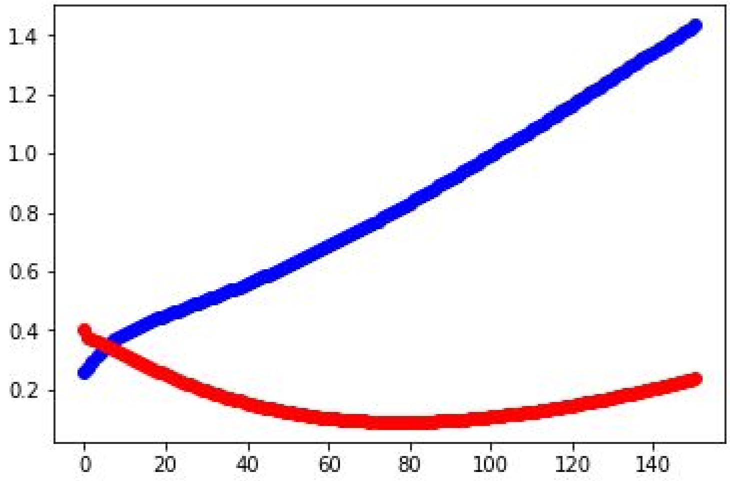
1. Implement L2 regularized linear regression algorithm with λ ranging from 0 to 150 (integers only). For each of the 6 dataset, plot both the training set MSE and the test set MSE as a function of λ (x-axis) in one graph.

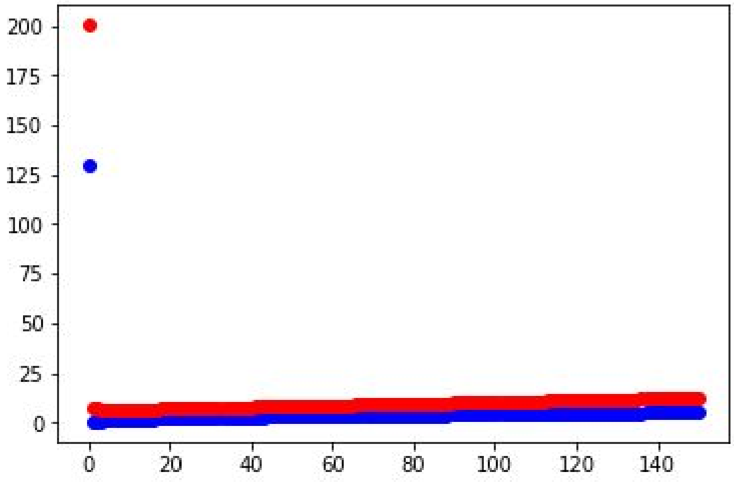
(a) For each dataset, which λ value gives the least test set MSE?

1)For train-100-10 and test-100-10



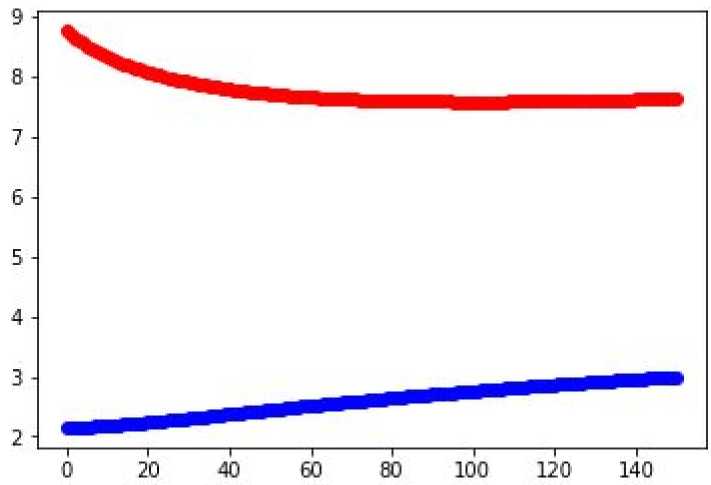
λ = 78 gives the least test set MSE.

2)For train-100-100 and test-100-100



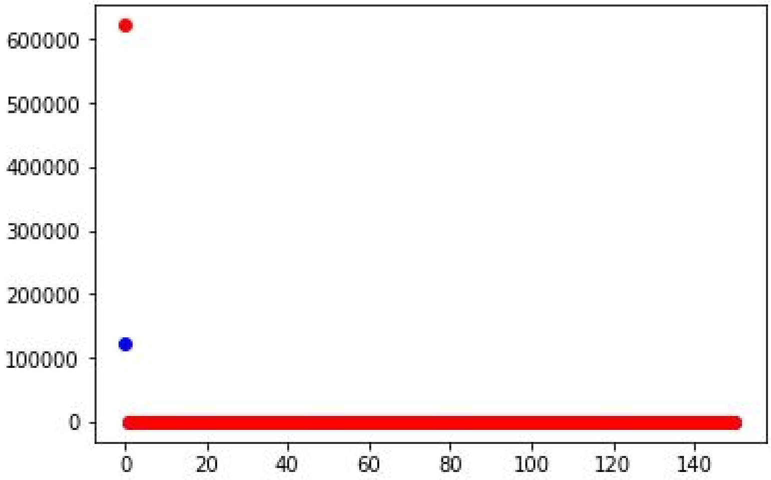
λ = 6 gives the least test set MSE.

3)For train-1000-100 and test-1000-100



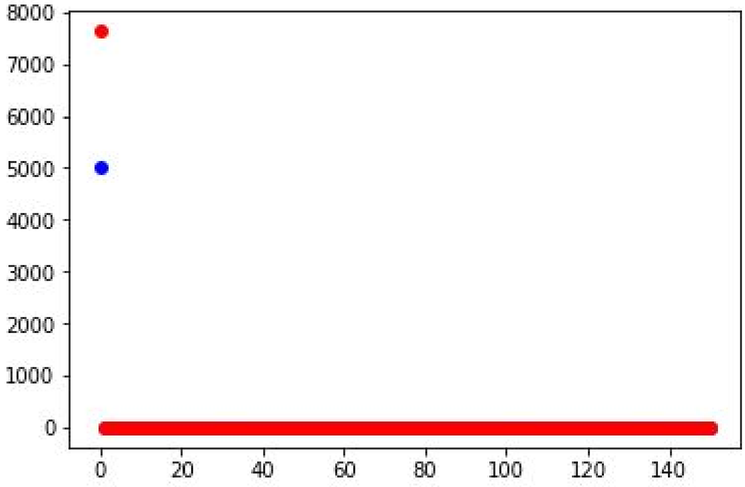
λ = 101 gives the least test set MSE.

4)For train-50(1000)-100 and test-1000-100



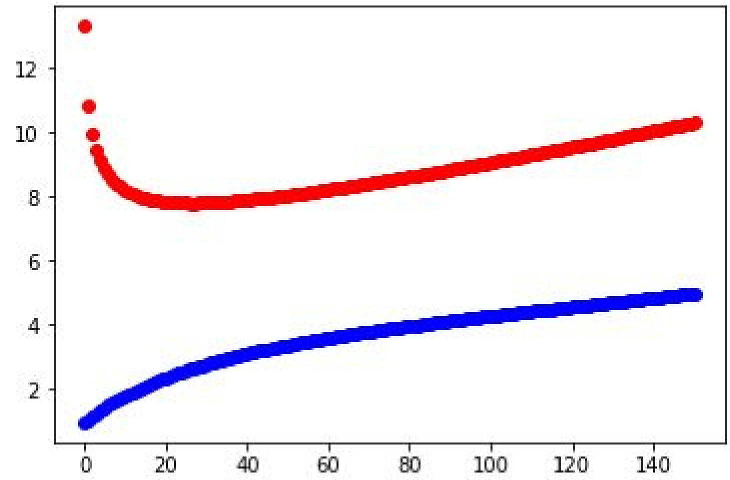
λ = 5 gives the least test set MSE.

5)For train-100(1000)-100 and test-1000-100



λ = 17 gives the least test set MSE.

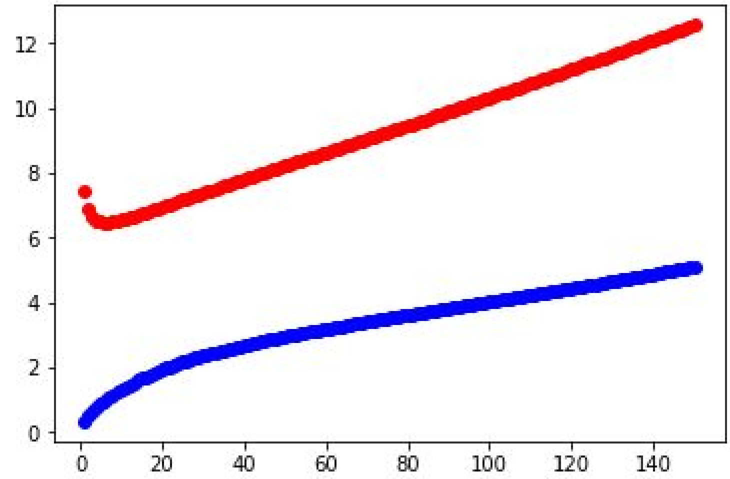
6)For train-150(1000)-100 and test-1000-100



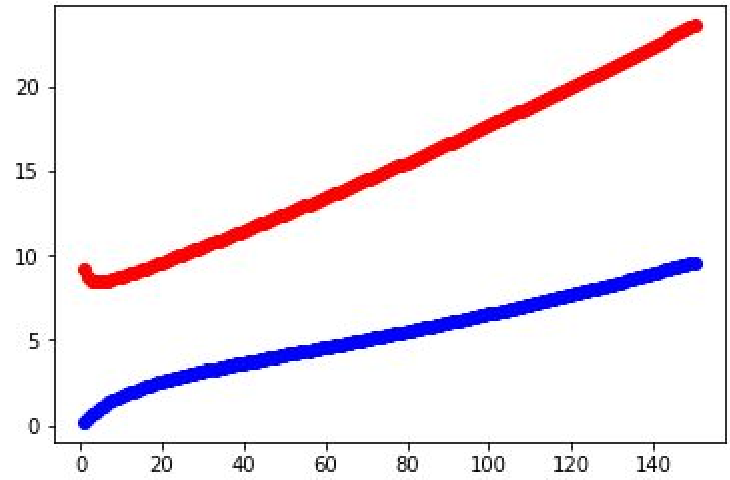
λ = 26 gives the least test set MSE.

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

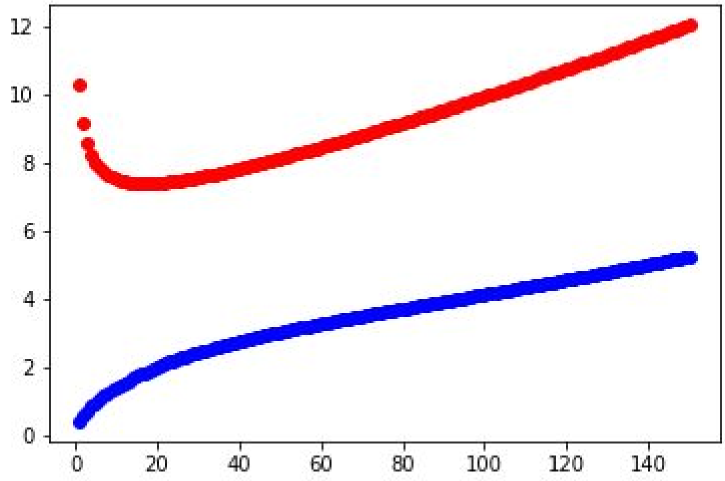
For train-100-100 and test-100-100



For train-50(1000)-100 and test-1000-100



For train-100(1000)-100 and test-1000-100



(c) Explain why λ = 0 (i.e., no regularization) gives abnormally large MSEs for those three datasets in (b).
with λ = 0, there will be no regularization, which means that all the features will be trained and fitted to the model. In our example datasets, the training data is not large enough, so the bias will not be small. In this case, the test MSE and train MSE both will be large as graph shows.

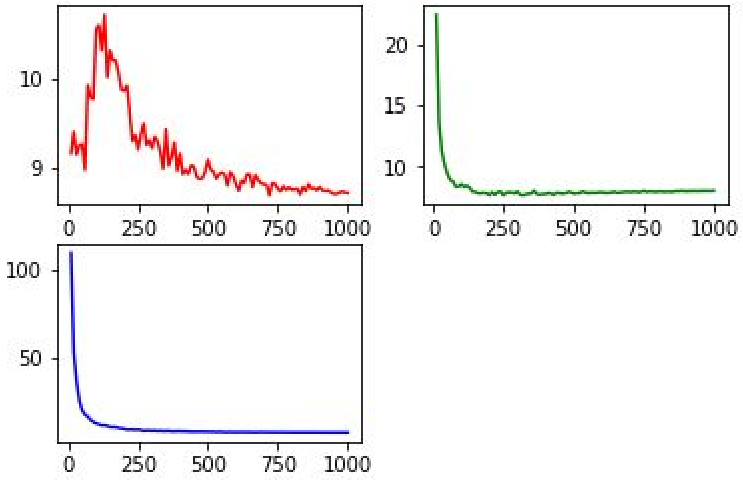
2.(a)
best choice λ and test MSE
(0, 0.40339559550015713)
(4, 6.542356531931802)
(9, 8.368462925852166)
(12, 8.900955082692358)
(10, 7.508724156327268)
(9, 8.304332627274558)

(b) λ and MSE in question 2 is totally different from λ and MSE in question 1, MSE seems to be a little big larger using CV technique.
(c)
1.takes too much time for calculating
2.because CV does not use the whole train data for training, bias is a little bit larger than LOOCV
3.need to choose a proper K

(d)

1. the choice of K
2. the size of training data

3.
the learning curve for λ= 1,25,150



10
9
0
250
500
750
1000
20
15
10
0
250
500
750
1000
100
50
0
250
500
750
1000