CSE574 Introduction to Machine Learning

Programming Assignment 1

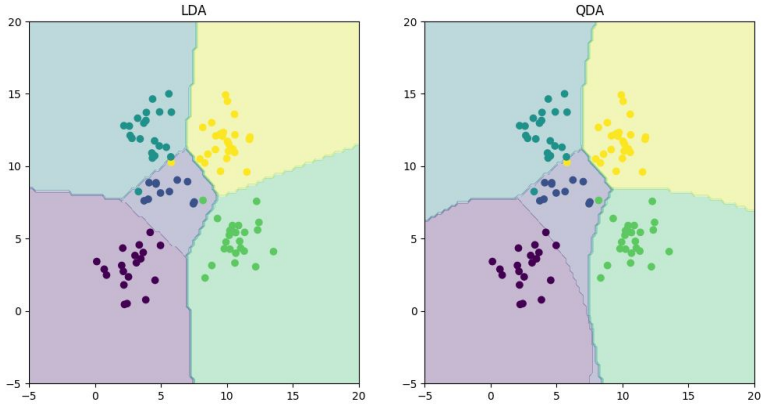
Classication and Regression

yuxiang liu <yliu268@buffalo.edu>
shaoming xu <shaoming@buffalo.edu>

REPORT 1

Train both methods using the sample training data (sample train). Report the accuracy of LDA and QDA on the provided test data set (sample test). Also, plot the discriminating boundary for linear and quadratic discriminators. The code to plot the boundaries is already provided in the base code. Explain why there is a difference in the two boundaries.

1. LDA Accuracy = 0.97
2. QDA Accuracy = 0.96
3. Discriminating boundary for linear and quadratic discriminators:



4. QDA computes non-diagonal covariance matrix for each class in qdalearn() and uses the covariance matrix of the specific class to compute the its MLE in qdaTest(). But LDA only computes a non-diagonal covariance matrix of all training samples in ldalearn() and uses it to compute all class MLE in ldaTest(). The fact on using different covariance matrixs between QDA and LDA is the reason why there is a difference in the two boundaries.

REPORT 2.

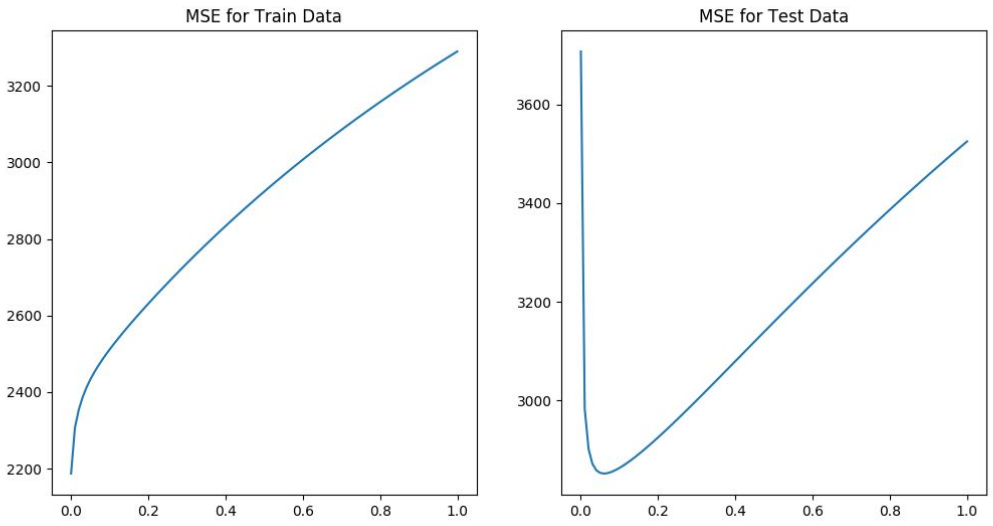
Calculate and report the MSE for training and test data for two cases: first, without using an intercept (or bias) term, and second with using an intercept. Which one is better?

1. For Test data:
    a. MSE without intercept 106775.36156138896
    b. MSE with intercept 3707.8401816746155
    c. MSE with intercept is better
2. For Training data
    a. MSE without intercept 19099.44684457091
    b. MSE with intercept 2187.1602949303865
    c. MSE with intercept is better

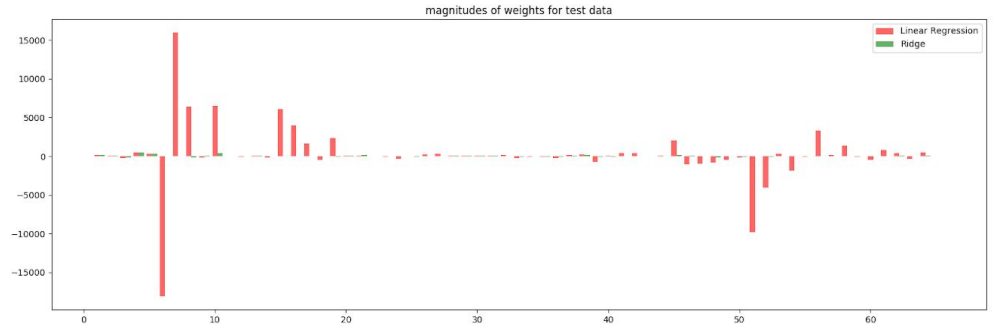
REPORT 3.

Calculate and report the MSE for training and test data using ridge regression parameters and using the testOLERegression function that you implemented in Problem 2. Use data with intercept. Plot the errors on train and test data for different values of lambda. Vary lambd from 0 (no regularization) to 1 in steps of 0.01. Compare the relative magnitudes of weights learnt using OLE (Problem 2) and weights learnt using ridge regression. Compare the two approaches in terms of errors on train and test data. What is the optimal value for lambda and why?

1. Plot of the errors on train and test data for different values of lambda.



2. Compare the relative magnitudes of weights learnt using OLE (Problem 2) and weights learnt using ridge regression
   From the plot we can see ridge regression tends to minimize magnitudes of weights.



3. Compare the two approaches in terms of errors on train and test data. What is the optimal value for lambda and why?
   a. For Test data
      i. Linear regression
         1. MSE with intercept: 3707.8401816746155
      ii. Ridge regression
         1. MSE with intercept: 2851.33021344
         2. optimal value of lambd: 0.06

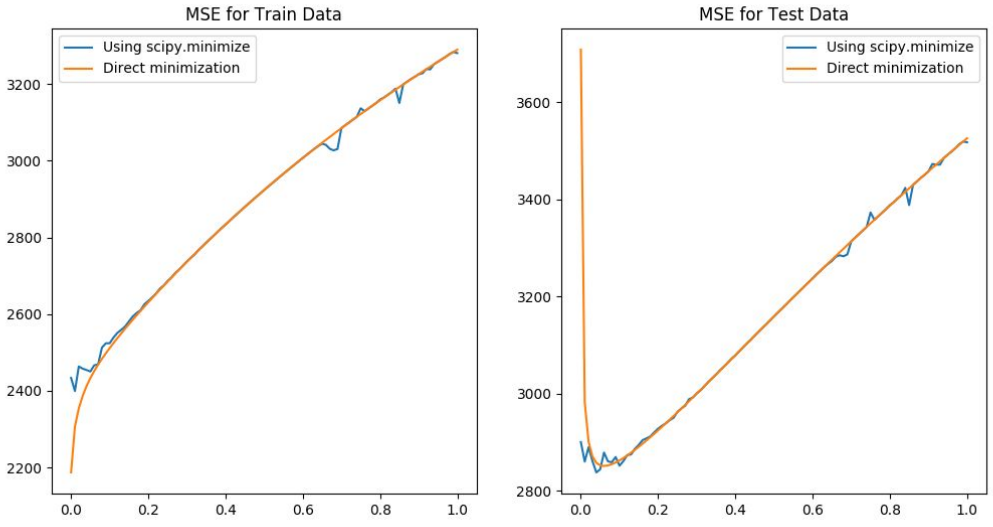
b. For Train data
    i. Linear regression
        1. MSE with intercept 2187.1602949303865
    ii. Ridge regression
        1. MSE with intercept: 2187.16029493
        2. optimal value of lambda: 0.0

When lambda is small, the weights learnt is overfitting. So although we get a low MSE for trainning data, we will have a high MSE, (mean worse prediction) for test data. So when we increase the lambda to minimize overfitting, the weights we learnt will perform worse on training data, but will give us a better result on test data. However when the lambda is too large, the penalty will have a larger weight than actual model, so the weight learnt is underfit. Both training data and test data will give a high MSE. So the lambda gives lowest MSE for test data is the best lambda, because it reduce overfitting without causing too much underfitting.

REPORT 4.

Plot the errors on train and test data obtained by using the gradient descent based learning by varying the regularization parameter . Compare with the results obtained in Problem 3.

1. Plot



2. Compare with the results obtained in Problem 3
    a. We can see the shape of plot is nearly identical for Problem 4 and 3.
    b. For Test data
        i. Problem3
            1. MSE with intercept: 2851.33021344
            2. optimal value of lambd: 0.06
        ii. Problem4
            1. MSE with intercept: 2832.81744288
            2. optimal value of lambd: 0.03
    c. For train data
        i. Problem3
            1. MSE with intercept: 2187.16029493
            2. optimal value of lambda: 0.0

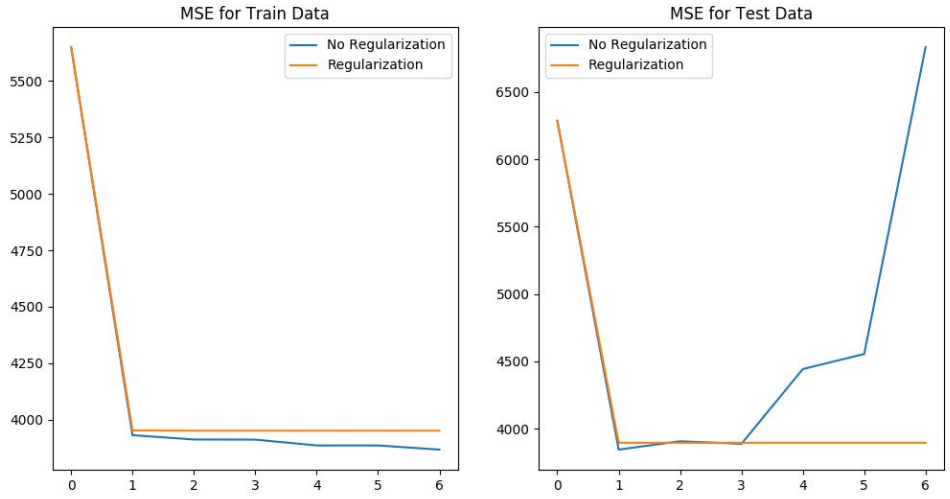
ii. Problem4
    1. MSE with intercept: 2393.28057997
    2. optimal value of lambda: 0.01

REPORT 5.

Using the lambda = 0 and the optimal value of lambda found in Problem 3, train ridge regression weights using the non-linear mapping of the data. Vary p from 0 to 6. Note that p = 0 means using a horizontal line as the regression line, p = 1 is the same as linear ridge regression. Compute the errors on train and test data. Compare the results for both values of lambda. What is the optimal value of p in terms of test error in each setting? Plot the curve for the optimal value of p for both values of lambda and compare.

1. Compute the errors on train and test data.
    a. train data
        i. lambda = 0
        [5650.71 3930.92 3911.84 3911.19 3885.47 3885.41 3866.88]
        ii. lambda = 0.06
        [5650.71 3951.84 3950.69 3950.68 3950.68 3950.68 3950.68]]
    b. test data
        i. lambda = 0
        [6286.4 3845.03 3907.13 3887.98 4443.33 4554.83 6833.46]
        ii. lambda = 0.06
        [6286.88 3895.86 3895.58 3895.58 3895.58 3895.58 3895.58]

2. Compare the results for both values of lambda.
    a. In training data the MSE of lambda = 0 is decreasing more rapidly than the MSE of the lambda = 0.06. This phenomenon is caused by overfitting of training data when lambda = 0.
    b. We can see the Test data. In the lambda = 0 case , the MSE starts oscillating when p in [1,3) and increasing rapidly after p in [3,6]. Overfitting happens. But In the lambda = 0.06 case, the MSE decreases rapidly when p in [0,1) and keeps stable after p in [3,6]. So the Overfitting is avoided.
3. What is the optimal value of p in terms of test error in each setting?
    a. For the training data
        i. For No Regularization case the p is 6
        ii. For Regularization case the p is 6
    b. For the test data
        i. For No Regularization case the p is 1
        ii. For Regularization case the p is 4
4. Plot the curve for the optimal value of p for both values of and compare.
    a. In training data the MSE of No Regularization case is decreasing more rapidly than the MSE of the Regularization case. This phenomenon is caused by overfitting of training data in the No Regularization case.
    b. We can see the Test data. In the No Regularization case, the MSE starts oscillating when p in [1,3) and increasing rapidly after p in [3,6]. Overfitting happens. But In the Regularization case, the MSE decreases rapidly when p in [0,1) and keeps stable after p in [3,6]. So the Overfitting is avoided.



Problem 6: Interpreting Results

Using the results obtained for previous 4 problems, make final recommendations for anyone using regression for predicting diabetes level using the input features.

REPORT 6.

Compare the various approaches in terms of training and testing error. What methods should be used to choose the best setting?

Here are several rules we can keep in mind:

1. Adding intercept is better than not.
2. Regularization is better than No Regularization.
3. Lambda should be chose carefully.
4. Ridge regression is better than Linear regression.
5. The fact your model fits training data too well might bite you.
6. Worried about singular matrix or performance for large data set? Using Gradiant descent!
7. Order of Non-linear Regression should be chose carefully.
8. Increasing order avoids underfitting. Decreasing order or regularization avoid overfitting.