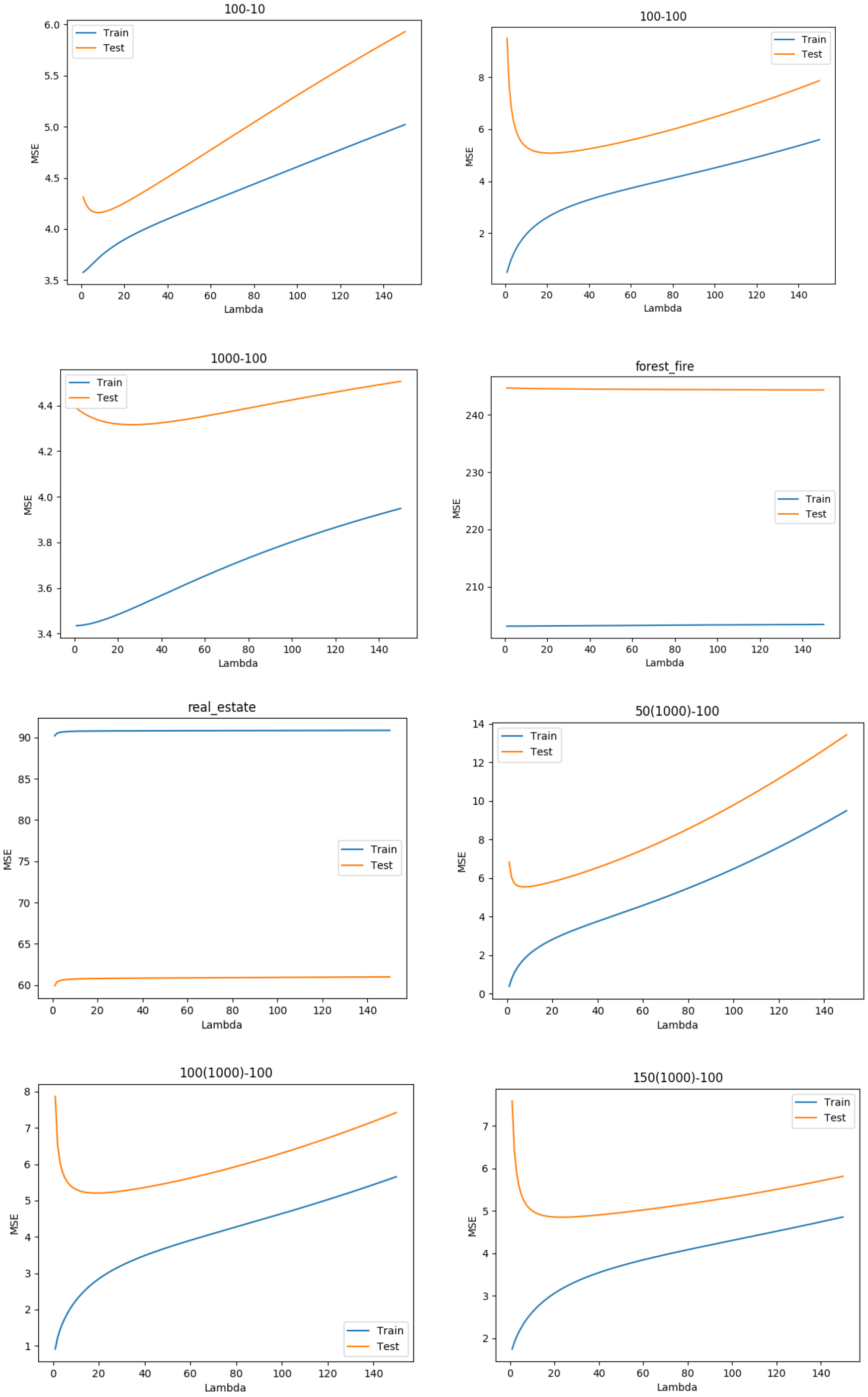
Christian Zinck
15 hours

# Programming Project 2

## Task 1:



*Note: Lambda = 0 has been omitted from the graphs for visibility purposes*

| | Optimal Lambda | Resulting MSE |
|---|---|---|
| 100-10 | 8 | 4.159678509482877 |
| 100-100 | 22 | 5.078299800593872 |
| 1000-100 | 27 | 4.315570630318436 |
| forest_fire | 150 | 244.36480655282236 |
| real_estate | 0 | 52.00475385468231 |
| 50(1000)-100 | 8 | 5.5409022291854075 |
| 100(1000)-100 | 19 | 5.205911957333213 |
| 150(1000)-100 | 23 | 4.848943053347755 |

From the above graphs, it can be clearly seen that the training mean squared error is strictly increasing with respect to lambda in all eight graphs. This makes sense, because the error is being calculated on the same data used to train the regression model, so no regularization will produce optimal MSE. It is for this reason that the training data cannot be used to do model selection on the regularization parameter.

The first three graphs have optimal lambda values of 8, 22, and 27 with resulting test MSE values of 4.16, 5.08 and 4.32, only slightly above the true MSE values of 3.78, 3.78 and 4.015 respectively. From these numbers, two conclusions can be drawn. Firstly, to achieve a low MSE, the number of examples must outnumber the number of features. In other words, as the ratio of examples to features increases, the MSE decreases. Secondly, as the size of the problem increases, the more regularization is needed to produce optimal MSE. The last three graphs corroborate these conclusions because as the number of examples, and thereby the ratio of examples to features, increases, the MSE also decreases. Furthermore, as the size of the data set increases, the more regularization is needed, as the optimal lambda values are 8, 19 and 23 respectively.

The forest fire and real estate graphs show that no amount of parameter optimization will cause a data set that does not fit a linear regression to be modeled appropriately by a linear regression.

Task 2:

| | Optimal Lambda | Resulting MSE |
|---|---|---|
| 100-10 | 5 | 5.92969408967808 |
| 100-100 | 41 | 7.869193641231156 |
| 1000-100 | 150 | 4.50562444416904 |
| forest_fire | 150 | 244.36480655282236 |
| real_estate | 150 | 61.00199308221777 |
| 50(1000)-100 | 25 | 13.432009521283886 |
| 100(1000)-100 | 42 | 7.420858413999264 |
| 150(1000)-100 | 56 | 5.812422230099203 |

In general, the optimal lambda values found using 10 fold cross validation on the training data are larger than the optimal lambda values found using the test data, with the exception of the 100-10 data set. Additionally, the MSE from cross validation is larger than the MSE from test data for all data sets. I propose that this occurs because the number of examples used for testing in 10 fold cross validation is 1/10 of the number of examples used for testing when an entire test set is used.

If we assume that calculating MSE is $O(n^2)$ where n is the larger dimension of the data matrix, then model selection of lambda for using l values and f fold cross validation, then the time complexity is $O(lfn^2)$. Increasing the number of examples, number of features, number of l values, and number of folds will all increase the time complexity, however only increasing the number of examples and number of l values can improve the quality of the model selection, while increasing number of features and number of folds will decrease the quality of the model selection.

Task 3:

| | Optimal Alpha | Optimal Beta | Resulting MSE |
|---|---|---|---|
| 100-10 | 11273999.411778929 | 0.0009183573231667 | 13.431120335707789 |
| 100-100 | 0.2019101166585637 | 0.0006936584682662 | 16.75844768504774 |
| 1000-100 | 0.2188095449302109 | .00007554936025004 | 12.990715137530454 |
| forest_fire | 47931857.26680571 | .00000353394434723 | 687.6097463800669 |
| real_estate | 150152093.99171722 | .00000186250294029 | 1647.0249351789566 |
| 50(1000)-100 | 0.2191820811737866 | 0.0015507221137088 | 15.734034657644061 |
| 100(1000)-100 | 0.219754993498758 | 0.0007552255881888 | 14.906661535175356 |
| 150(1000)-100 | 0.2226629052283457 | 0.0005271623575228 | 13.608836499775556 |

From this table, it can seen that the trends previously discussed continue because a higher ratio of examples to features leads to lower MSE. Additionally, with the exception of the 100-10, all of the artificial data sets have comparatively low MSE and very similar optimal alpha and beta values, which suggests that there is some alpha/beta ratio that is optimal. Compared to the MSE values from tasks 1 and 2, this approach has significantly worse MSE values across the board. This is especially true of the forest fire and real estate data sets, which have already been determined to not fit a linear regression model.

Since the speed of convergence of this model selection algorithm is determined by the initial parameters and the cutoff at which convergence is deemed to be reached, I cannot provide a specific time complexity. However, it seems that each iteration costs O($n^2$) due to matrix operations.

Task 4:

Comparing 10 fold cross validation and Bayesian convergence as model selection methods, it can be seen that 10 fold cross validation results in MSE values closer to the true values. However, cross validation requires MSE to be calculated for each fold and for each lambda value, while the Bayesian method only requires one MSE calculation per data set, making cross validation a much more expensive model in terms of time complexity. For both methods, a high number of examples to number of features ratio will lead to a lower MSE. Also, regardless of model selection, if the data set doesn’t fit a linear regression, the MSE will be high