UNIVERSITY AT BUFFALO

Project Report

On

<u>CSE 574 (Machine Learning)</u>

<u>Assignment-1</u>

By

Group-86

Members:
Arshabh Semwal
Person No. 50419031

Utkarsh Bansal
Person No. 50415035

Diwanshu Chouragade
Person No. 50419387

Abstract

This is the report of the programming assignment 1 in CSE 474/574 Introduction to Machine Learning offered in Fall 2021 under Dr. Mingchen Gao at University at Buffalo. In this report we will look at various learning techniques such as Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Linear Regression, Ridge Regression, and non-linear Ridge Regression for classification and regression problems. We will then compare these techniques on the given datasets.

Introduction

We are given 2 datasets, sample.pickle and diabetes.pickle. More description about these datasets will be provided later. The goal is to analyse both these datasets and make predictions based on them. We use LDA and QDA techniques to classify entries from sample.pickle. Linear Regression and Ridge Regression are used to analyze diabetes.pickle.

Dataset

Sample.pickle is a data matrix containing 250 datapoints of 2 features and 1 target variable. The target variable has values in the range [1, 5]. Our goal is to train a model which based on the values of two features, can predict the class to which that datapoint belongs to.

Diabetes.pickle is a data matrix containing 442 datapoints of 64 features and 1 target variable. The target variable tells about the level of diabetes in a patient. Our goal is to train a model which based on the values of 64 features, can predict the level of diabetes for a particular patient.

Problem 1: Experiment with Gaussian Discriminators

- Accuracy of Linear Discriminant Analysis is **97%**.
- Accuracy of Quadratic Discriminant Analysis is **96%**.

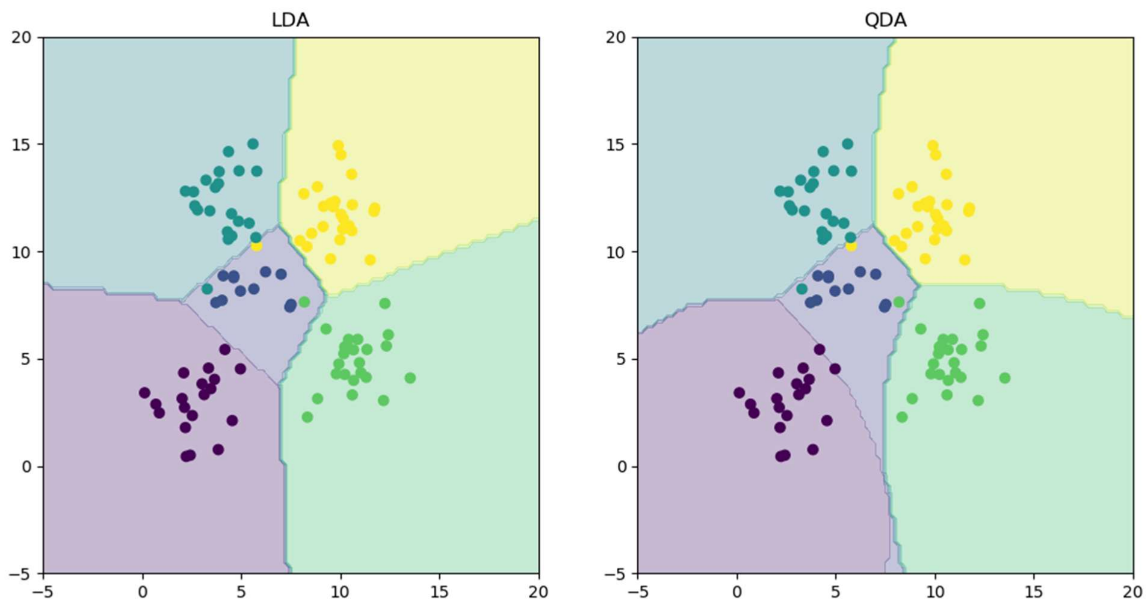

Figure 1: Classification using LDA and QDA.

In LDA,

- Every class is assumed to be normally distributed.
- Every class has its own mean.
- Only one covariance matrix is obtained for the data.

In QDA,

- Every class has its own mean.
- Every class has its own covariance matrix.

QDA has low bias and high variance, hence it is better for large datasets.

There is a difference in the two boundaries because covariance matrices used for LDA and QDA are different.

Problem 2: Experiment with Linear Regression

- By not having an intercept in our linear model, we are forcing the line to pass through the origin. This, in most cases, will give higher MSE than the case where we use the intercept term.
- MSE for training data without intercept: 19099.4468. MSE for training data with intercept: 2187.1602.
- MSE for test data without intercept: 106775.3614. MSE for test data with intercept: 3707.8401.
- From this we can conclude that it is always better to have an intercept term whenever we train our model. If, for some data, the intercept term is not required then the algorithm will automatically reduce the weight associated with it.
- Most optimum results are achieved when we use intercept.

Problem 3: Experiment with Ridge Regression

- The parameter *lambda* is chosen for which MSE of test data comes out to be minimum. In our case, the optimum value of *lambda* is found to be 0.06.
- MSE of training data for optimum lambda: 2451.528. MSE of test data for optimum lambda: 2851.33.
- L2 norm of weight vector found using linear regression: 1.5508×10^{10}. L2 norm of weight vector found using ridge regression: 920281.3569.
- Comparison of MSE for training and test data for different values of *lambda* can be seen in figure below.

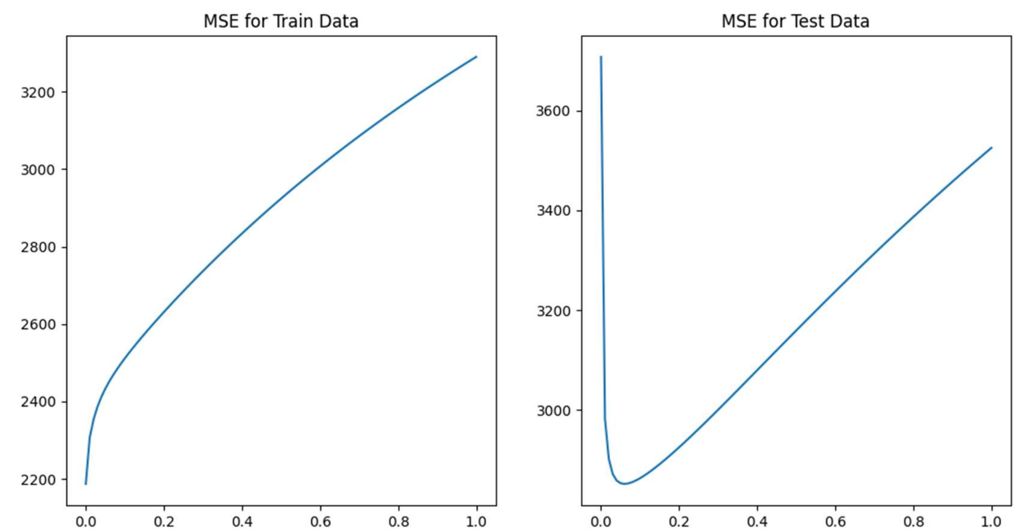

Figure 2: Comparison of MSE for different values of lambda on training and test data.

Problem 4: Using Gradient Descent for Ridge Regression Learning

- The MSE curve obtained by gradient descent is similar to the MSE curve obtained in Problem 3.
- The graph can be seen in figure below.

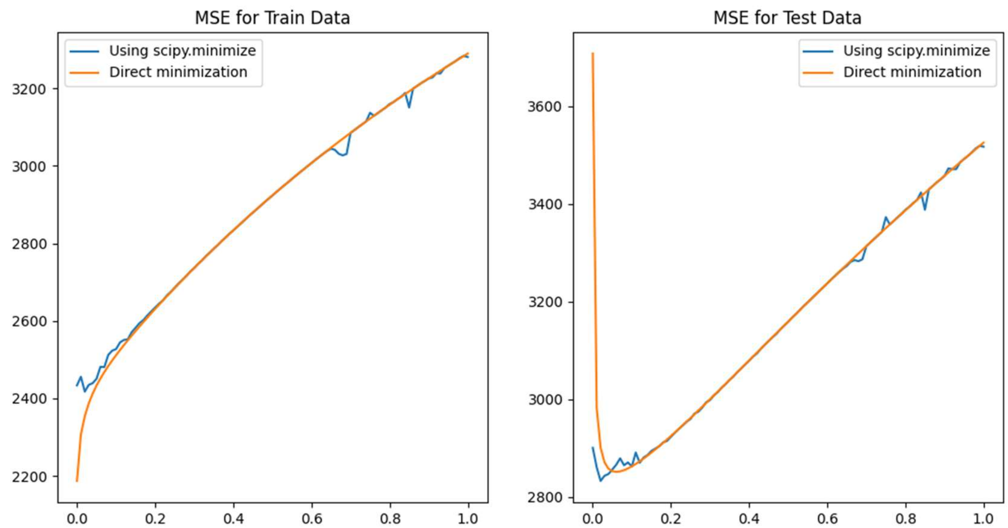

Figure 3: Comparison of MSE using gradient descent and direct minimization.

Problem 5: Non-linear Regression

- For *lambda* = 0, the lowest MSE for training data is observed at $p = 6$ (MSE = 3866.8834). This is simply because higher order polynomials without regularization tend to overfit the data.
- Similarly, the lowest MSE for test data is observed at $p = 1$ (MSE = 3845.0347). This means the linear equation is the best option to choose for the current data if we don't want to use regularization.
- For optimum lambda, *lambda* = 0.06, the MSE for training data has remained almost constant for $p \geq 1$ (MSE = 3950.6823). The lowest MSE for test data is observed at $p = 4$ (MSE = 3895.5826).
- The comparison for MSE values with/without regularization on training and test data for different values of p has been provided in figure below.

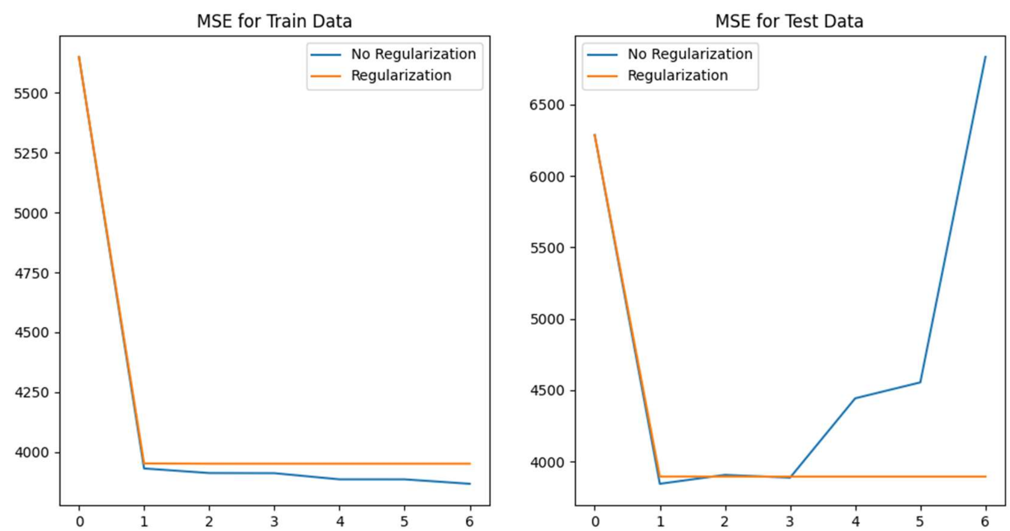

Figure 4: Comparison of MSE on different values of p and different regularization parameters.

Problem 6: Interpreting Results

- From results of problem 2 and problem 3 we have found that it is always a better idea to use an intercept term while training the model.
- Using regularization can improve the model considerably by not only preventing overfitting but by also reducing the magnitude of the weight vector learned. Concluding ridge regression is better than linear regression.
- We have also found that the weight vector learned from gradient descent is similar to the weight vector obtained by direct matrix minimization. It depends on the problem in hand to decide which method to use.
- For this problem, the best results were achieved when regularization parameter *lambda* was equal to 0.06.