Question 1:

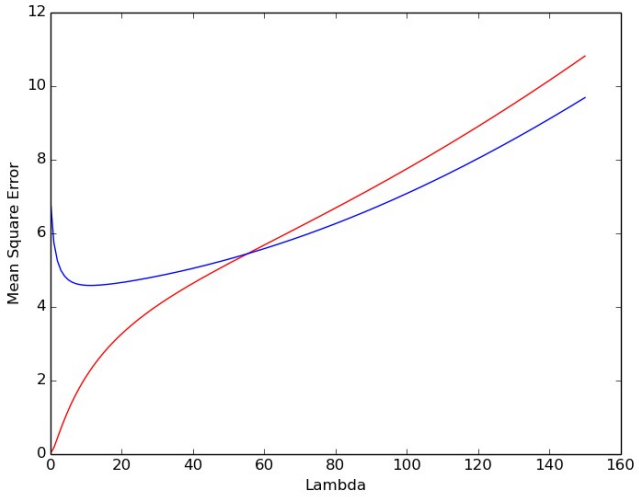

Figure-1 Head-50 vs 1000-100-test

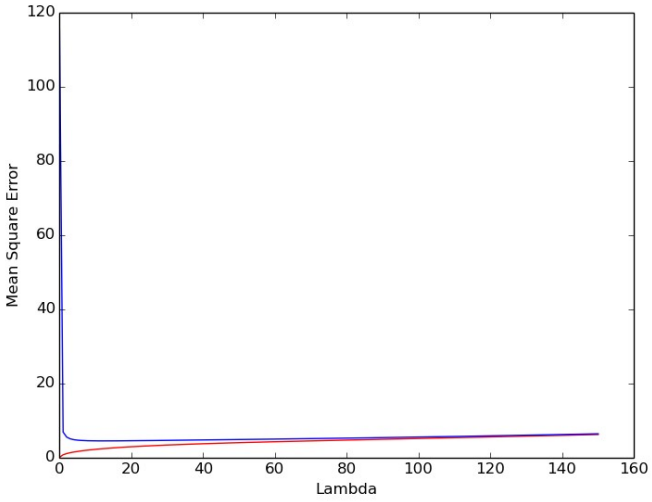

Figure-2 Head-100 vs 1000-100-test

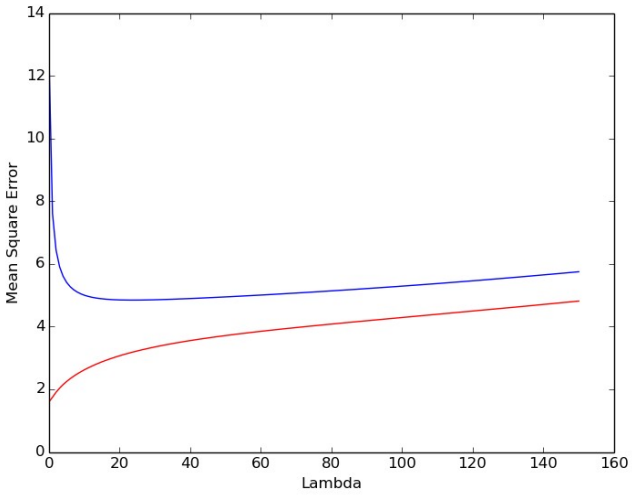

Figure-3 Head-150 vs 1000-100-test

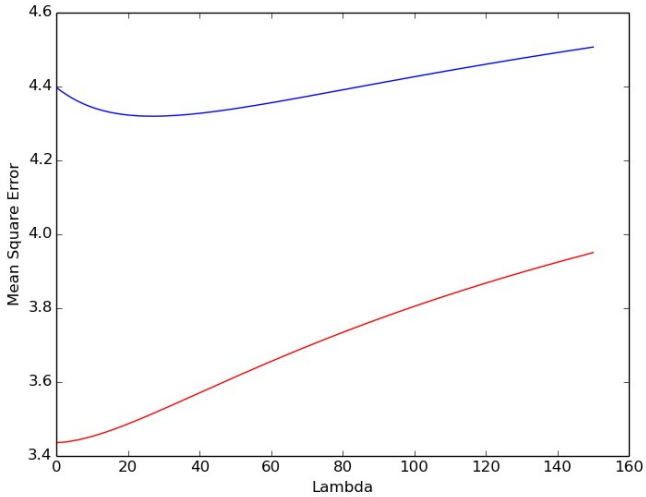

Figure-4 1000-100-train vs 1000-100-test

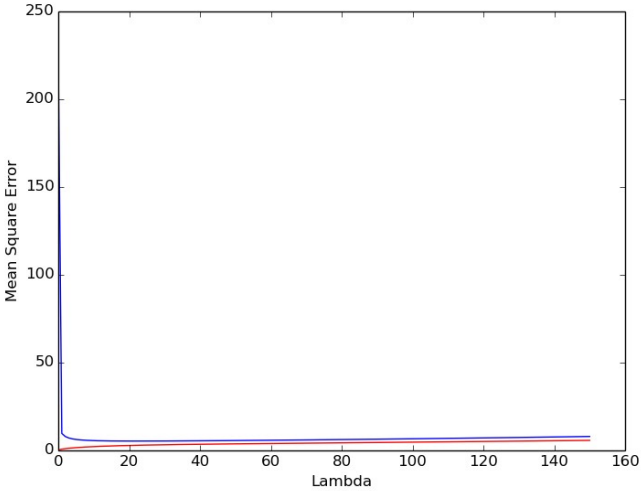

Figure-5 100-100-train vs 100-100-test

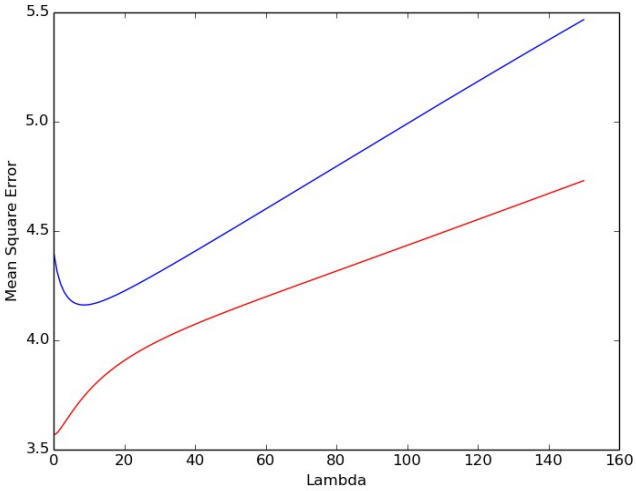

Figure-6 100-10-train vs 100-10-test

Above figures show as the value of λ increases from 0 to 150, how the training error and test error

change for each train/test pairs. The red line indicates the training error and the blue line shows the testing error.

If the above figures are unclear, you can find the graphs under the “pic/q1” directory, where the names of the photos refer to:

PNG Name	Data Sets
h50_100.png	50(1000)_100_train / 1000_100_test
h100_100.png	100(1000)_100_train / 1000_100_test
h150_100.png	150(1000)_100_train / 1000_100_test
1000_100.png	1000_100_train / 1000_100_test
100_100.png	100_100_train / 100_100_test
100_10.png	100_10_train / 100_10_test

And here is the best λ for all the train/test pairs:

Best λ for H50_100_TRAIN: 55 (50 examples, 100 features)
Test MSE: 5.42424429762

Best λ for H100_100_TRAIN: 150 (100 examples, 100 features)
Test MSE: 6.34432065511

Best λ for H150_100_TRAIN: 150 (150 examples, 100 features)
Test MSE: 5.74841020606

Best λ for 1000_100_TRAIN: 150 (1000 examples, 100 features)
Test MSE: 4.50556188087

Best λ for 100_10_TRAIN: 26 (100 examples, 10 features)
Test MSE: 4.27494222663

Best λ for 100_100_TRAIN: 66 (100 examples, 100 features)
Test MSE: 5.64751367191

Q: How does λ affect the MSE in general?
A: When λ is relatively small, the model is complex and it tends to overfit the data. In this case, the training error will be very low but the testing error will be high.

When λ keeps increasing, the training error tends to be higher and the testing error will be lower and lower.

When λ increases to the best value, the training error might be higher than the models using small λ, but the good thing is the testing error achieves the lowest point.

However, as λ keeps increasing after reaching the best value, both training and testing error will go up because the model now becomes too simple to predict the true labels.

Q: How does the choice of λ depend on the number of features vs. examples?
A: (1) When (number of features) * (number of examples) is larger, then best value of λ tends to be larger. This can been seen from the result listed in the following table:

Features * Examples	Best λ
100 * 10	26
50(head) * 100	55
100 * 100	66
100(head) * 100	150
150(head) * 100	150
1000 * 100	150

The last three datasets have the same best λ, because we only tried λ samller than 151.
And the reason for this phenomenon might be when the training dataset has more examples and more features, the trained model tends to be more complex to fit the training data. Hence there is more potentiality for λ to increase to the best value, before the training error and testing error both go up.

(2) When the number of examples is fixed, the best value of λ increases as the number of features go up. The following table shows the best λ for datasts containing 100 examples:

Number of Features	Best λ
10	26
100	66
100 (100 head of 1000 records)	150

Q: How does λ change with number of examples when the number of features is fixed?
A: Here is a table showing the list of best λ for training set with **100 features**:

Number of Examples	Best λ
50 (head)	55
100 (head)	150
100	66
150 (head)	150
1000	150

As the number of examples goes up, the value of best λ goes up as well.

Question 2:

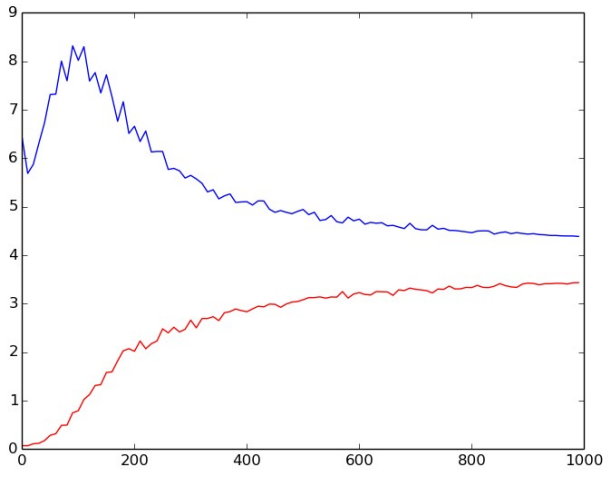

Figure-7 λ = 1

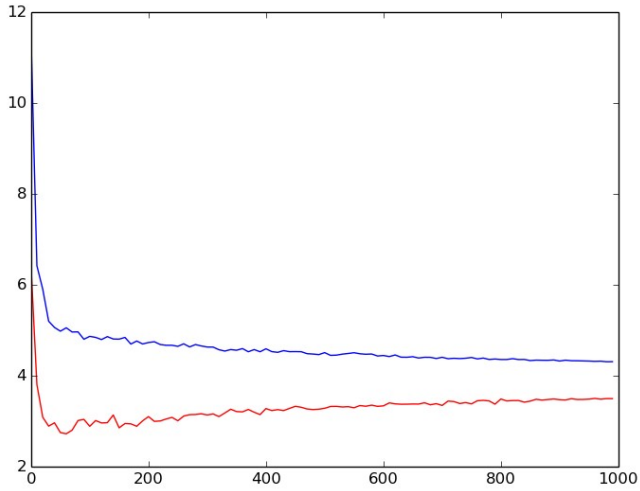

Figure-8 λ = 25

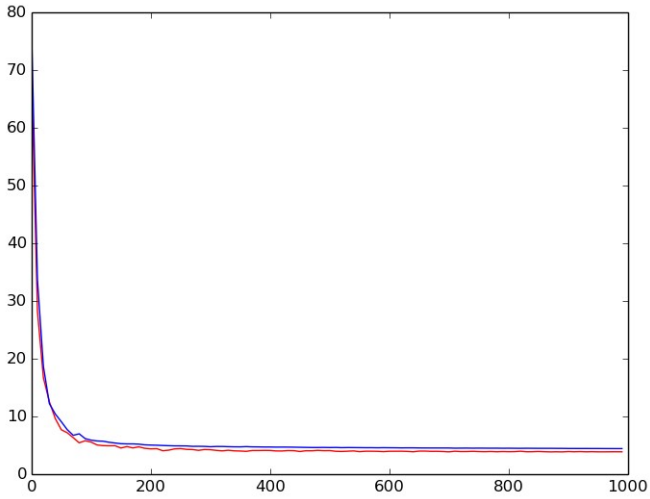

Figure-9 λ = 150

The three graphs can also be found in “pic/q2”. They all represent how the training/testing error change as the size of training set goes up. The red line displays the trend of training error and the blue line shows the testing error.

Question 3:

Output of q3.py:

Best lamda for H50_100_TRAIN: 0 (head 50 examples, 100 features)
1000_100_TEST MSE for this lamda: 6.94904858122

Best lamda for H100_100_TRAIN: 18 (head 100 examples, 100 features)
1000_100_TEST MSE for this lamda: 4.45023214048

Best lamda for H150_100_TRAIN: 47 (head 150 examples, 100 features)
1000_100_TEST MSE for this lamda: 4.92900317377

Best lamda for 1000_100_TRAIN: 39 (1000 examples, 100 features)
1000_100_TEST MSE for this lamda: 4.32518286252

Best lamda for 100_10_TRAIN: 13 (100 examples, 10 features)
100_10_TEST MSE for this lamda: 4.17350735396

Best lamda for 100_100_TRAIN: 20 (100 examples, 100 features)
100_100_TEST MSE for this lamda: 5.07675140801

Q: How do the values for λ and MSE obtained from CV compare to the ones in question 1?
A: In question 1, the values of λ is obtained by calculating the minimum distance of training mean square error and testing mean square error. But in question 3, there is no testing set when computing the best λ. For each different training dataset, we do cross-validation over it. We try different values of λ and calculate the average “test fold error”. The λ which introduces lowest “test fold mean square error” will be considered as the best choice.

Q: What are the drawbacks of CV?
A: (1) Although for each fold, the training data set is different but actually they are overlapped.
(2) It is sometimes difficult to decide how many folds you want to split.

Q: What are the factors affecting the performance of CV?
A: (1) The choice of number of folds. The larger the number, more loops you are going to run.
(2) If you are using cross-validation to calculate the best parameter for your model – like what we did in question 3, then the outer loop will be determined by the range of the parameter you will consider. The larger the range, the more loops you are going to run.