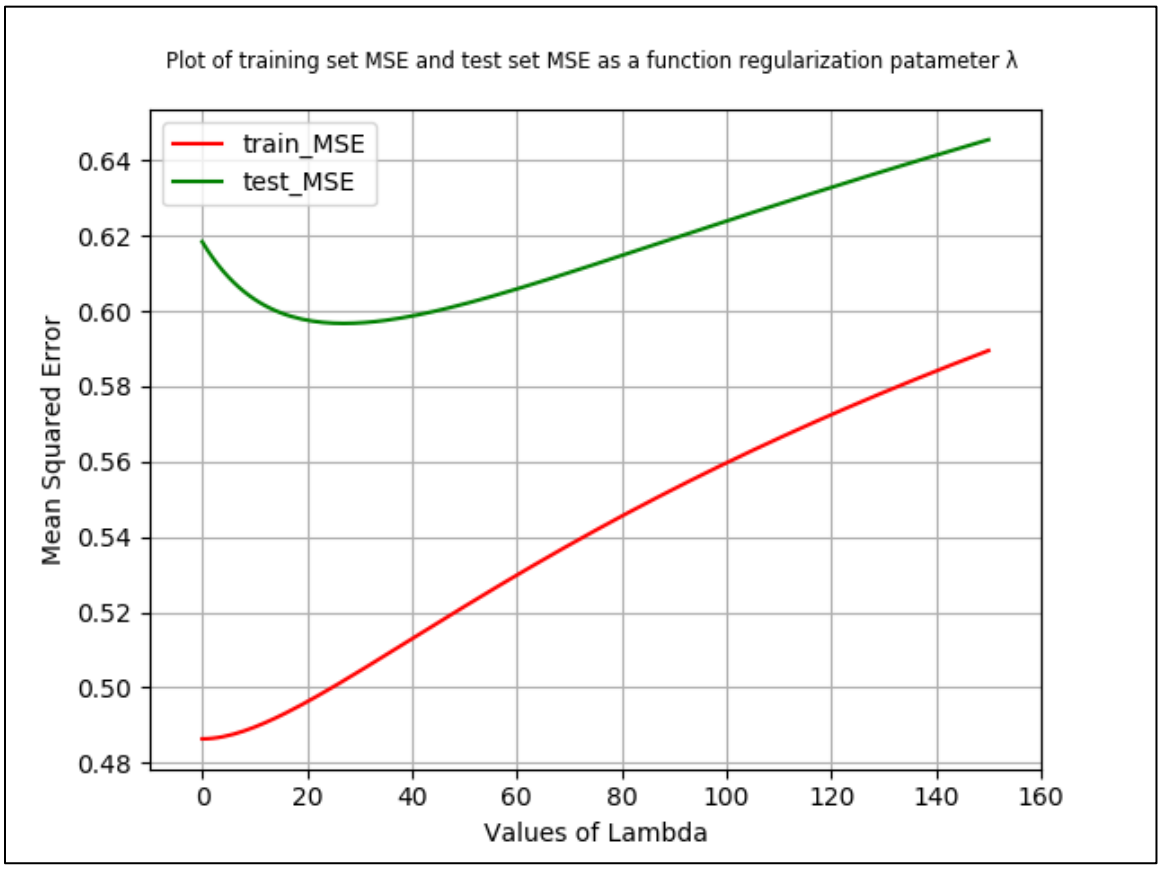


Figure 3: Plot for 1000_100 data set

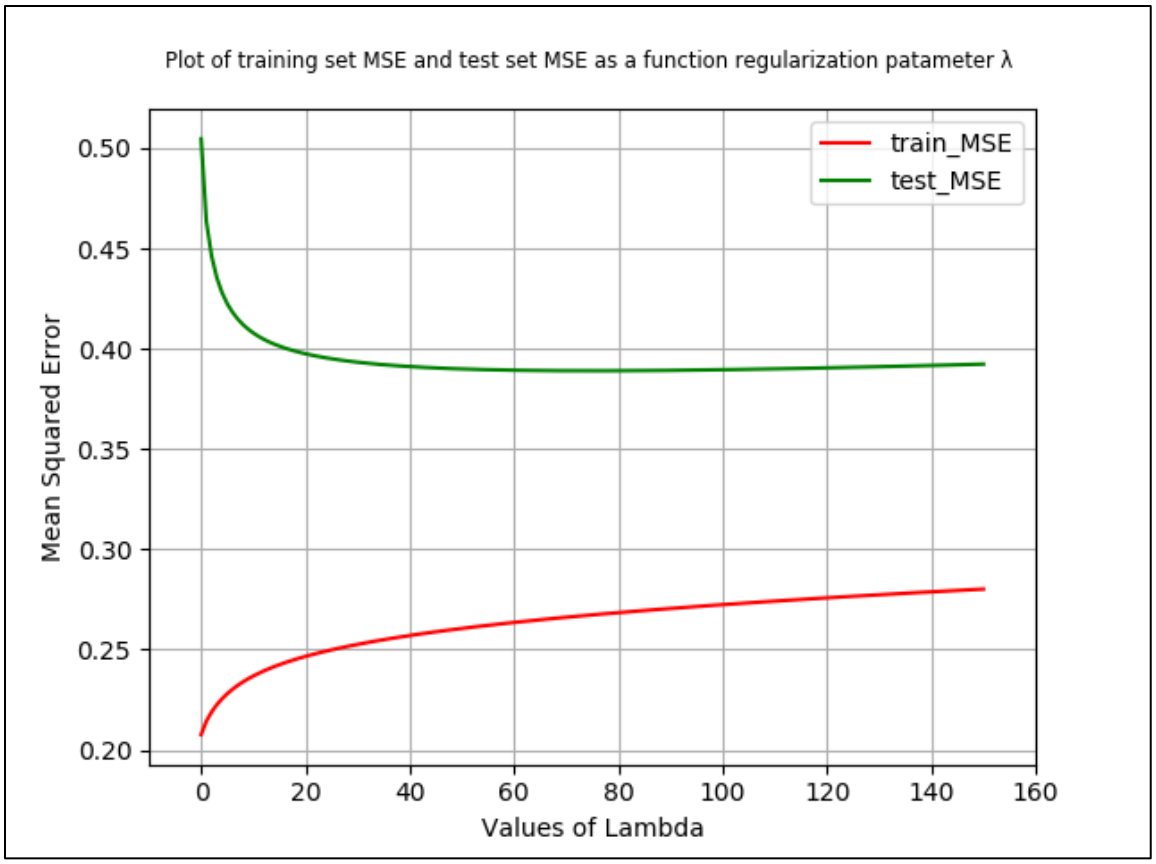


Figure 4: Plot for crime data set

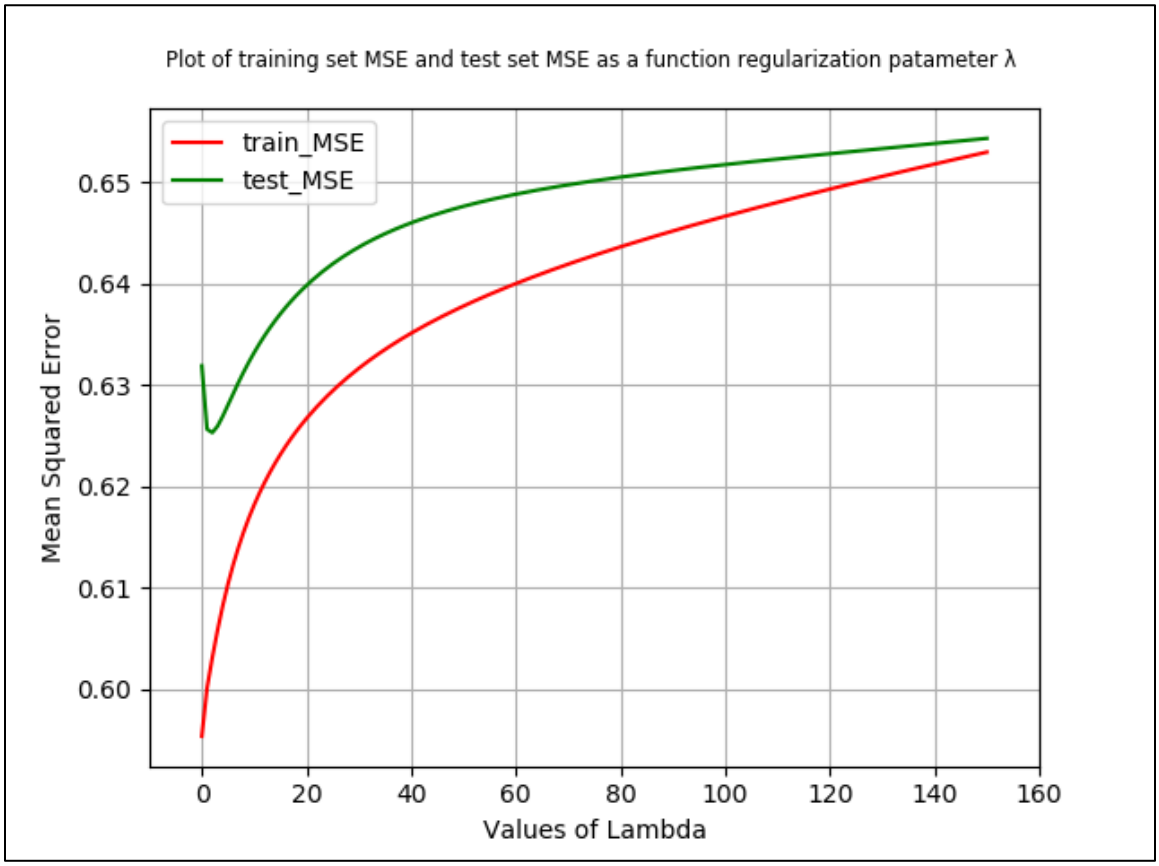


Figure 5: Plot for wine data set

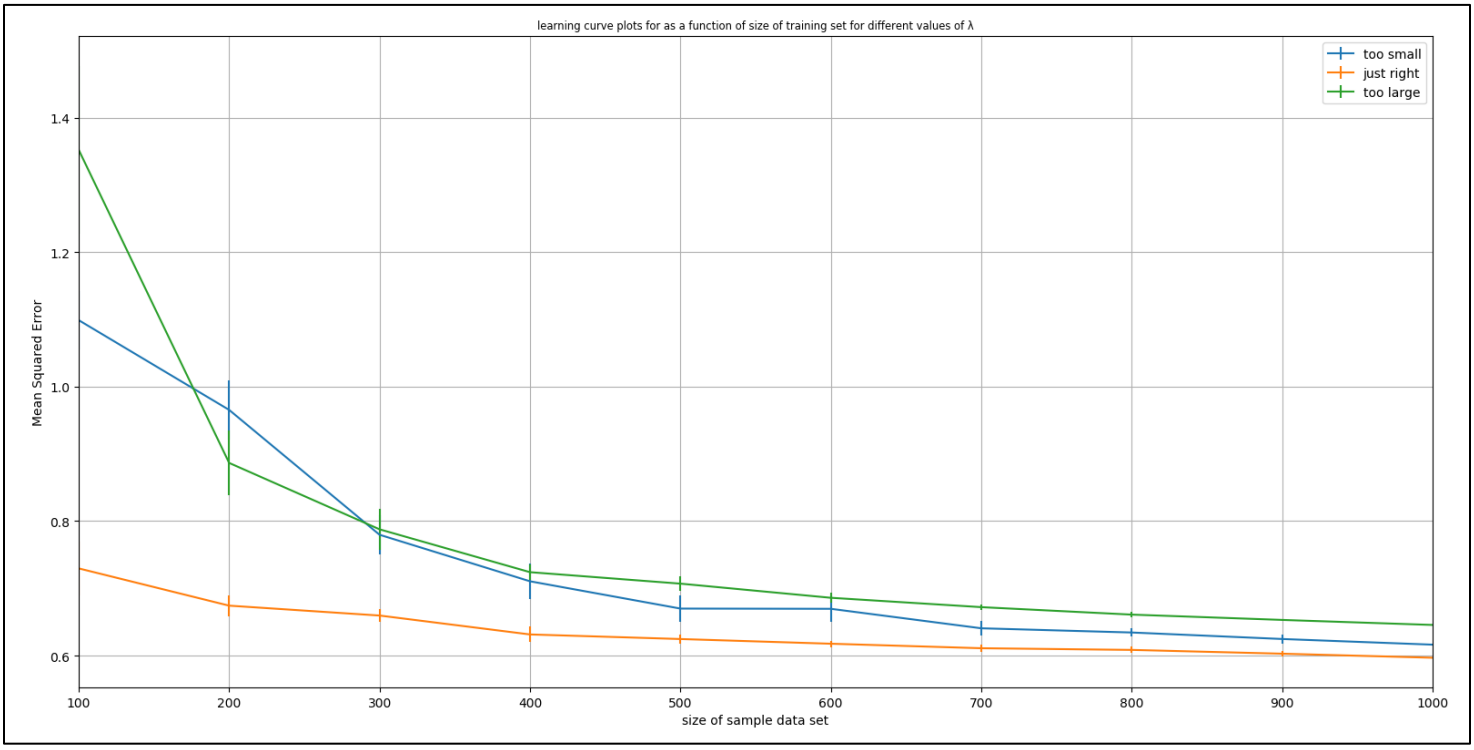
Conclusions:

- As the values of λ increases, the error on test data error will first decrease, then reach a minimum and then will keep on increasing. The minimum point is the best value of λ as we have the lowest MSE in that case.
- Training error is less than the test error because the same data used to fit the model is used to assess its training error, so it easily adapts to training data leading to overfitting. Looking at this in other way, if we use training error to select λ, we will get lowest MSE for λ=0, for which our regularized linear regression will turn into normal linear regression with overfitting problems. Thus, training error should not be used to select the value of λ. The above conclusion is same for all the data sets. The above conclusion holds true for all the data sets.

Task 2: Learning Curves

Learning curves plots performance (in our case MSE) of an algorithm as a function of training data size. For this we work only with 1000-100 (because no. of data points is more compared to other data sets) where we select 3 values of **λ:** "too small", "just right" and "too large" from task 1 performed. To produce these curves, draw random subsets are drawn of the training set (of increasing sizes) and performance is recorded (on the fixed test set) when training on these subsets.

Results:



Too Small (λ_1): 1 Just Right (λ_2): 27 Too Large (λ_3): 150

Conclusions:

Looking at the graph we can see that as the no. of training data increases, MSE decreases which is obvious because with increases the no. of data points the model has more data to learn from which means that the performance of the algorithm is improving. Another observation that can be made is λ_2 has low mean squared error compared to λ_1 and λ_3. For all the values of λ variance decreases with increases in size of sample data set and λ_2 has lowest variance compare to other values of lambda.

Experiment 2:

To investigate two methods for model selection in linear regression (Bayesian Model Selection and cross validation.)

Task 1: 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. Following numerical results were obtained:

Results:

Dataset	Best value of λ	Test Error	Run Time
100-10	15	6.214	0.622
100-100	18	0.7202	1.69
1000-100	23	0.597	11.41
Crime	150	0.392	4.12
Wine	1	0.625	2.75

Following are the results obtained from task 1 of the assignment, corresponds to the graphs shown in Task1:

Dataset	Best value of λ	Test Error
100-10	4	6.084
100-100	18	0.7202
1000-100	27	0.596
Crime	75	0.389
Wine	2	0.625

Conclusion:

From the above tables we can compare the values of **λ** and the test error obtained from two different experiments. We can see that the test error is almost around the same values for both the experiments. Also, comparing the values of lambda, for most of the dataset, we have got the values that are near to each other.

Task 2: Bayesian Model Selection

In this task we consider the formulation of Bayesian Linear Regression using a prior W. We calculate the hyperparameters of W such that the evidence function is maximized. Recall that the evidence function gives a method to pick the parameters α and β, which we can use to find lambda (**λ**).
Λ = α/β. Following numerical results are obtained from this experiment:

Results:

Dataset	Alpha (α)	Beta (β)	Effective λ (=α/β)	Test Error	Run Time
100-10	0.8820	0.1651	5.34	6.087	0.0009
100-100	5.154	3.154	1.63	1.063	0.0072
1000-100	10.28	1.860	5.52	0.608	0.0061
Crime	425.64	3.250	130.95	0.391	0.0047
Wine	6.16	1.609	3.82	0.626	0.0052

Following are the results obtained from part 1 of the assignment, corresponds to the graphs shown in Task1:

Dataset	Best value of λ	Test Error
100-10	4	6.084
100-100	18	0.7202
1000-100	27	0.596
Crime	75	0.389
Wine	2	0.625

Conclusion:

Comparing the results from Task 1 and Task 3.2 we find that there is a drastic difference between the values of λ. The reason for such variation is that in the first part we calculate test error first and then find the best lambda where the MSE is minimum. Whereas in the Task 3.2 we are calculating hyperparameters of the prior W, using which we calculated the effective λ. The hyperparameters α, β are calculated using only the train data set and the test data is unknown to the model. To verify this method of model selection we can look at the test MSE in both the case, we find that the values of test error are comparable.

Task 3: Comparison of both the methods of Model Selection

- In terms of run time The Bayesian Model selection method performs far better than the model selection done using K-fold Cross Validation. While in k-fold Cross Validation it takes over 1500 iterations to find the best value for lambda, in Bayesian Model selection method the values alpha and beta converges in few iterations, giving the optimum value for lambda.
- In term of effective λ and test set MSE comparing the both the methods we don't find any particular trends. While in one case it's Cross Validation has low MSE and in other Bayesian method has lower MSE.
- In condition when no. of features is increased (in our case from 10 to 100) Bayesian Model selection performs better than the Cross-Validation method. And, in condition when no. of features is increased (from 100 to 1000) cross validation method performs better.