CSE574 Introduction to Machine Learning

Programming Assignment 3

Classification and Regression

Team members:

Sahithya Arveti Nagaraju (50559752 – sarvetin@buffalo.edu)

Sushmitha Manjunatha (50560530 – smanjuna@buffalo.edu)

Problem 1: Experiment with Gaussian Discriminators

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 between the two boundaries.

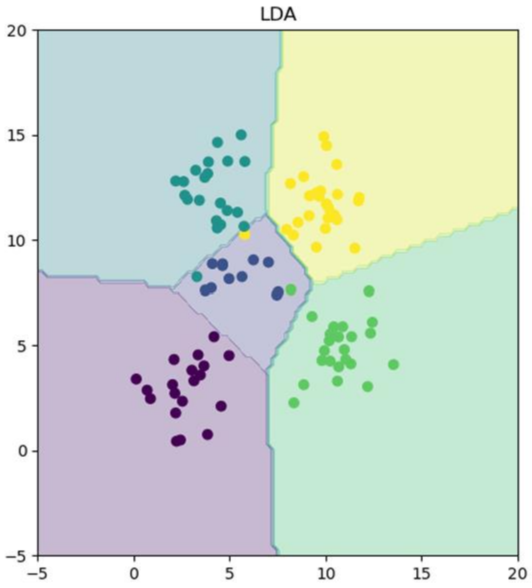
- **LDA** and **QDA** implemented with functions:
  - ldaLearn and qdaLearn to estimate means and covariance.
  - ldaTest and qdaTest for predictions and accuracy.

Results:

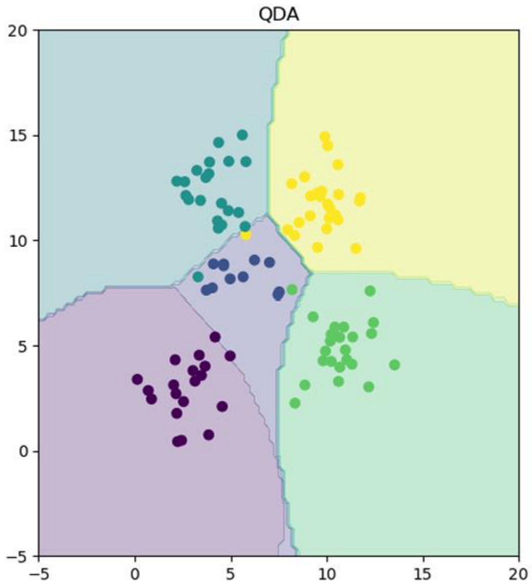
- LDA accuracy: 97.0%
- QDA accuracy: 96.0%

Boundary Plots:

- Linear decision boundary for LDA.



- Quadratic decision boundary for QDA.



The separation boundary between LDA and QDA is due to their treatment of covariance:

- LDA assumes equal covariance matrices across different classes, which creates a linear decision boundary. LDA, in effect, separates classes using only the mean of the classes.
- QDA lets each class have different covariance matrices, hence giving a quadratic boundary. This will allow various shapes and orientations of class distributions.

That being said, LDA works well with an equal covariance assumption while QDA is apt for cases where class covariances are different, hence the different boundary shapes.

Problem 2: Experiment with Linear Regression

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?

MSE without intercept on test data 106775.36155764468

MSE with intercept on test data 3707.8401815786037

MSE without intercept on train data19099.446844570528

MSE with intercept on train data 2187.160294930391

MSE Type	Train Data	Test Data
Without Intercept	19099.446844570528	106775.36155764468
With Intercept	2187.160294930391	3707.8401815786037

The model with an intercept is better because it achieves significantly lower mean squared error (MSE) on both the training and test datasets. The model with an intercept can potentially account for a baseline value of the data, hence resulting in much better predictions. In the model without an intercept, it is assumed that the target should be zero when all features are zero. That gives it biased predictions and higher MSE. Adding the intercept helps to better fit and generalize the model on data. Therefore, this is the better choice.

Problem 3: Experiment with Ridge Regression

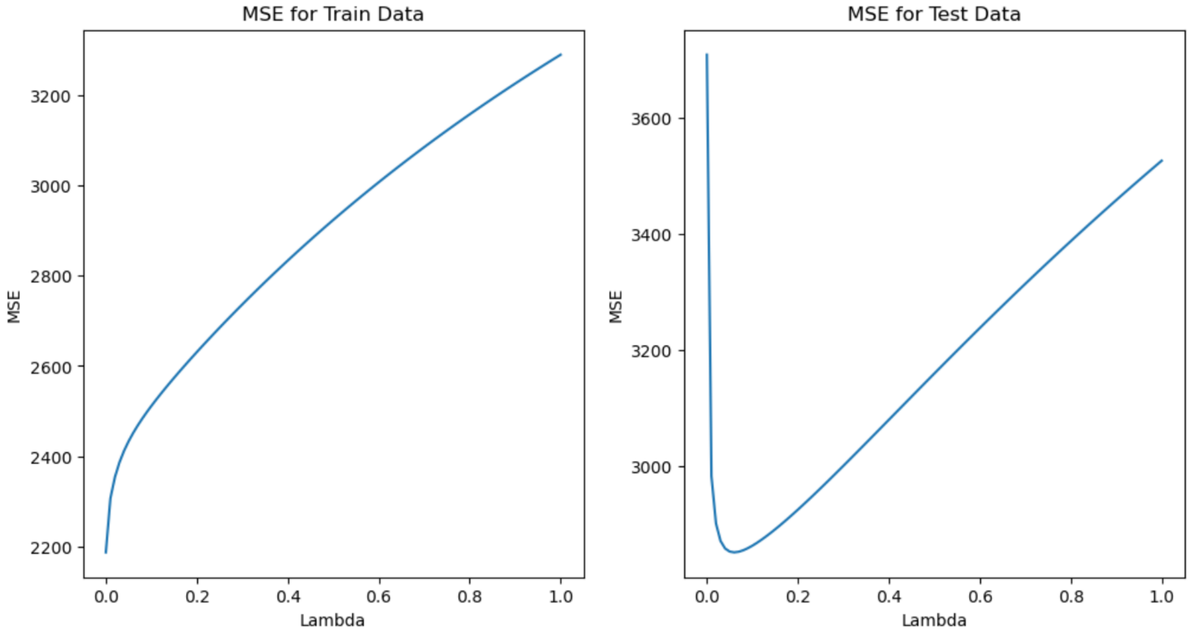
Calculate and report the MSE for training and test data using ridge regression parameters using the the testOLERegression function that you implemented in Problem 2. Use data with intercept. Plot the errors on train and test data for different values of λ. Vary λ from 0 (no regularization) to 1 in steps 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 λ and why?

Below is the report of the MSE for training and test data using Ridge regression parameters using testOLERegression function from Problem 2 by varying λ from 0 to 1 in steps of 0.01,

lambda	Train Data	Test Data
0.0	2187.1603	3707.8402
0.01	2306.8322	2982.4461

0.02	2354.0713	2900.9736
0.03	2386.7802	2870.9416
0.04	2412.1190	2858.0004
0.05	2433.1744	2852.6657
0.06	2451.5285	2851.3302
0.07	2468.0776	2852.3500
0.08	2483.3656	2854.8797
0.09	2497.7403	2858.4444
0.10	2511.4323	2862.7579
0.20	2630.8728	2924.7532
40.30	2736.4726	3000.1158
0.40	2833.6641	3079.3852
0.50	2923.6301	3159.0140
0.75	3121.8865	3350.4239
1.00	3289.7613	3525.3946

Plot of the errors on train and test data for different values of **λ,**



Observations:

- As λ increases beyond 0.06, the MSE on the test data increases, which indicates that larger regularization is reducing the model's ability to fit the data well.
- On the other hand, with λ=0, we see a very low training MSE (2187.1603) but a higher test MSE (3707.8402), suggesting overfitting without regularization.

Therefore, **the optimal value of λ is 0.06**, as it minimizes the MSE on the test data (2851.3302), providing the best generalization. Although λ=0 results in the lowest training error (2187.1603), it leads to overfitting with a higher test error (3707.8402). Therefore, λ=0.06 strikes the best balance between training and test errors.

Comparison of OLE and Ridge Weights:

1. OLE Regression Weights:

- **L2 Norm**: 1.5508×10^{10}.
- The OLE weights are large, indicating potential overfitting. OLE doesn't have regularization, so the weights can grow excessively, especially in noisy or multicollinear data.

2. Ridge Regression Weights:

- **L2 Norm**: 185010.691292986
- Ridge regression applies regularization, which shrinks the weights, reducing overfitting. The smaller L2 norm shows that the weights are more controlled compared to OLE.

Comparison of OLE and Ridge Errors:

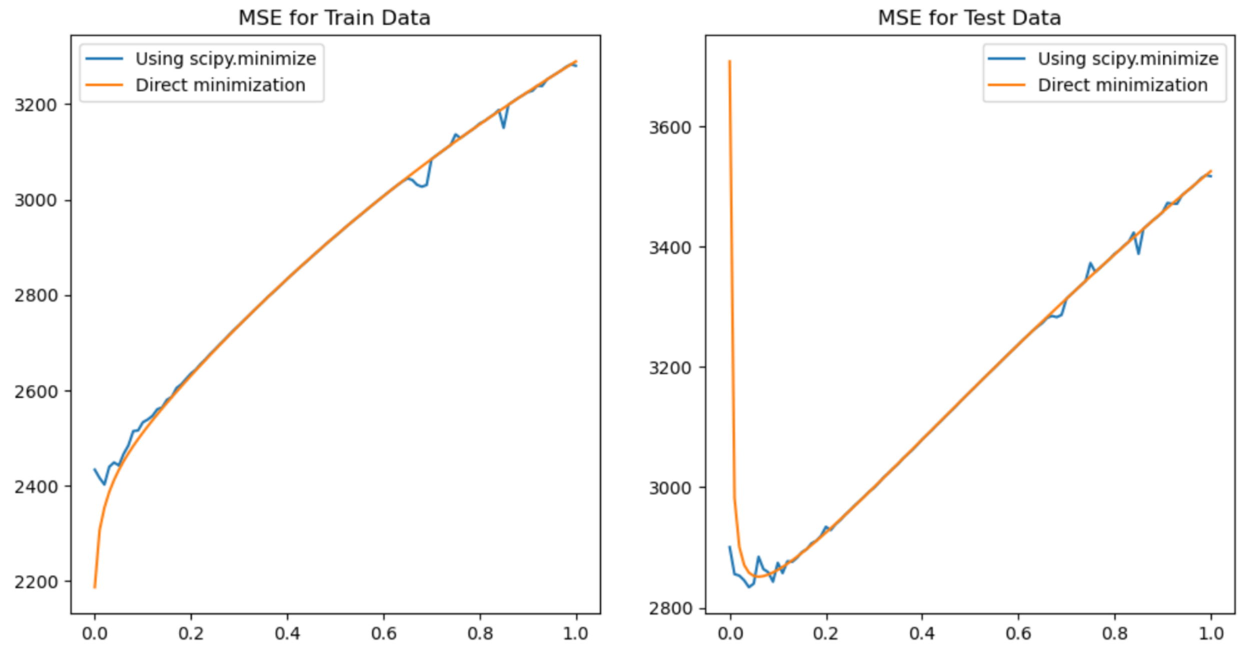
- **Training Data:** Ridge regression achieves a lower MSE (2451.5285) compared to OLE regression, which has a much higher MSE (2187.1603).
- **Test Data:** Ridge regression also performs better on the test data, with MSE = 2851.3302 for λ=0.06, compared to OLE's MSE = 3707.8401.

Based on the results above, it is clear that the testing error is lower when using ridge regression. Therefore, ridge regression is a more effective method than OLE for the current problem.

Problem 4: Using Gradient Descent for Ridge Regression Learning

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.

Plot of errors on train and test data obtained using gradient descent-based learning by varying regularization parameter lambda,



```
Gradient Descent MSE for Train Data (optimal lambda): [2448.82704127]
Gradient Descent MSE for Test Data (optimal lambda): [2833.85711219]
```

Below is the report of the MSE for training and test data using Ridge regression and using gradient descent,

	Ridge Regression		**Gradient descent for Ridge Regression**	
lambda	**Train Data**	**Test Data**	**Train Data**	**Test Data**
0.0	2187.1603	3707.8402	2433.664	2900.546
0.01	2306.8322	2982.4461	2416.256	2855.648
0.02	2354.0713	2900.9736	2402.464	2853.176
0.03	2386.7802	2870.9416	2439.536	2845.414
0.04	2412.1190	2858.0004	2448.827	2833.857
0.05	2433.1744	2852.6657	2442.534	2839.897
0.06	2451.5285	2851.3302	2466.877	2884.648
0.07	2468.0776	2852.3500	2484.732	2864.05
0.08	2483.3656	2854.8797	2514.581	2858.907
0.09	2497.7403	2858.4444	2515.926	2843.039
0.10	2511.4323	2862.7579	2433.664	2900.546
0.20	2630.8728	2924.7532	2635.398	2934.381
0.30	2736.4726	3000.1158	2736.836	2999.128
0.40	2833.6641	3079.3852	2833.638	3079.432
0.50	2923.6301	3159.0140	2923.2	3158.909
0.75	3121.8865	3350.4239	3114.584	3343.026

1.00	3289.7613	3525.3946	3280.783	3517.192

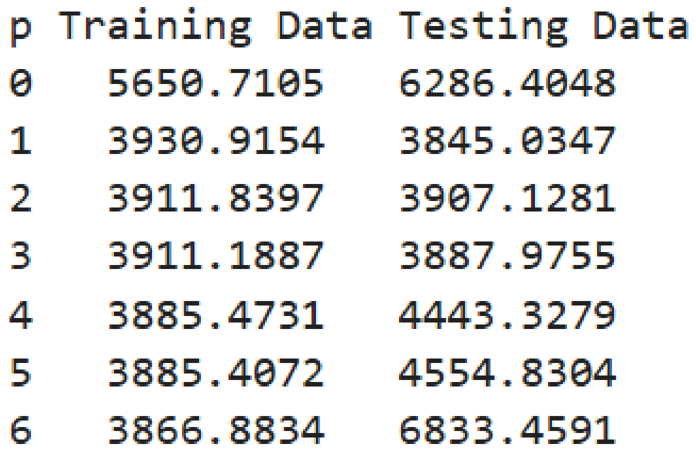
Observations:

For both Ridge regression and Gradient Descent for Ridge regression, similar error patterns are observed. For Ridge regression, the optimal value of λ is 0.06. For this value, the minimum MSE on the training data is 2187.16, and on the test data, it is 2851.33. By comparison, with Gradient Descent, the minimum training MSE is 2402.464 at λ = 0.02, while the minimum test MSE is 2833.857 at λ = 0.04. This implies that, in terms of both training and test data error, Ridge regression using gradient descent with λ=0.04 slightly outperforms Ridge Regression, while the errors are developing similarly for both methods in general.

Problem 5: Non-linear Regression

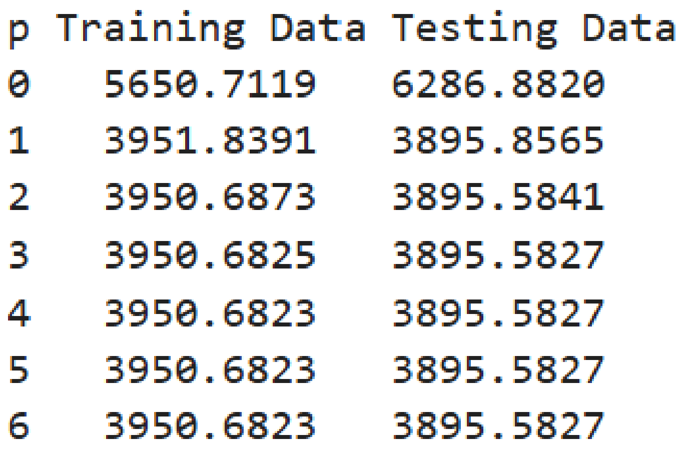
Using the λ = 0 and the optimal value of λ found in Problem 3, train ridge regression weights using the non-linear mapping of the data. Vary p from 0 to 6. Note that p = 0 means using a horizontal line as the regression line, p = 1 is the same as linear ridge regression. Compute the errors on train and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

MSE values when λ = 0,



```
p Training Data Testing Data
0   5650.7105   6286.4048
1   3930.9154   3845.0347
2   3911.8397   3907.1281
3   3911.1887   3887.9755
4   3885.4731   4443.3279
5   3885.4072   4554.8304
6   3866.8834   6833.4591
```

MSE values when λ = 0.06,

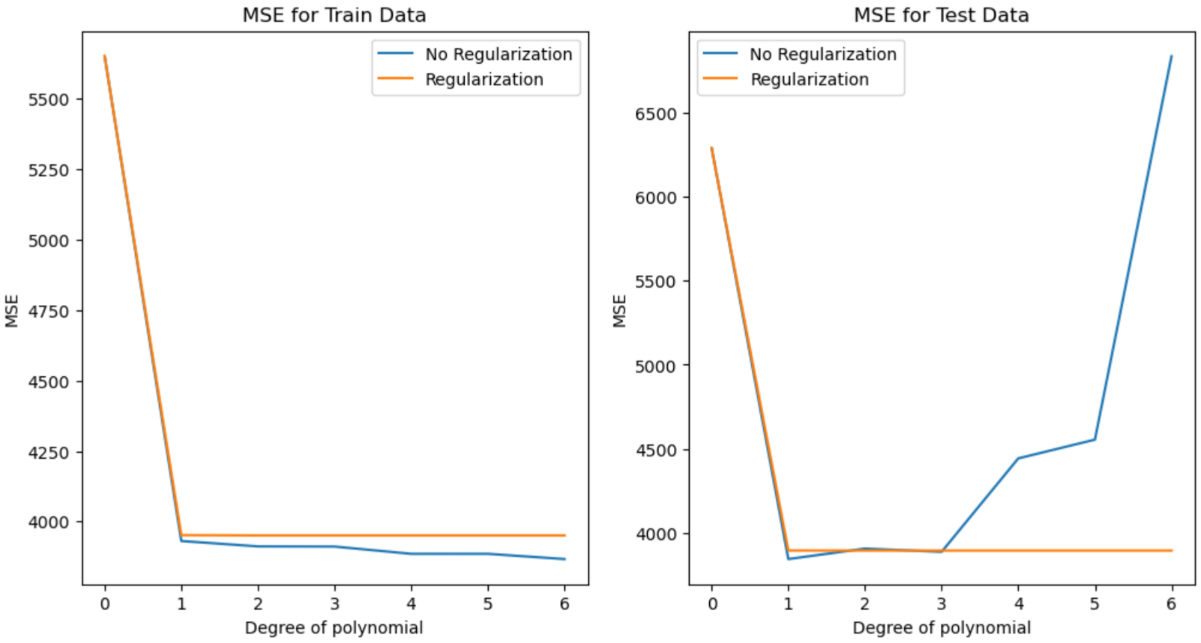


```
p Training Data Testing Data
0   5650.7119   6286.8820
1   3951.8391   3895.8565
2   3950.6873   3895.5841
3   3950.6825   3895.5827
4   3950.6823   3895.5827
5   3950.6823   3895.5827
6   3950.6823   3895.5827
```

Key Observations:

1. **λ = 0 (No Regularization)**:
   - Training data: MSE decreases significantly for p=1 and stabilizes as p increases further.
   - Testing data: MSE follows a similar trend but increases sharply for higher p values (especially p>=4).
2. **λ = 0.06 (With Regularization)**:
   - Training data: MSE slightly increases compared to λ = 0, but it remains stable across all values of p.
   - Testing data: MSE is very stable and low across all p values. No increase for higher p.

Plot of the MSE for **λ** = 0 and **λ** =0.06 respectively by varying p,



Optimal Value of p:

- For **λ = 0**, the testing error is minimal at p=1 (MSE ≈ 3845.0347).
- For **λ = 0.06**, the testing error remains almost constant regardless of p, with the lowest MSE at p=4 (MSE ≈ 3895.5827).

Problem 6: Interpreting Results

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

Table to show the MSE using different approaches,

Approach	MSE on Train Data	MSE on Test Data
Linear Regression without intercept	19099.4468	106775.3615
Liner Regression with intercept	2187.1602	3707.8402
Ridge Regression (λ = 0.06)	2451.5285	2851.3302
Ridge Regression (Gradient descent λ = 0.04)	2448.827	2833.857
Non-linear Regression with λ = 0.0 (p=1)	3930.9154	3845.0347
Non-linear Regression with λ = 0.06 (p=4)	3950.6825	3895.5827

Observations:

- Linear Regression without intercept has the poorest training and testing error, thus being a bad fit for the data.
- Linear Regression with intercept significantly improves, bringing the training and testing errors lower. The latter is still much higher on the testing dataset than in Ridge Regression.
- Both Ridge Regression models, for both λ = 0.06 and λ = 0.04, had lower test errors than Linear Regression, while having slightly higher train errors due to regularization. Both models generalize better when it comes to unseen data.
- Non-linear Regression has higher testing errors compared to Ridge Regression; this could indicate that these settings are not good for the current dataset.

Best Setting:

Mean Squared Error (MSE) on the test data should be the primary metric for selecting the best model, as it reflects how well the model generalizes to unseen data.

Based on the lowest MSE on the test data, the Ridge Regression using Gradient Descent with λ = 0.04 provides the best performance (MSE: 2833.857), closely followed by Ridge Regression with λ = 0.06 (MSE: 2851.3302).