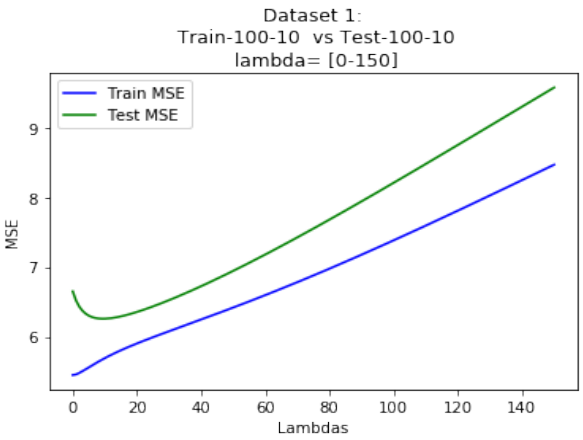
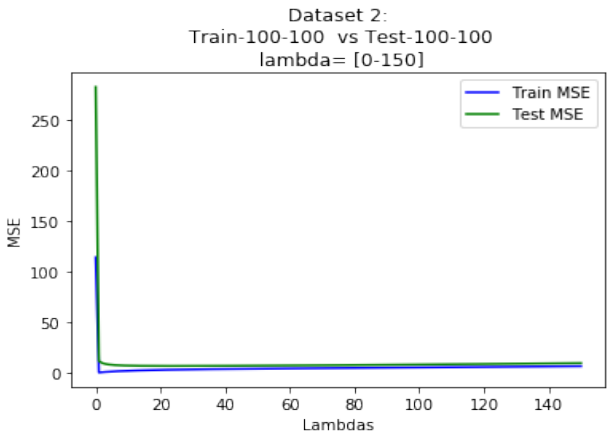
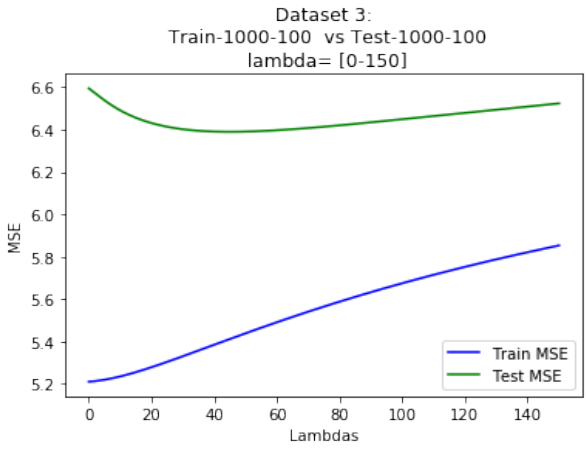
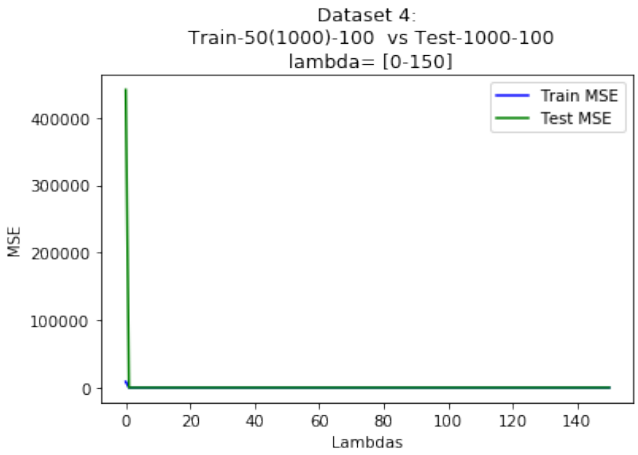
Angela Krontiris
CISC 6930: Data Mining
Assignment 1
Due: September 28

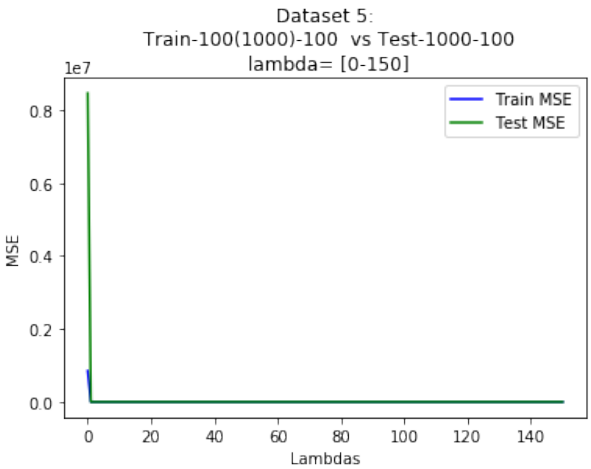
1) Below are the plots for both the train and test MSE as a function of lambda for 6 datasets. The lambda values range from 0 to 150.

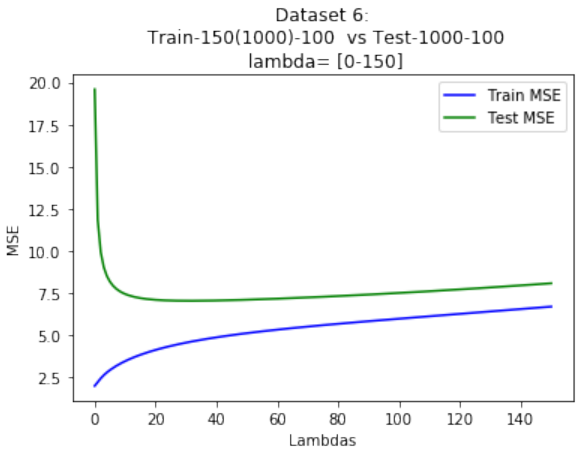








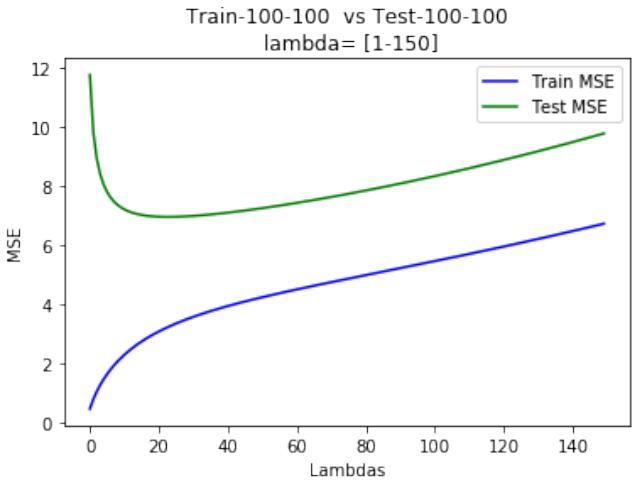


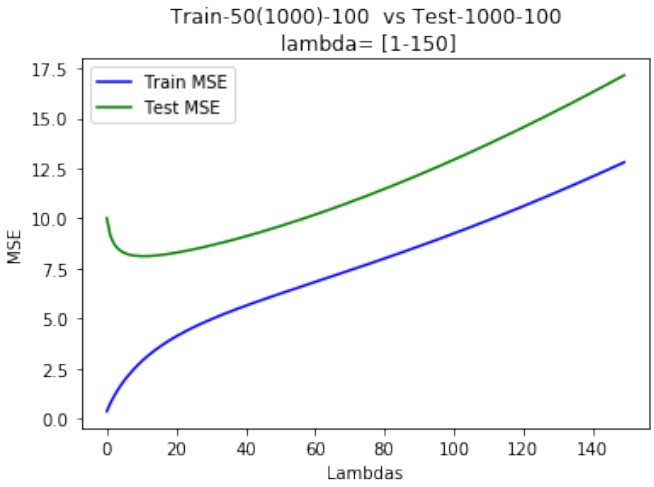


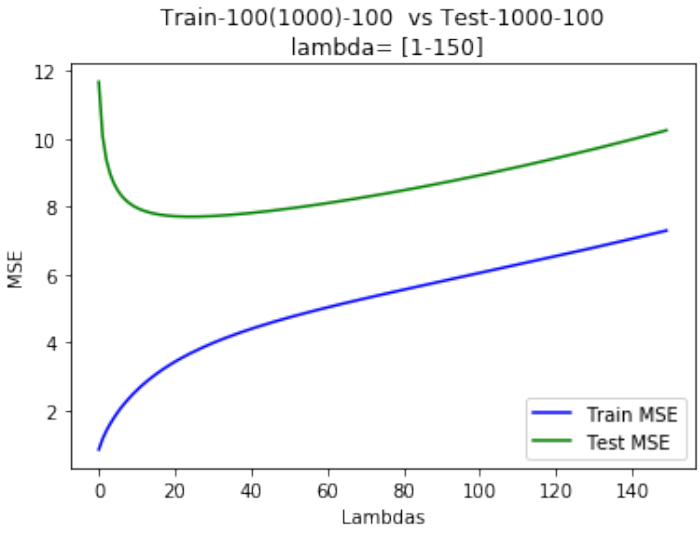
1a) For each dataset, which lambda gives the least test set MSE?

Dataset	MSE (Test set)	Lambda Value
100-10	6.258	9
100-100	6.965	24
1000-100	6.389	45
50(1000)-100	8.112	12
100(1000)-100	7.699	25
150(1000)-100	7.052	31

1b) For each of the datasets, 100-100, 50(1000)-100, 100(1000)-100, provide an additional graph with lambda ranging from 1 to 150







1c) Explain why lambda=0)i.e., no regularization) gives abnormally large MSEs for those three datasets in (b).

Setting lambda to zero removes regularization completely. In this case, training focuses exclusively on minimizing loss, which poses the highest possible over fitting risk.

2a) Using the CV technique, what is the best choice of lambda value and the corresponding test MSE for each of the six datasets?

Dataset	MSE (Test set)	Lambda Value
100-10	6.266	11
100-100	5.150	11
1000-100	6.265	59
50(1000)-100	7.456	13
100(1000)-100	5.888	15
150(1000)-100	6.829	38

2b) How do the values for lambda and MSE obtained from CV compare to the choice of lambda and MSE in question 1(a)?

In part 1a, the choice of lambdas range from 9 to 31. The least test MSE came out to 6.258 for dataset 1 (100-10) when lambda = 9. The lambda values obtained from CV range higher, from 11 to 59. After CV, the least test MSE came out to 5.150 for dataset 2 (100-100) when lambda = 11.

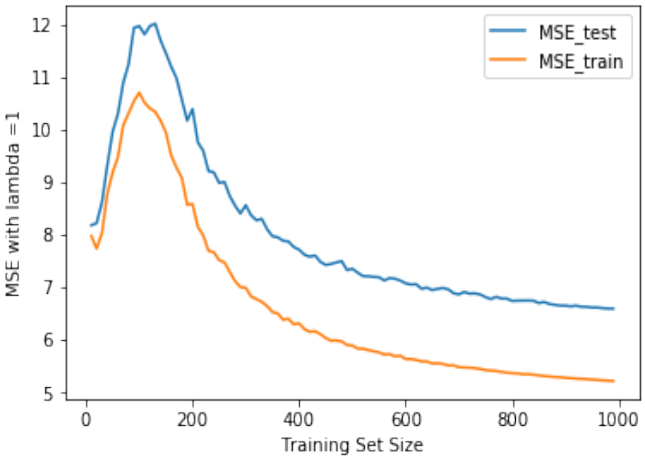
2c) What are the drawbacks of CV?

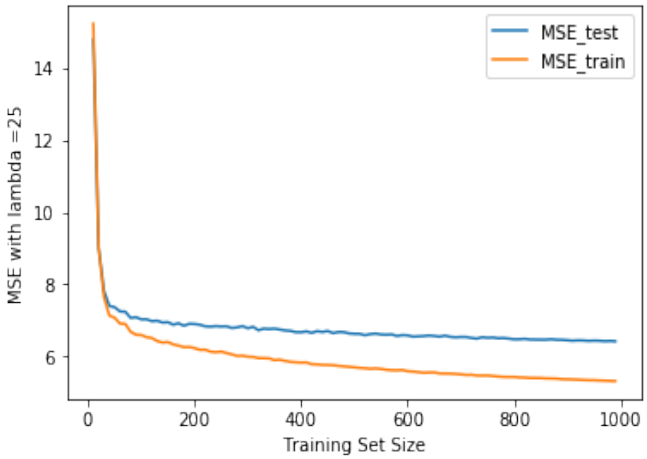
The disadvantage of using CV is that the training algorithm has to be rerun from scratch *k* times, which means it takes *k* times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set *k* different times.

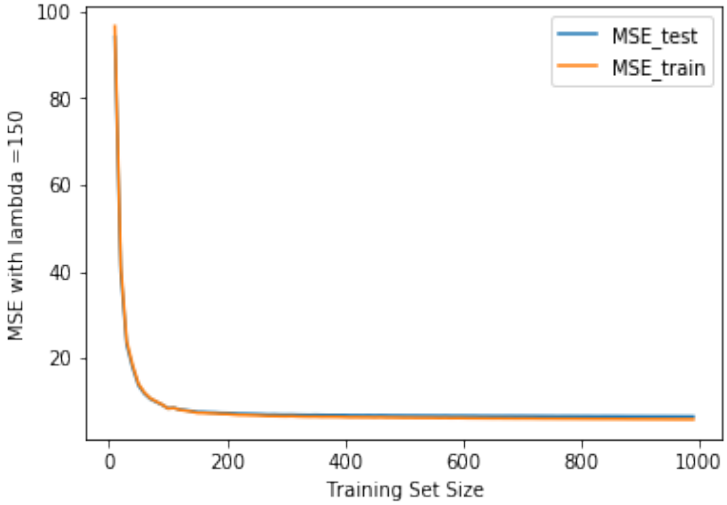
2d) What are the factors affecting the performance of CV?

The two factors affecting the performance measure of CV are the training and test set. The training set affects the measurement indirectly through the learning algorithm, whereas the composition of the test set has a direct impact on the performance measure. There must not be any overlap between the data used for learning and the data used for validation in the same run.

3) Fix lambda= 1, 25, 150. For each of these values, plot a learning curve for the algorithm using the dataset 1000-100.







As lambda increased, the algorithm performed the best on the 1000-100 dataset.