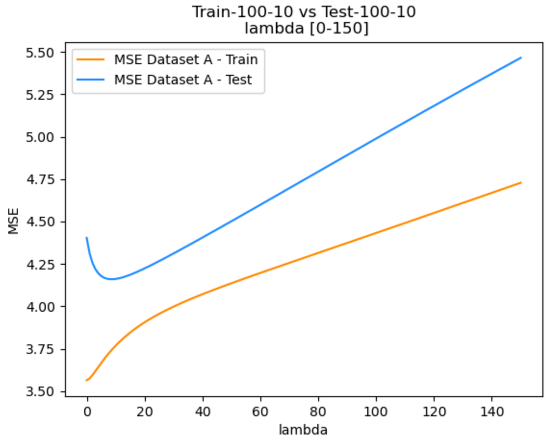
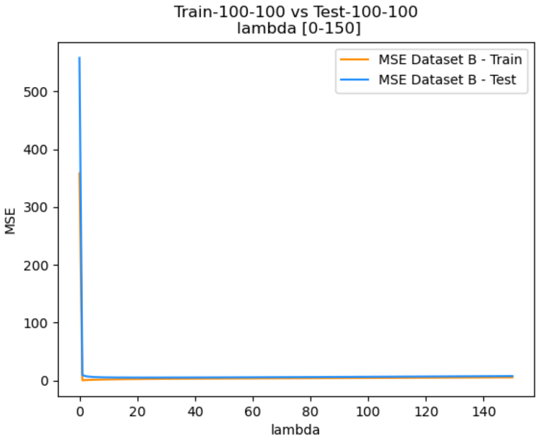
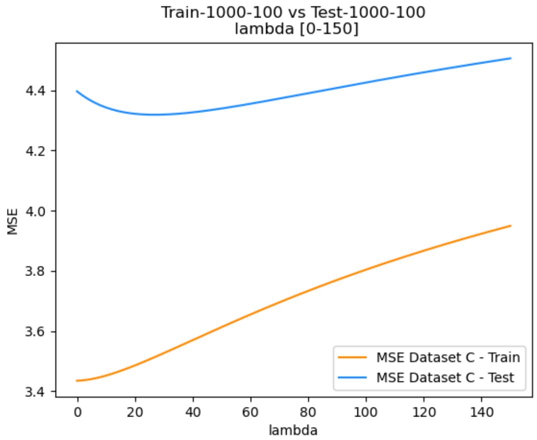
Luisa Rosa
CISC5790
Dr. Zhao
February 20th, 2024

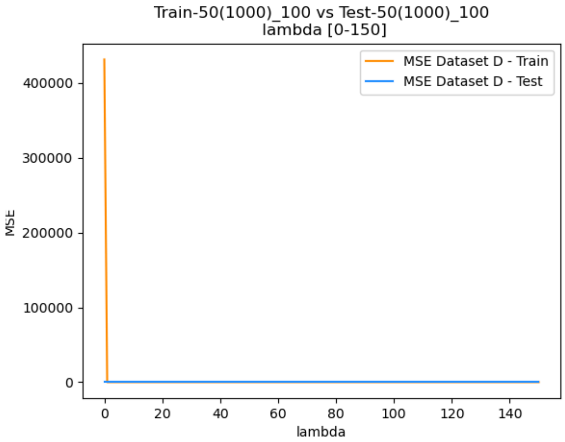
Homework #1 - Data Mining

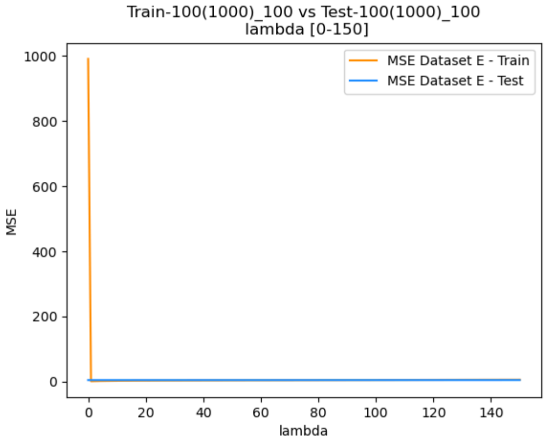
1) Implement L2 regularized linear regression algorithm with λ ranging from 0 to 150(integers only). For each of the 6 datasets, plot both the training set MSE and the test set MSE as a function of λ (x-axis) in one graph.

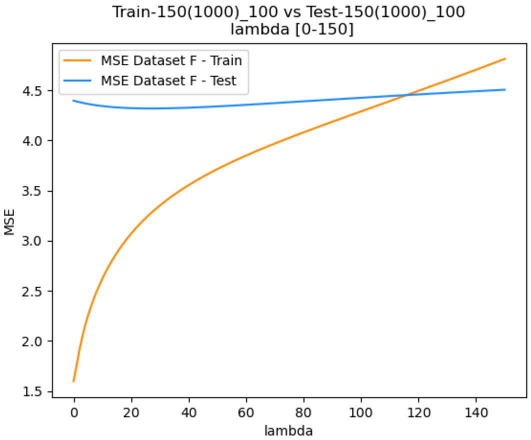

Train-100-10 vs Test-100-10
lambda [0-150]


Train-100-100 vs Test-100-100
lambda [0-150]


Train-1000-100 vs Test-1000-100
lambda [0-150]


Train-50(1000)_100 vs Test-50(1000)_100
lambda [0-150]


Train-100(1000)_100 vs Test-100(1000)_100
lambda [0-150]


Train-150(1000)_100 vs Test-150(1000)_100
lambda [0-150]

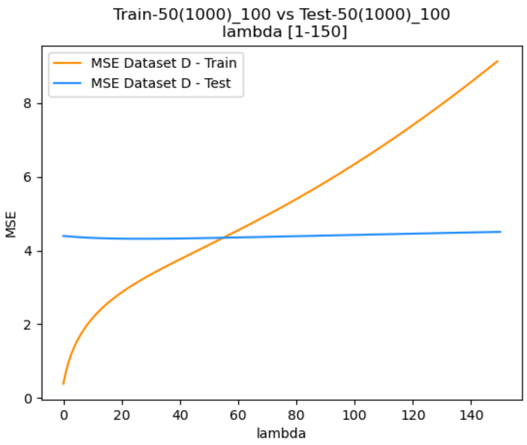
Dataset A = 100-10
Dataset B = 100-100
Dataset C = 1000-100

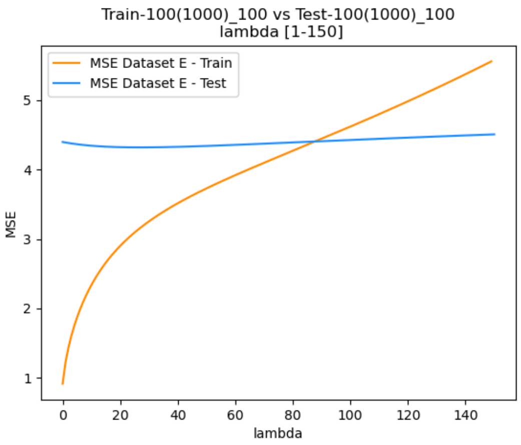
Dataset D = 50(1000)-100
Dataset E = 100(1000)-100
Dataset F = 150(1000)-100

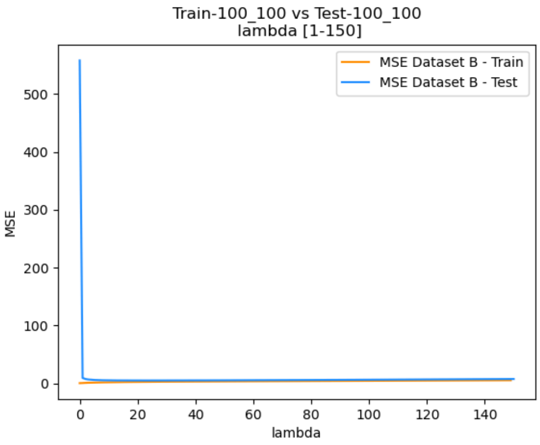
a) For each dataset, which λ value gives the least test set MSE?

- For the test dataset A **lambda 9** gives the least **MSE 4.159663927778 0625**
- For the test dataset B **lambda 22** gives the least **MSE 5.072750457735282**
- For the test dataset C **lambda 27** gives the least **MSE 4.318370456639974**
- For the test dataset D (lambda: 0-150) **lambda 8** gives the least **MSE 5.512273909883558**
- For the test dataset E (lambda: 0-150) **lambda 19** gives the least **MSE 5.196199710503634**
- For the test dataset F (lambda: 0-150) **lambda 24** gives the least **MSE 4.843720381414155**

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







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

When lambda = 0, there is no regularization, which means that there is no "penalty term" for model complexity and training focuses exclusively on minimizing loss. This leads to a high risk of overfitting. Another important thing to keep in mind is that a small evaluation set, gives innacurate estimated error.

2) From the plots in question 1, we can tell which value of λ is best for each dataset once we know the test data and its labels. This is not realistic in real-world applications. In this part, we use cross-validation (CV) to set the value for λ. Implement the 10-fold CV technique discussed in class (pseudo code given in Appendix A) to select the best λ value from the training set.

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

- For the test dataset **A lambda 13** gives the least **MSE 4.186549495447378**
- For the test dataset **B lambda 20** gives the least **MSE 4.466572219197872**
- For the test dataset **C lambda 39** gives the least **MSE 4.139641074529679**
- For the test dataset **D lambda 24** gives the least **MSE 5.285221355859347**
- For the test dataset **E lambda 31** gives the least **MSE 4.852209825819767**
- For the test dataset **F lambda 47** gives the least **MSE 4.876912890852046**

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

The lambda and MSE values we generated on question 1(a) differed with the use of cross-validation. In question 1, we will know which value of λ is best for each dataset once we know the test data and its labels. However, with a 10 fold CV we are finding the best λ value from the training set. CV provides a more robust and data-driven approach to selecting the optimal lambda value. It takes into account the dataset's characteristics and helps avoid overfitting.In question 1, lambda varied from 9 to 27, while in question 2, lambda varied from 13 to 47.

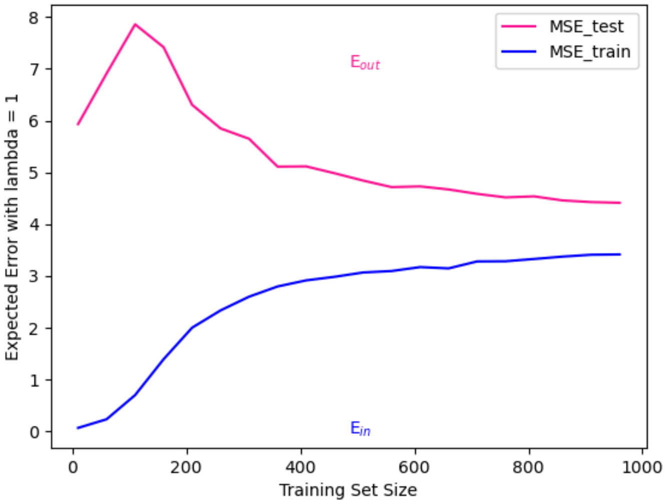
c) What are the drawbacks of CV?

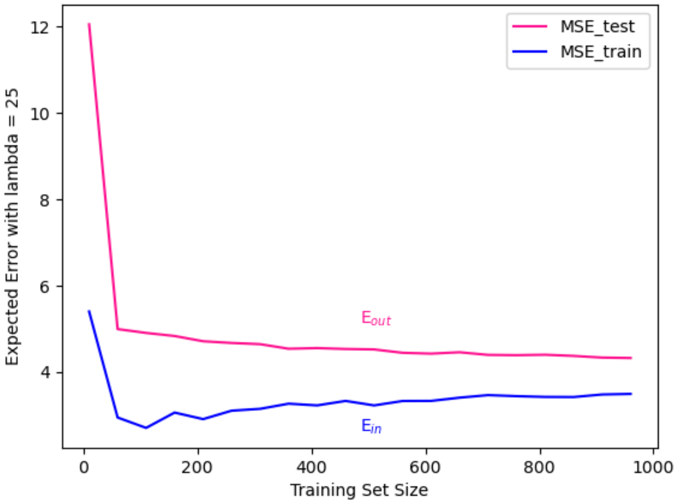
Cross Validation increases computational cost and time, as it requires training and testing the model k times. Cost of Computation = K folds x choices of lambda.

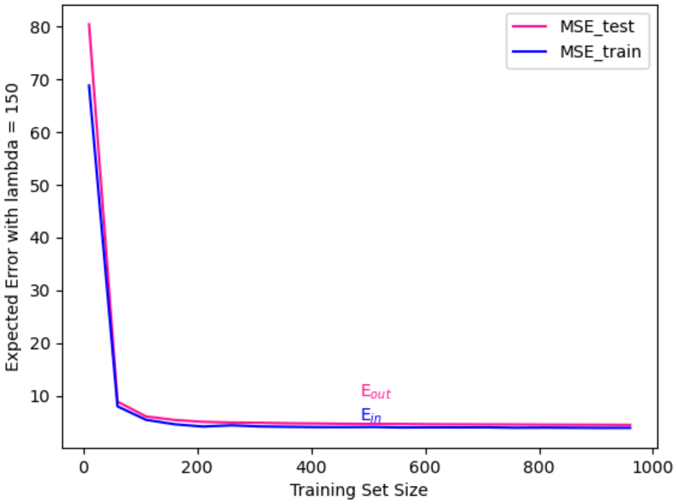
d) What are the factors affecting the performance of CV?

The performance of CV can be affected by several factors, including the choice of CV technique (k-fold, leave-one-out), the size of the dataset, the variability and complexity of the model being evaluated, and the presence of error or missing data (noise).

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


MSE_test
MSE_train
E_{out}
E_{in}
Expected Error with lambda = 1
Training Set Size


MSE_test
MSE_train
E_{out}
E_{in}
Expected Error with lambda = 25
Training Set Size


MSE_test
MSE_train
E_{out}
E_{in}
Expected Error with lambda = 150
Training Set Size