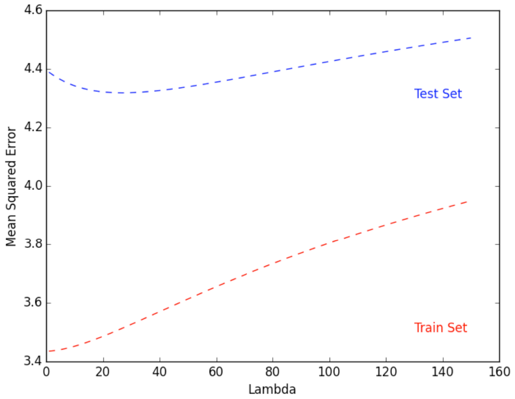
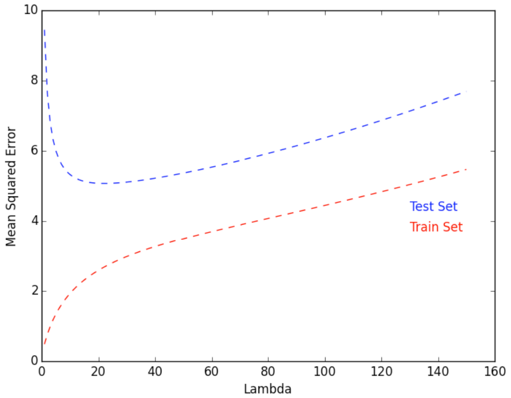
Kevin Soucy
Homework 1
CS6220: Data Mining
Spring 2015

1.

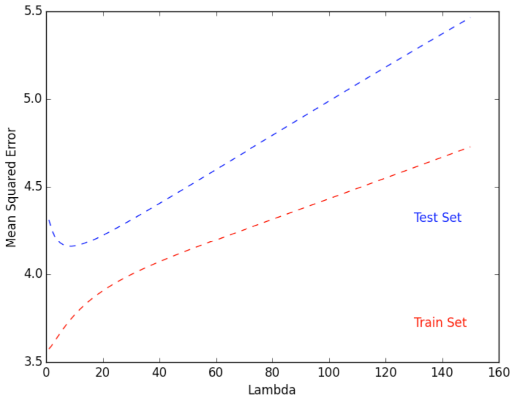
Model	Training Set	Testing Set	Optimum Alpha	Test MSE
1	1000-100-train.csv	1000-100-test.csv	27	4.32
2	50(1000)-100-train.csv	1000-100-test.csv	22	5.07
3	100(1000)-100-train.csv	1000-100-test.csv	9	4.16
4	150(1000)-100-train.csv	1000-100-test.csv	8	5.51
5	100-100-train.csv	100-100-test.csv	19	5.2
6	100-10-train.csv	100-10-test.csv	24	4.84



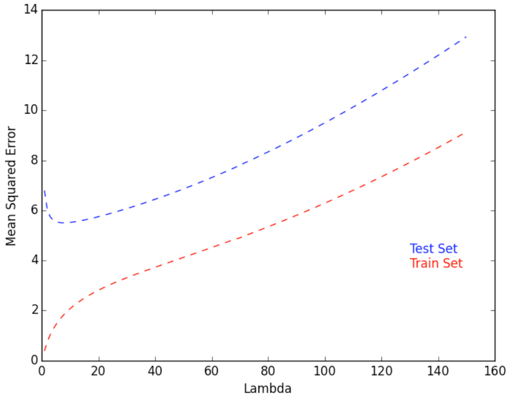
Model 1



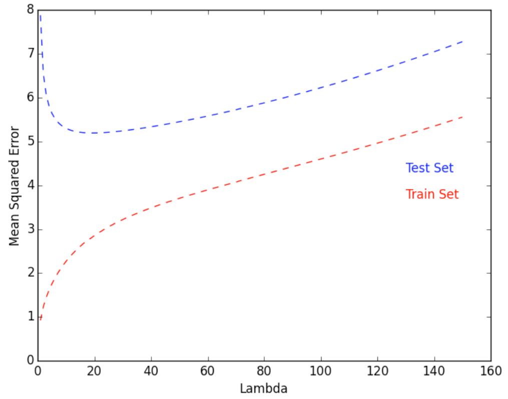
Model 2



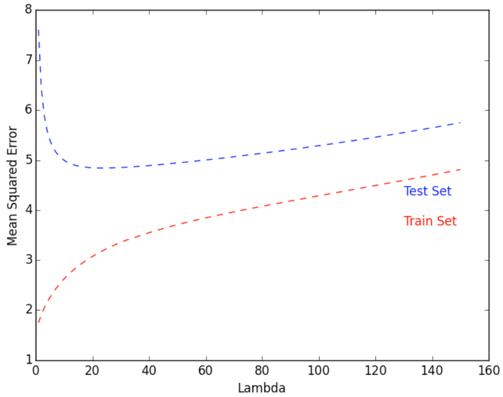
Model 3



Model 4



Model 5

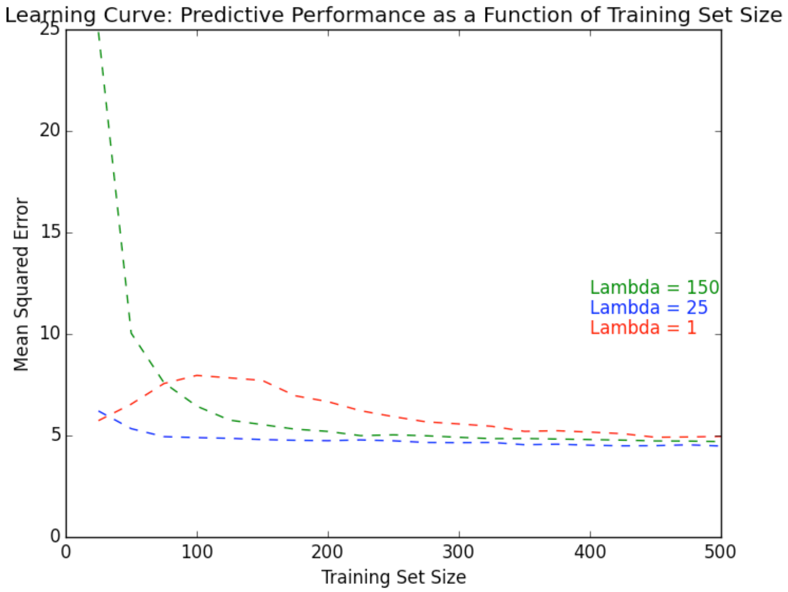


Model 6

At first, from alpha values 0 to around 10 to 20, the testing MSE sharply declines until it reaches an optimal point that minimizes MSE. From this optimal point, the MSE gradually increases as alpha increases. For the training set, MSE increases pretty consistently over the range of alpha values.

When the number of features decreases, such as when comparing Model 2 to Model 3, the optimal alpha value decreases as well. As the number of examples in the training set decreases, such as from Model 1 to 2, the optimal alpha decreases in value while the training MSE increases more sharply for low values of alpha.

2. Fix λ = 1, 25, 150. For each of these values, plot a learning curve for the algorithm using the dataset 1000 100.csv. To produce the curve, you need to draw random subsets (of increasing sizes) and record performance (MSE) on the corresponding test set when training on these sub- sets. In order to get smooth curves, you should repeat the process at least 10 times and average the results.



3. Implement the CV technique given in the class slides. For each dataset, compared the values of λ and MSE with the values in question 1). How do the values for λ and MSE obtained from CV compare to the ones in question 1? What are the drawbacks of CV? What are the factors affecting the performance of CV?

Model	CV Optimum Alpha	CV Test MSE	Q1 Optimum Alpha	Q1 Test MSE
1	29	4.26	27	4.32
2	9	5.66	22	5.07
3	7	4.12	9	4.16
4	8	6.37	8	5.51
5	15	5.64	19	5.2
6	20	5.17	24	4.84

On average, the CV alpha values are much lower than the method used in question 1 but the testing MSE values are on average slightly higher using cross validation. Cross validation is useful when you don't have much data and need to make the most out of the data you have without overfitting. The most important factor affecting CV is k. If k is too high, then regression coefficients will not be much different from those calculated using the entire dataset.

One drawback I noticed for CV is that the optimum alpha changes each time the CV is run. This is because each random sampling results in a different output and that depending on the random number generator's seed, the optimum alpha could vary by as much as 10 or more in either the positive or negative direction.