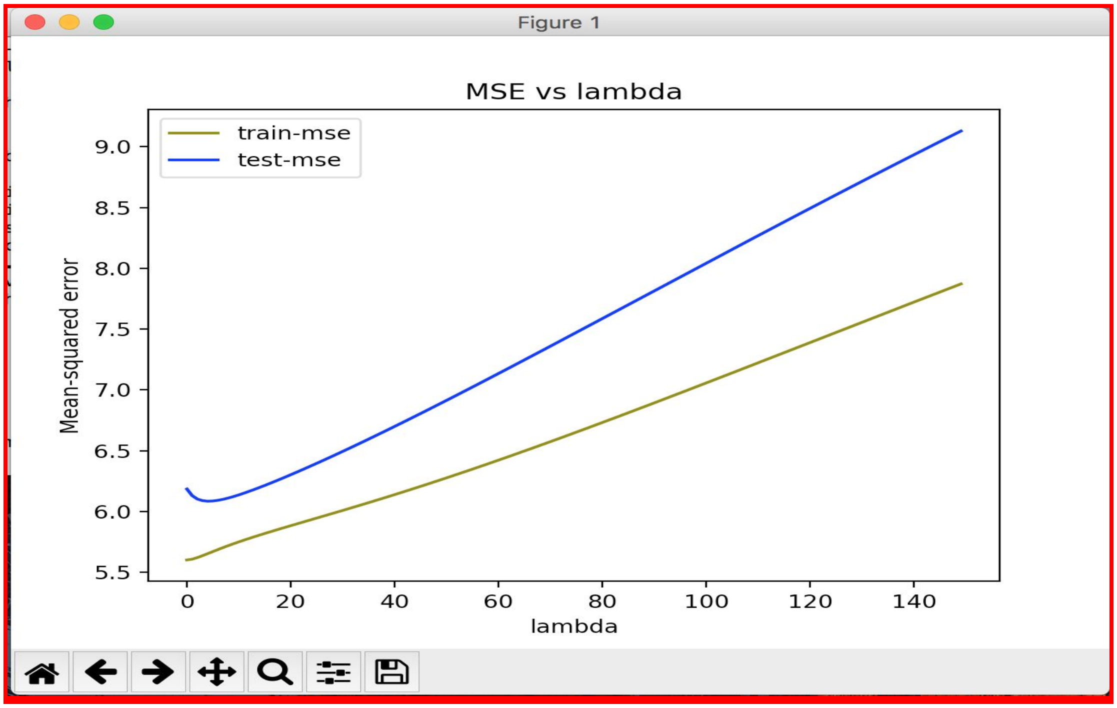
Machine Learning(B555) Programming Project-2:

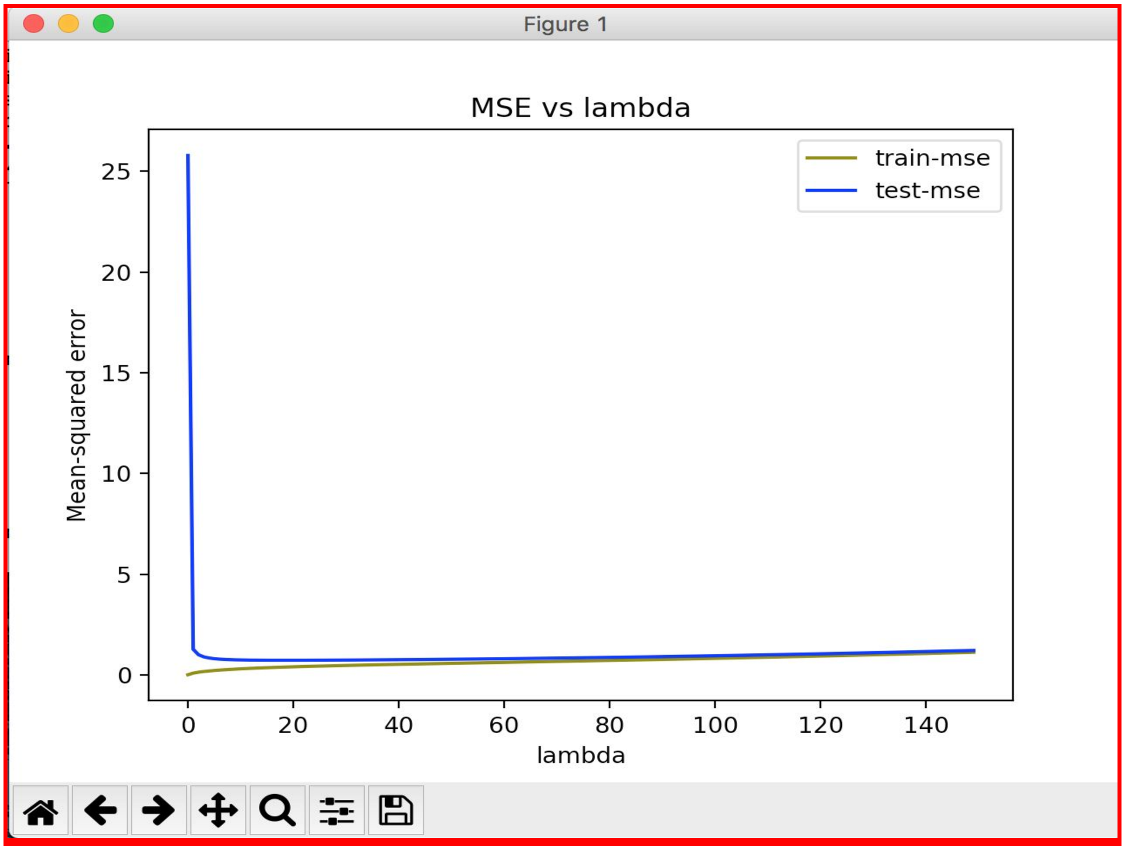
Task 1:Regularisation:

Result for the 100-10 dataset:



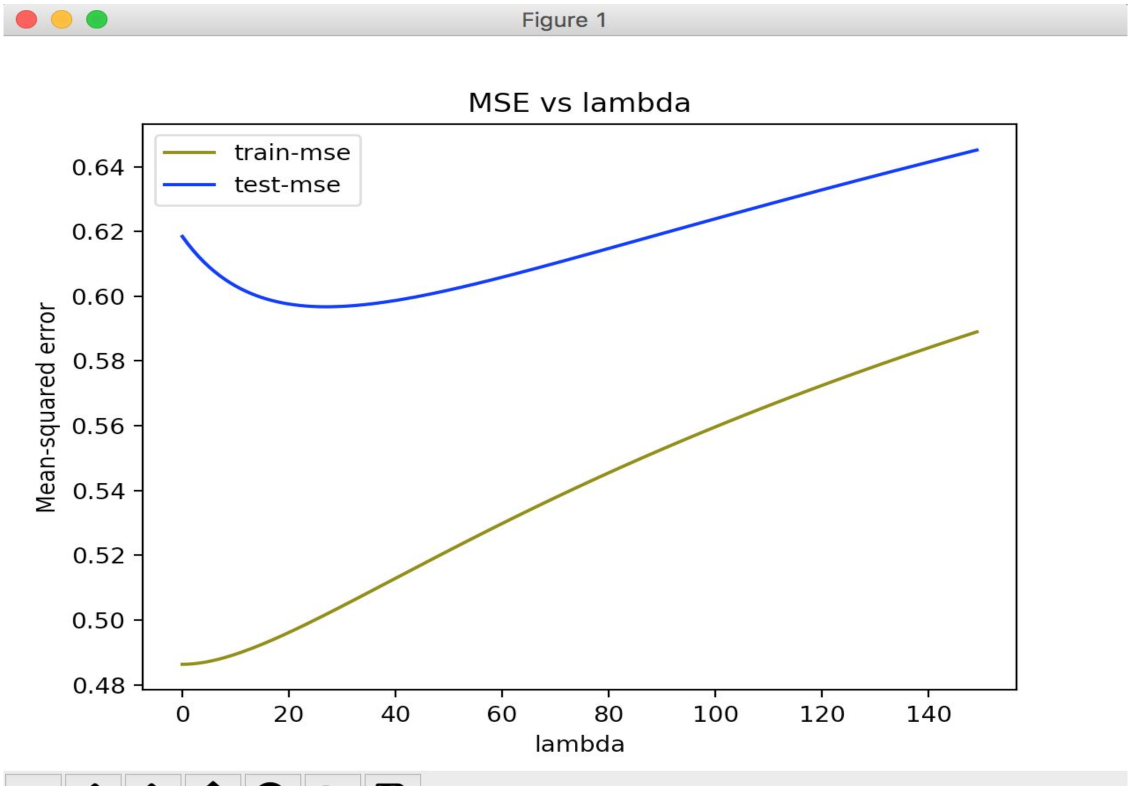
The MSE of the true function seems to be 5.714 and we can observe from The graph above that train set MSE is around 5.7.

Result for the 100-100 dataset:



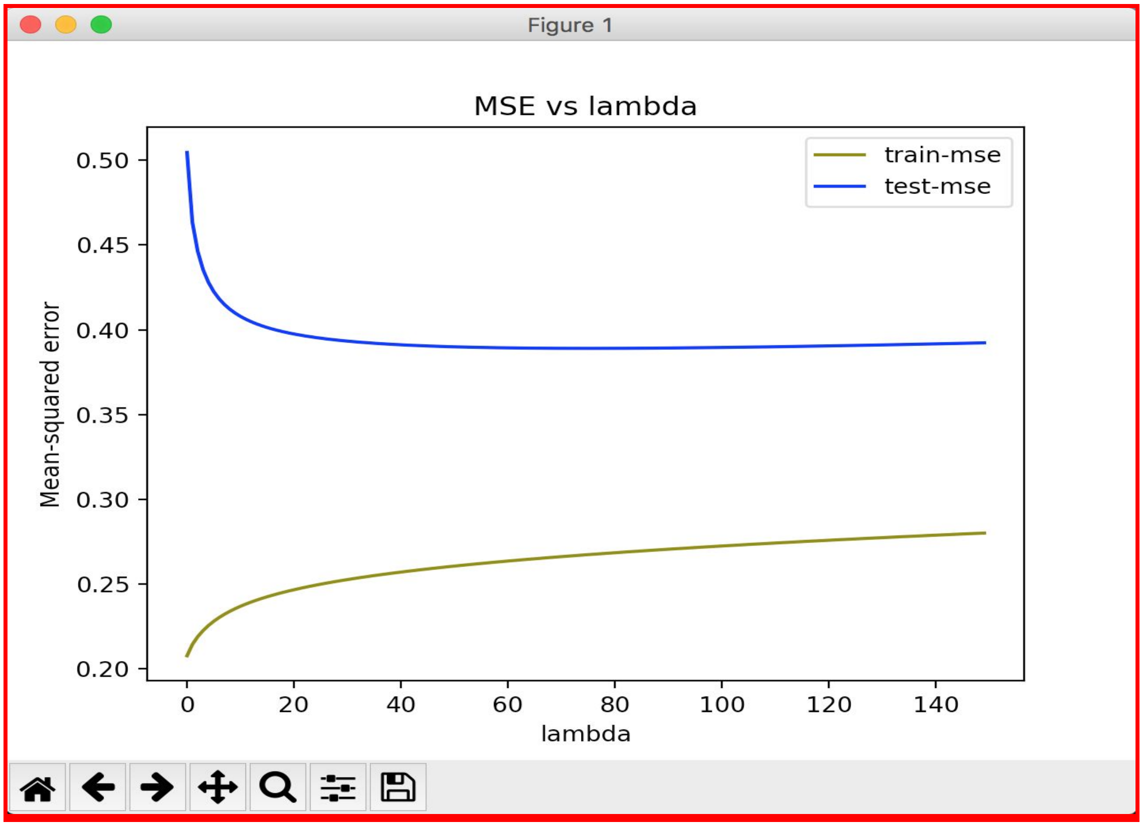
The MSE of the true function is 0.533. From the above plot, this is a little hard to observe. This is due to the fact that at lambda=0 the MSE is really high because of overfitting. The Y-axis intervals are large making it hard to observe a value close to 0.533.

Result for the 1000-100 dataset:

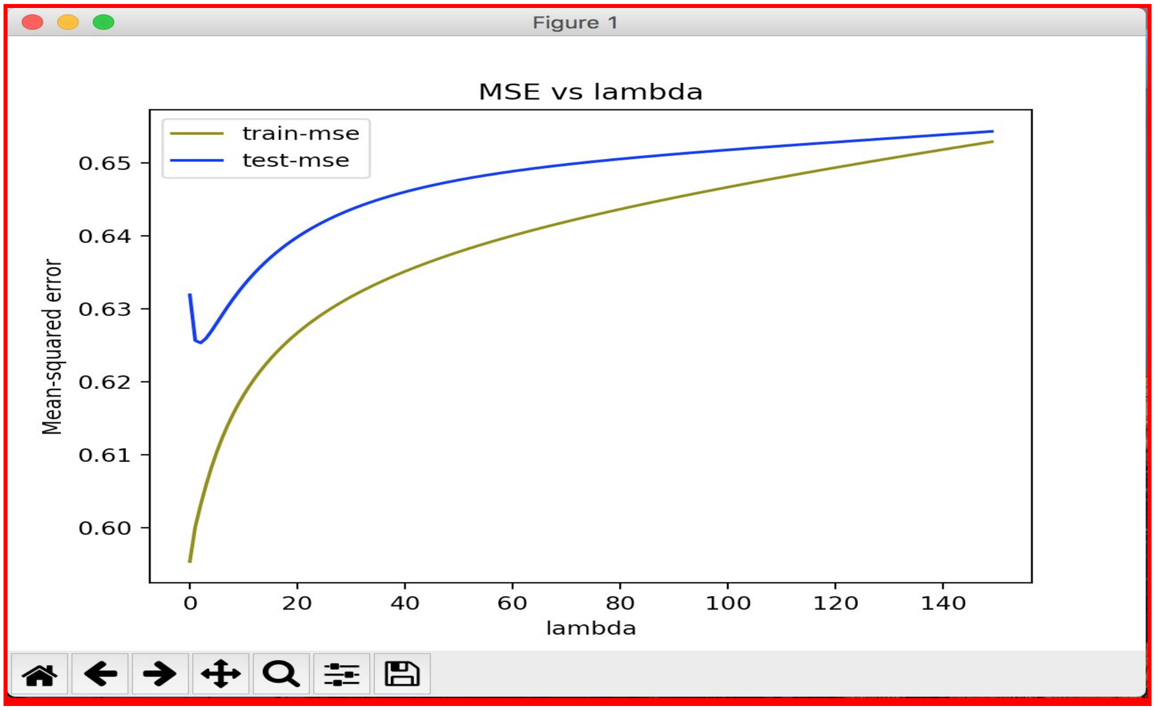


Here again, we see that true function MSE is approximately equal to the MSE obtained from the graph.

Result for the Crime dataset:



Result for the Wine dataset:



Questions for task-1:

1)Why can’t training set MSE be used to select λ?

This is because the hyperparameters have to be trained on unseen data otherwise it will cause our model to overfit. That is if we select the lambda using the train set then it will not be able to generalize to unseen data. For example, at lamda=0 the MSE on the train set is very low but MSE on the test set is pretty high. So we have to find a sweet spot where it performs decently on train and test set.

2)How does λ affect the error on the test set?

The MSE first decreases with an increase in lambda and then MSE increases with further increase in lambda. At lambda values close to zero there is no regularisation so due to overfitting on train set we get high MSE values on the test set. Then MSE reduces till an ideal value and then again starts to increase due to underfitting this time.

3) Does this differ for different datasets?

The variation of lambda for different datasets is different.

4) How do you explain these variations?

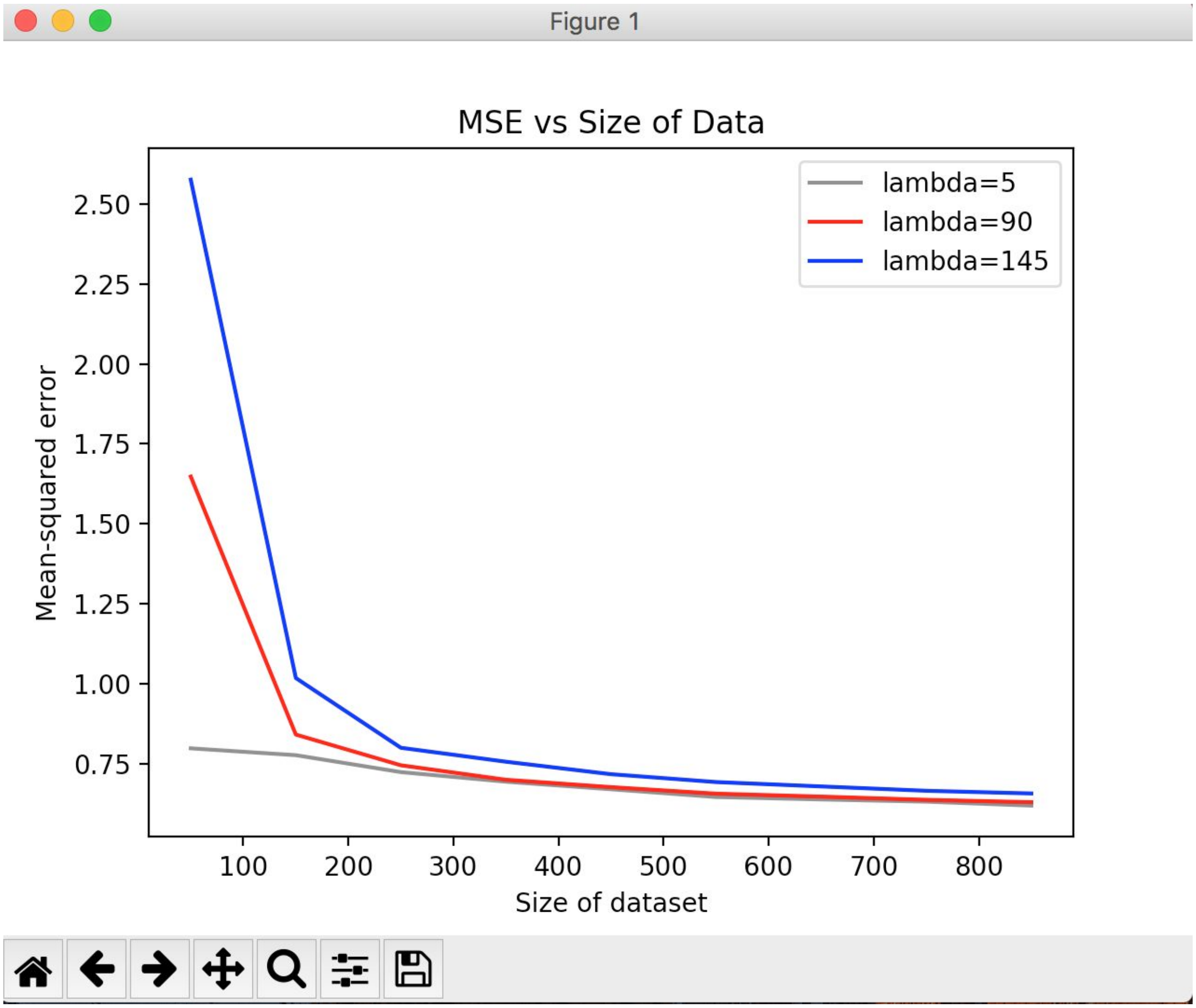
A pattern that can be observed from the graph is for a higher number of features we get a higher optimal lambda value. Also with larger amount of data the performance of the model is better, that is for the optimal lambda value we get a better performance on the MSE.
These variations can be explained using the following reason:

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

The above is the equation for regularised linear regression. In regularised linear regression the lambda parameter is added for making the covariance of the data matrix invertible. As the covariance matrix is different for the different datasets, the best lambda value will be different for different datasets.

Task 2:Learning Curve:

Plotted graphs for 3 values of lambda: lambda=5(too small), lambda= 90(just right), lambda= 145(too high) with the size of the dataset size.



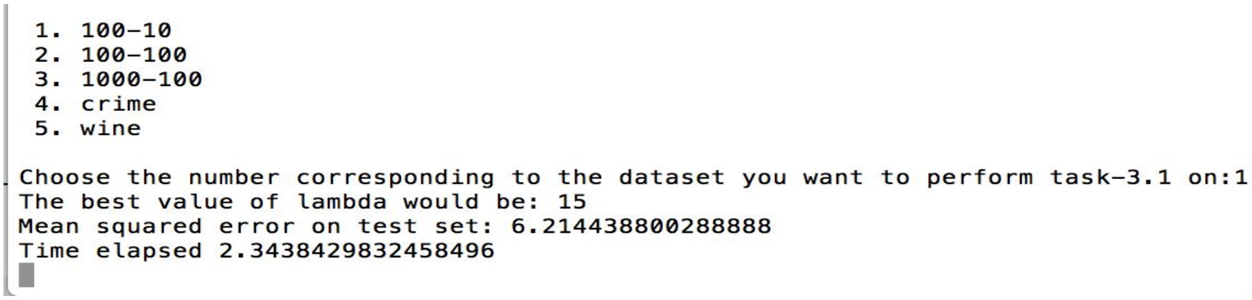
Questions for task-2:

Q)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?

Ans) We observe that for higher lambdas we generally have a more mean squared error for the same dataset. But, as we increase the size of the dataset the mean squared error for all the 3 values of lambda seems to converge to similar values.
This is because for small sizes of dataset the model will tend to underfit hence having higher values of lambda will only worsen this situation.

Task 3.1: Model Selection using Cross-Validation:

Result for the dataset 100-10:

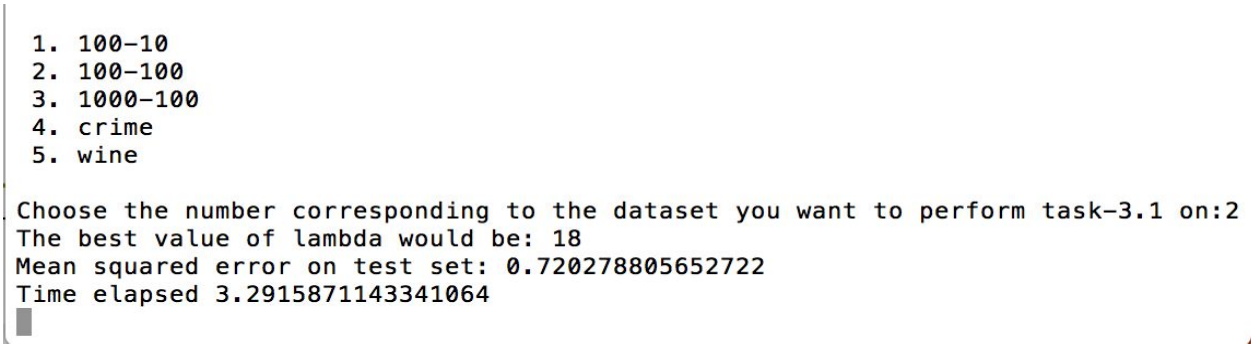


```
 1. 100-10
 2. 100-100
 3. 1000-100
 4. crime
 5. wine

Choose the number corresponding to the dataset you want to perform task-3.1 on:1
The best value of lambda would be: 15
Mean squared error on test set: 6.214438800288888
Time elapsed 2.3438429832458496
```

-> We can see from the result above that the best value of lambda would be 15 and the error on the test set at this point would be 6.2144.

Result for the dataset 100-100:

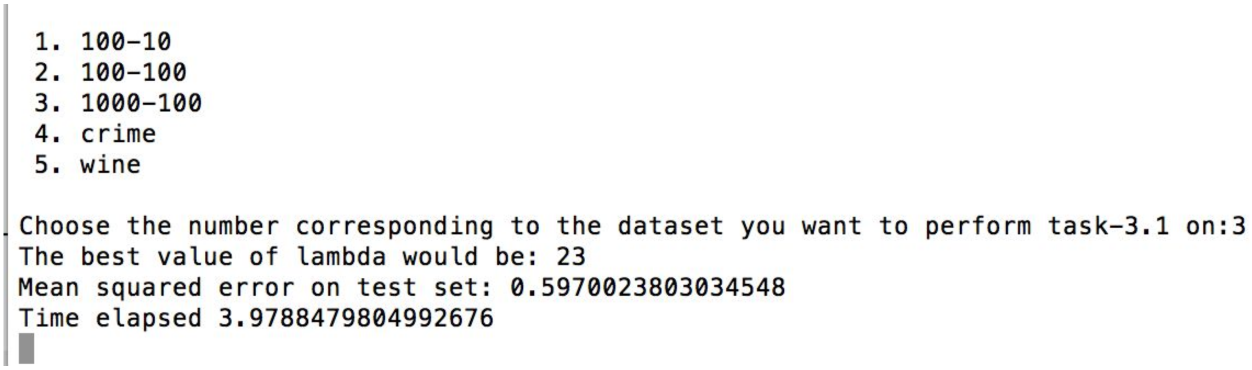


```
 1. 100-10
 2. 100-100
 3. 1000-100
 4. crime
 5. wine

Choose the number corresponding to the dataset you want to perform task-3.1 on:2
The best value of lambda would be: 18
Mean squared error on test set: 0.720278805652722
Time elapsed 3.2915871143341064
```

-> We can see from the result above that the best value of lambda would be 18 and the error on the test set at this point would be 0.72027.

Result for the dataset 1000-100:

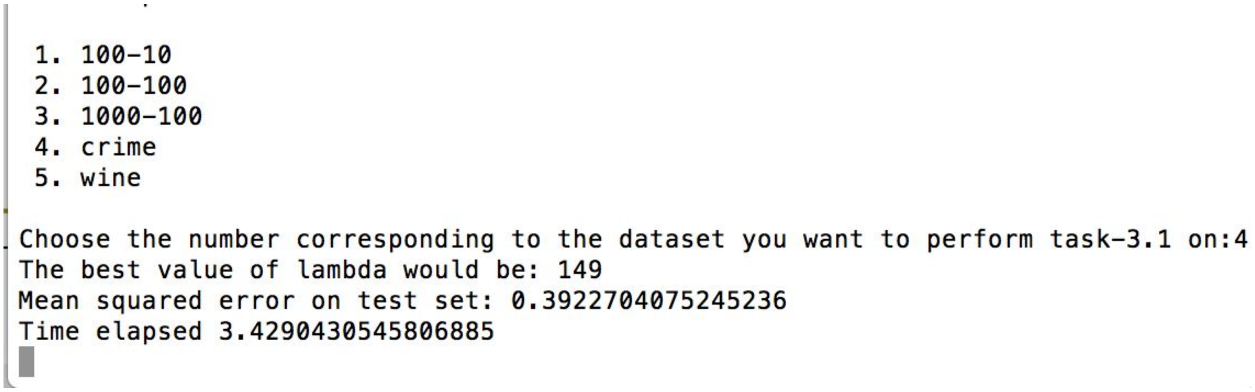


```
 1. 100-10
 2. 100-100
 3. 1000-100
 4. crime
 5. wine

Choose the number corresponding to the dataset you want to perform task-3.1 on:3
The best value of lambda would be: 23
Mean squared error on test set: 0.5970023803034548
Time elapsed 3.9788479804992676
```

-> We can see from the result above that the best value of lambda would be 18 and the error on the test set at this point would be 0.597.

<u>Result for the dataset Crime:</u>

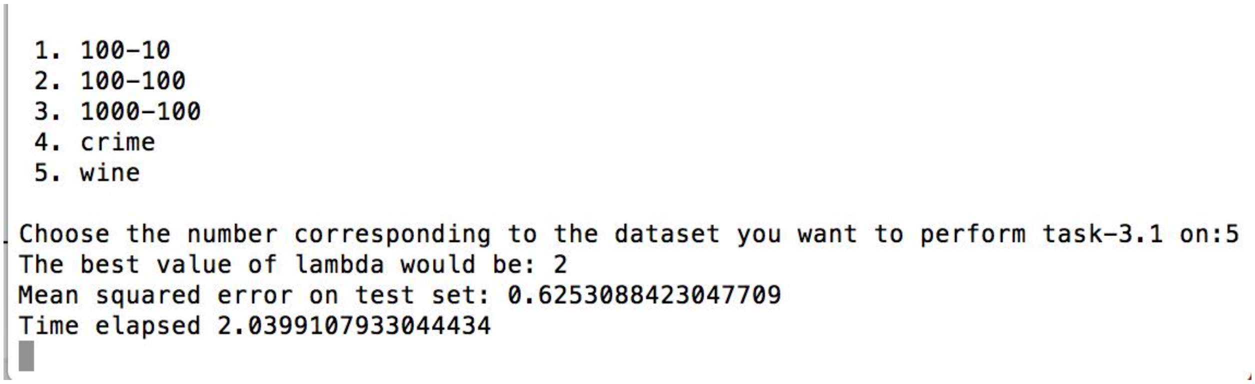


```
 1. 100-10
 2. 100-100
 3. 1000-100
 4. crime
 5. wine

Choose the number corresponding to the dataset you want to perform task-3.1 on:4
The best value of lambda would be: 149
Mean squared error on test set: 0.3922704075245236
Time elapsed 3.4290430545806885
```

-> We can see from the result above that the **best value of lambda would be 149** and the **error on the test set at this point would be 0.39227**

<u>Result for the dataset Wine:</u>



```
 1. 100-10
 2. 100-100
 3. 1000-100
 4. crime
 5. wine

Choose the number corresponding to the dataset you want to perform task-3.1 on:5
The best value of lambda would be: 2
Mean squared error on test set: 0.6253088423047709
Time elapsed 2.0399107933044434
```

-> We can see from the result above that the **best value of lambda would be 2** and the **error on the test set at this point would be 0.62530**.

<u>Questions of Task 3.1:</u>

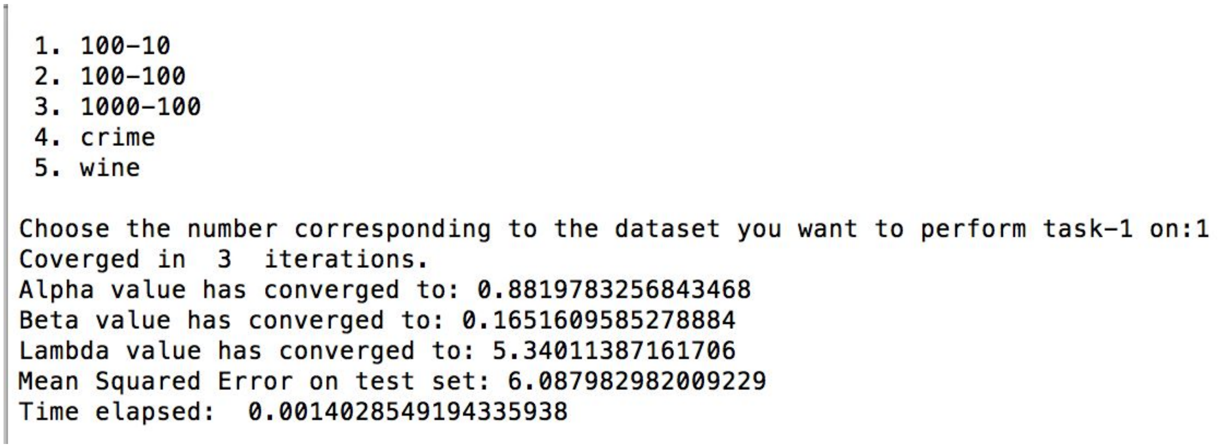
Q)How do the results compare to the best test-set results from part 1 both in terms of the choice of λ and test set MSE?

Ans)By comparing the results obtained in part-1 with these results, we can say that both lambda and MSE values are almost the same in all the cases.

We can see that the run time is pretty high here as there are multiple iterations that must be run for cross-validation..

Task 3.2: Bayesian Model Selection:

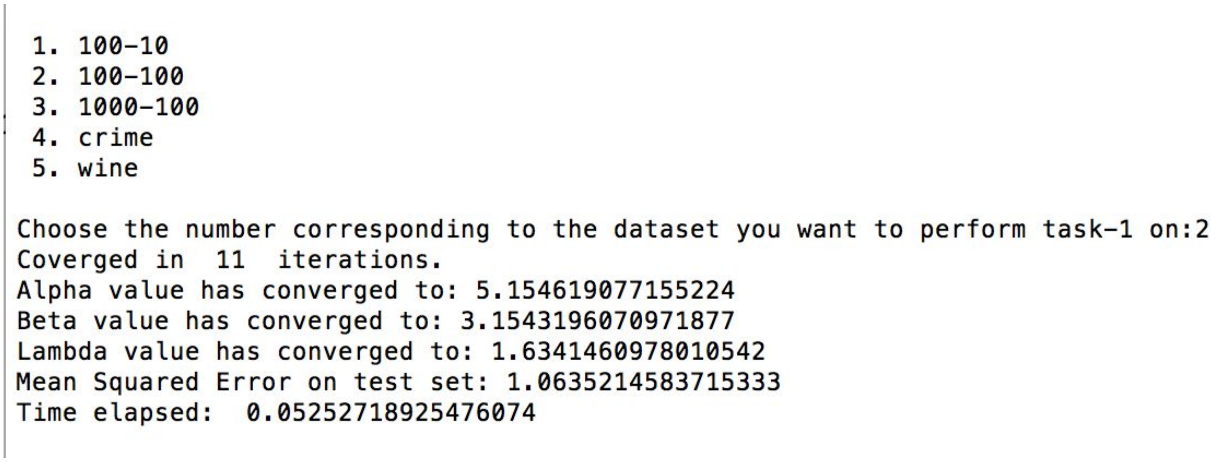
Results on the 100-10 Dataset:



```
 1. 100-10
 2. 100-100
 3. 1000-100
 4. crime
 5. wine

Choose the number corresponding to the dataset you want to perform task-1 on:1
Coverged in  3  iterations.
Alpha value has converged to: 0.8819783256843468
Beta value has converged to: 0.1651609585278884
Lambda value has converged to: 5.34011387161706
Mean Squared Error on test set: 6.087982982009229
Time elapsed:  0.0014028549194335938
```

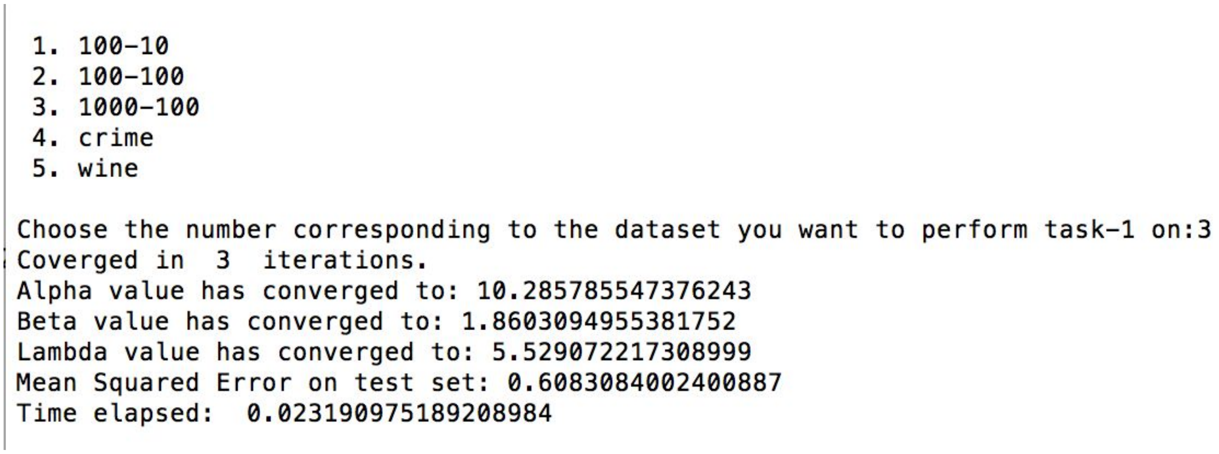
Results on the 100-100 Dataset:



```
 1. 100-10
 2. 100-100
 3. 1000-100
 4. crime
 5. wine

Choose the number corresponding to the dataset you want to perform task-1 on:2
Coverged in  11  iterations.
Alpha value has converged to: 5.154619077155224
Beta value has converged to: 3.1543196070971877
Lambda value has converged to: 1.6341460978010542
Mean Squared Error on test set: 1.0635214583715333
Time elapsed:  0.05252718925476074
```

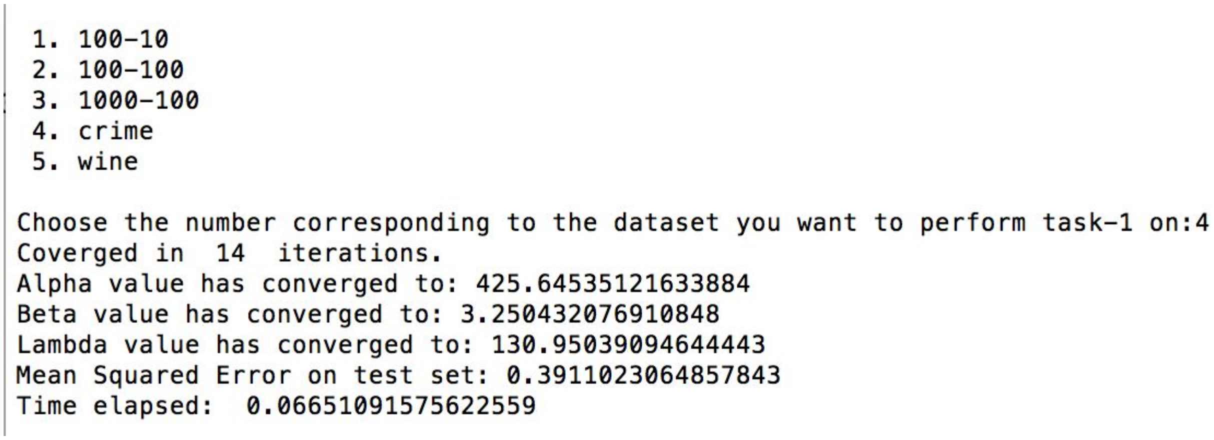
Results on the 1000-100 Dataset:



```
 1. 100-10
 2. 100-100
 3. 1000-100
 4. crime
 5. wine

Choose the number corresponding to the dataset you want to perform task-1 on:3
Coverged in  3  iterations.
Alpha value has converged to: 10.285785547376243
Beta value has converged to: 1.8603094955381752
Lambda value has converged to: 5.529072217308999
Mean Squared Error on test set: 0.6083084002400887
Time elapsed:  0.023190975189208984
```

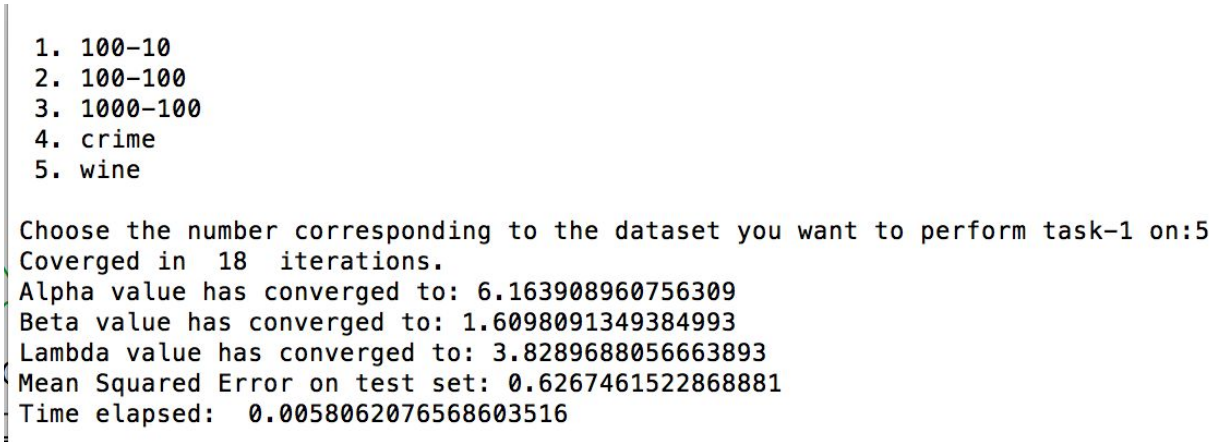
Results on the Crime Dataset:



```
 1. 100-10
 2. 100-100
 3. 1000-100
 4. crime
 5. wine

Choose the number corresponding to the dataset you want to perform task-1 on:4
Coverged in  14  iterations.
Alpha value has converged to: 425.64535121633884
Beta value has converged to: 3.250432076910848
Lambda value has converged to: 130.95039094644443
Mean Squared Error on test set: 0.3911023064857843
Time elapsed:  0.06651091575622559
```

Results on the Wine Dataset:



```
 1. 100-10
 2. 100-100
 3. 1000-100
 4. crime
 5. wine

Choose the number corresponding to the dataset you want to perform task-1 on:5
Coverged in  18  iterations.
Alpha value has converged to: 6.163908960756309
Beta value has converged to: 1.6098091349384993
Lambda value has converged to: 3.8289688056663893
Mean Squared Error on test set: 0.6267461522868881
Time elapsed:  0.0058062076568603516
```

Questions for task 3.2:

Q)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?

Ans)Although there are differences in lambda, most values of MSE values obtained here in the Bayesian model section are the same as those obtained in the part 1.

Though we calculate lambda based on only training set here, we get a performance comparable to part-1.

The Runtime is very low compared to cross-validation model selection. This is mainly because we are not performing cross validation, so we are avoiding a lot of iterations and still getting the optimal λ.

Task 3.3:

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?

Run-time:

In terms of run-time, we can observe that Bayesian model selection converges faster than the cross-validation method.

MSE:

The MSE values obtained from both methods are almost the same for all the datasets. Only in the 100-100 dataset, the MSE value varies a little between the 2 methods.

If we want the runtime to be lower, I believe Bayesian model selection is slightly better than cross-validation method.