CS6220-01 Data Mining Techniques – Fall 2016

Assignment 1

Name: Qian Cao

Date: Oct. 2nd,2016

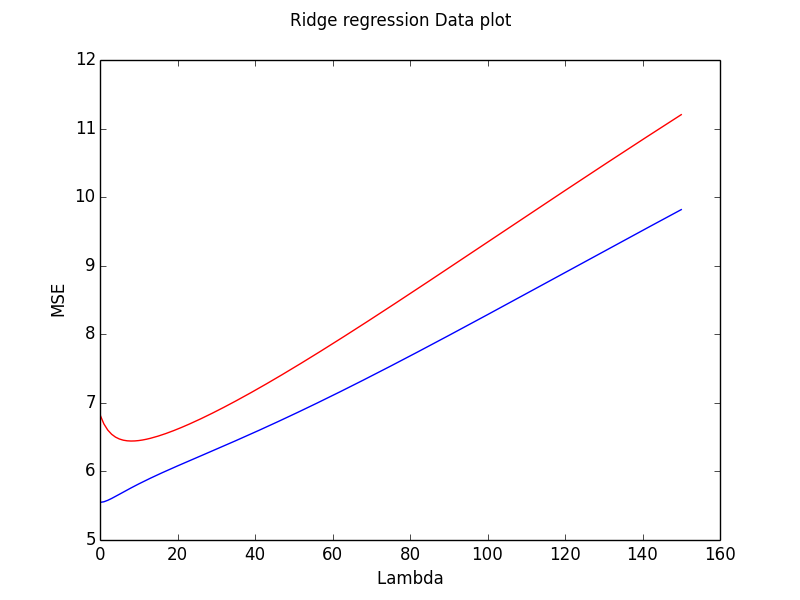
***Note: All red lines in image is for test data,***

***and all blue lines is for train data.***

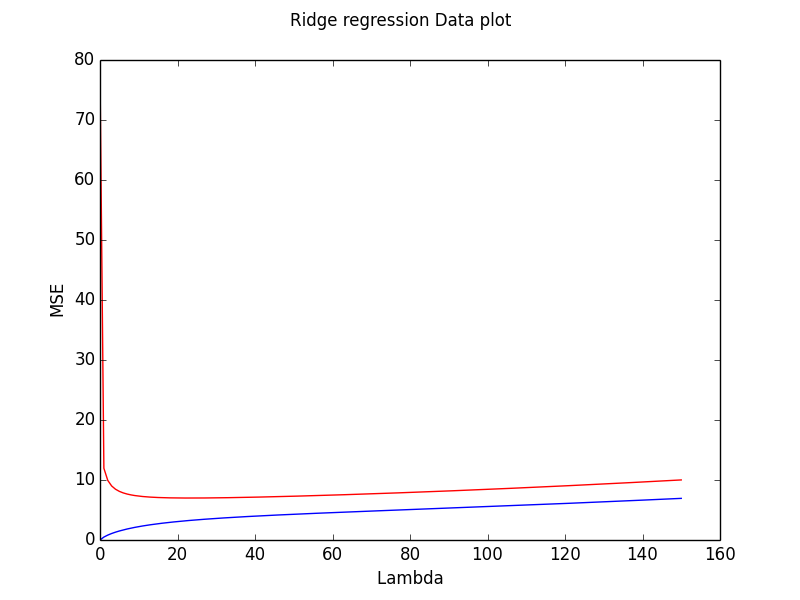
**Part 1**

**There are seven images in this part. The best value of lambda will be shown in Part 3 together.**

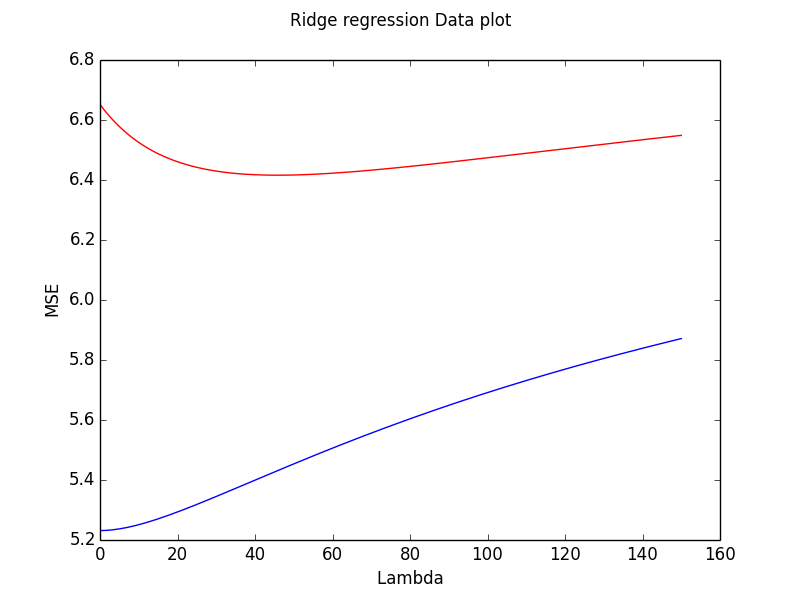
Run: python hw1\_Part1.py train-100-10.csv test-100-10.csv



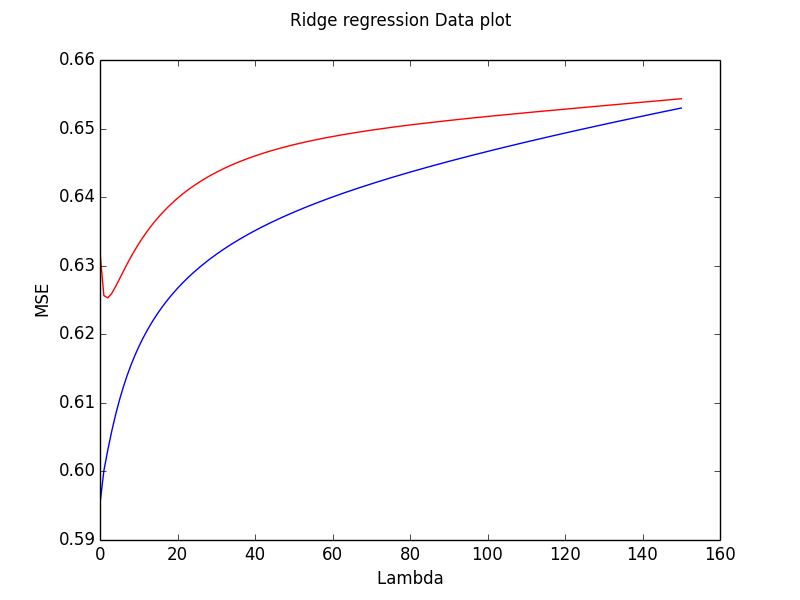
Run: python hw1\_Part1.py train-100-100.csv test-100-100.csv



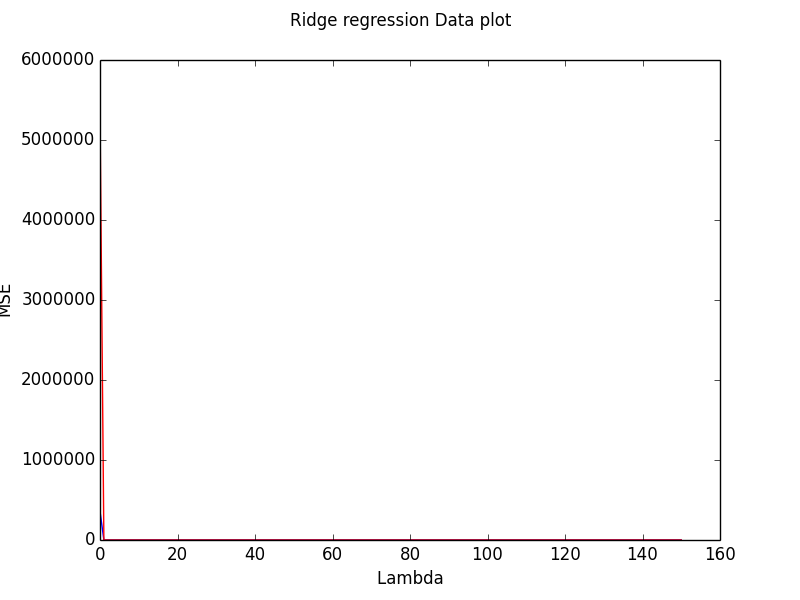
Run: python hw1\_Part1.py train-1000-100.csv test-1000-100.csv



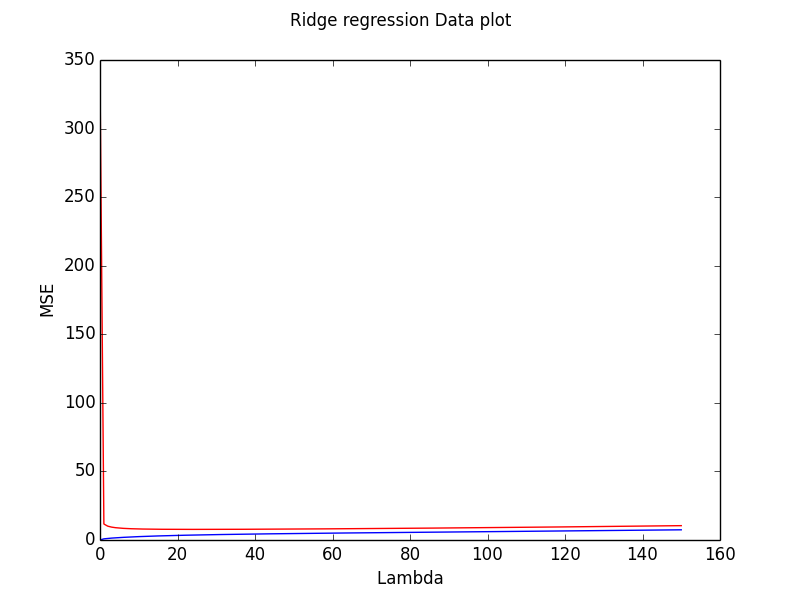
Run: python hw1\_Part1.py train-wine.csv test-wine.csv



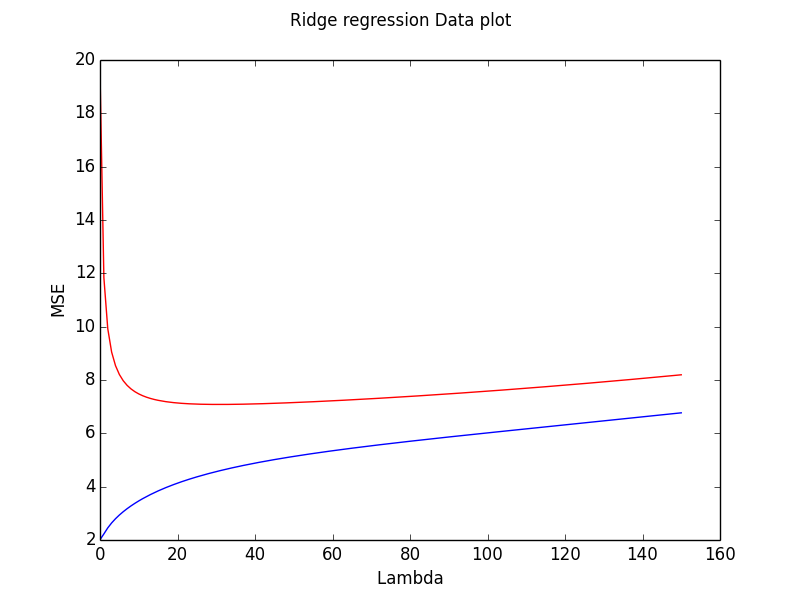
Run: python hw1\_Part1.py train-50of1000-100.csv test-1000-100.csv



Run: python hw1\_Part1.py train-100of1000-100.csv test-1000-100.csv



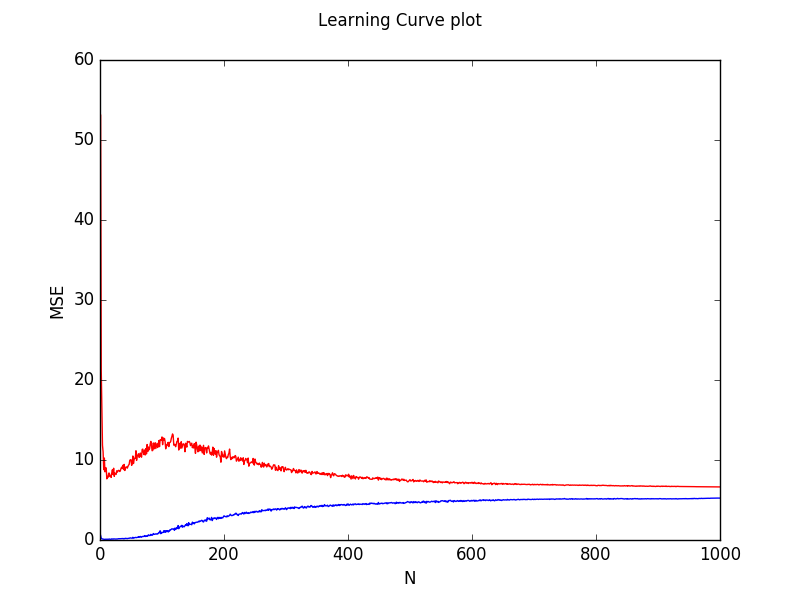
Run: python hw1\_Part1.py train-150of1000-100.csv test-1000-100.csv

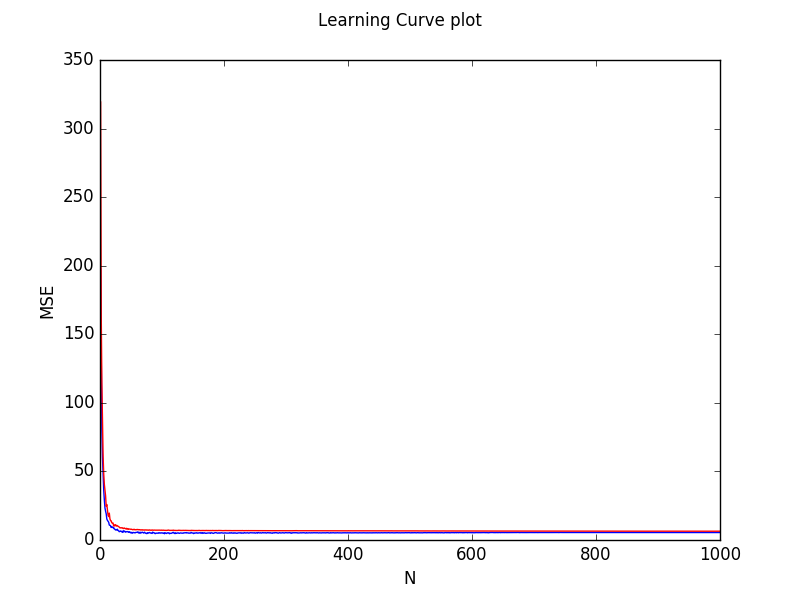


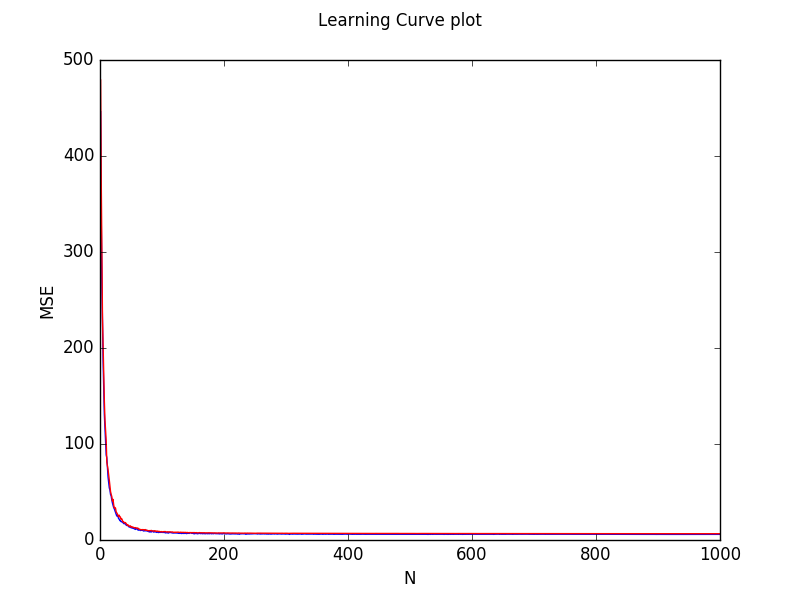
* Question: How does λ affect the MSE in general? How does the choice of λ depend on the number of features vs. examples? How does λ change with number of examples when the number of features is fixed?
* Answer: If lambda is bigger, the MSE is smaller at first and then bigger in general. We can get the best value of lambda as the MSE of test data is the smallest. The best value of lambda will be bigger if the number of features or the number of examples grows. If the number of features is fixed, the best value of lambda will be bigger with the number of examples increasing.

**Part 2**

**This part will show the learning curve for different value of lambda.**

lambda = 1 :

lambda = 46 (best value) :

lambda = 150 :

**Part 3**

**Here is a table for comparison of best value of lambda and MSE between Part 1 and Part 3.**

|  |  |  |
| --- | --- | --- |
| Data | Use CV (Part 3) | Use test data (Part 1) |
| 100-10 | 9, [ 6.39515867] | 8, [ 6.44323578] |
| 100-100 | 12, [ 5.24290646] | 23, [ 6.99424915] |
| 1000-100 | 59, [ 6.28384207] | 46, [ 6.41599941] |
| wine | 3, [ 0.64190936] | 2, [ 0.62530884] |
| 50of1000-100 | 13, [ 7.46514274] | 11, [ 8.17951264] |
| 100of1000-100 | 14, [ 5.86604203] | 25, [ 7.7409998] |
| 150of1000-100 | 37, [ 6.82466029] | 31, [ 7.08319072] |

* Question: How do the values for λ and MSE obtained from CV compare to the choice of λ and best test set MSE in question 1? What are the drawbacks of CV? What are the factors affecting the performance of CV?
* Answer: There is a little difference between the results from Part 1 and Part 3, but they are quite close to each other. If the number of examples is smaller, using CV will bring much bigger difference. The drawback of 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. The factors which affect the performance of CV are the number of examples, the number of features, the variance of data and the number k in k fold Cross Validation.