UNIVERSITY AT BUFFALO

The State University of New York

Introduction to Machine Learning

(CSE 574)

Report of Programming Assignment 2

Group #26, Members:

Ting Zhou

Xuan Han

Apr 12, 2017

Problem 1: Experiment with Gaussian Discriminators

LDA Accuracy = 97.0%
QDA Accuracy = 96.0%
When we fix the covariance matrix for all classes in LDA, the discriminant function of each class C_k is,

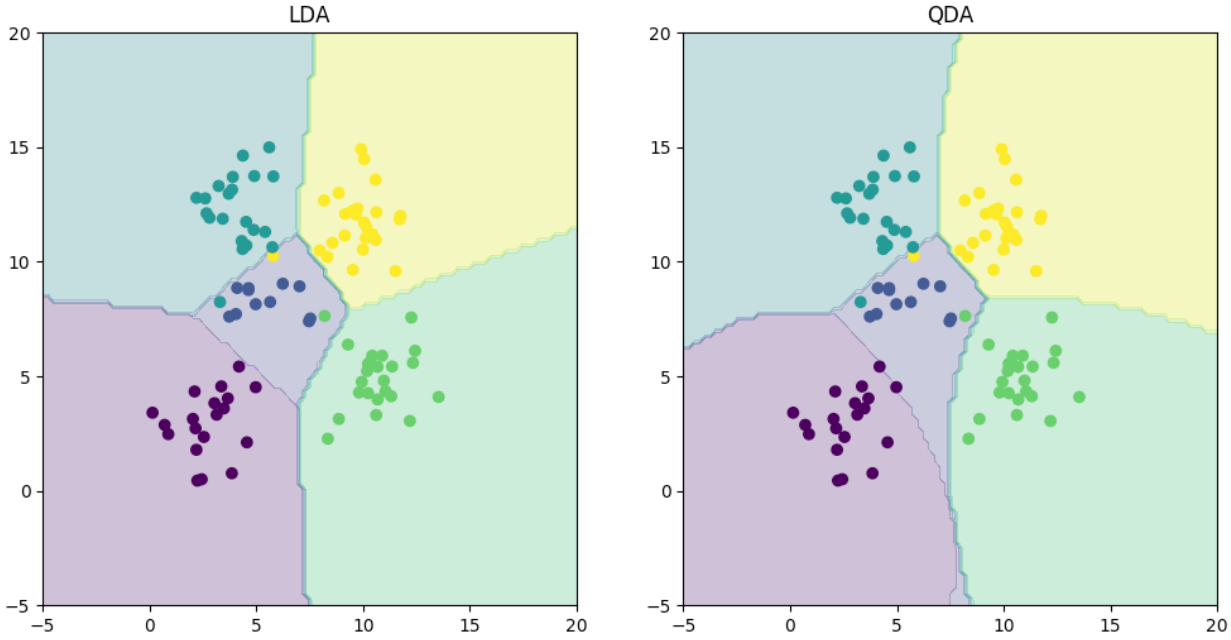
$$\delta_k(x) = -\frac{1}{2}x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + logP(C_k)$$

We can find that the discriminant function is linear. Thus the decision boundaries are linear in LDA.

QDA is a modification of LDA. In QDA, we allow the heterogeneity of the classes' covariance matrices. Then the discriminant function of each class C_k become to,

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + logP(C_k)$$

We can find that the discriminant function is quadratic. Thus the decision boundaries are quadratic in QDA.



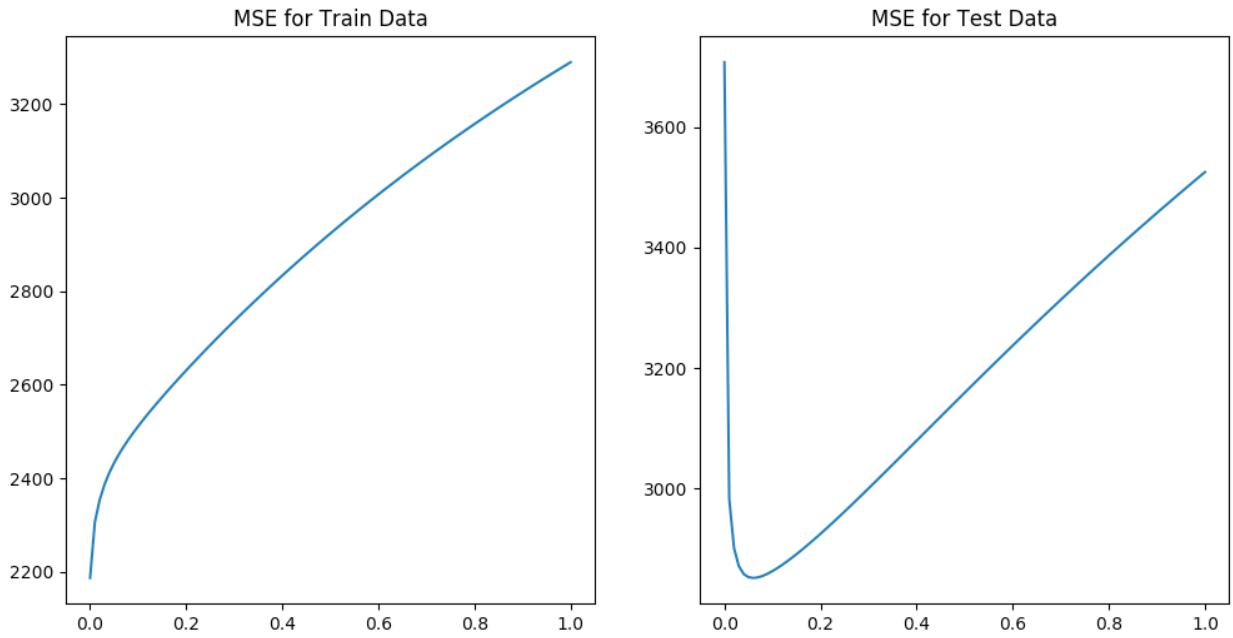
Problem 2: Experiment with Linear Regression

MSE without intercept = 106775.36155512
MSE with intercept = 3707.84018132
The linear regression model with intercept is better, because without the bias the regression model has to go through the original point. With the bias, the model can move freely and get a better regression.

Problem 3: Experiment with Ridge Regression

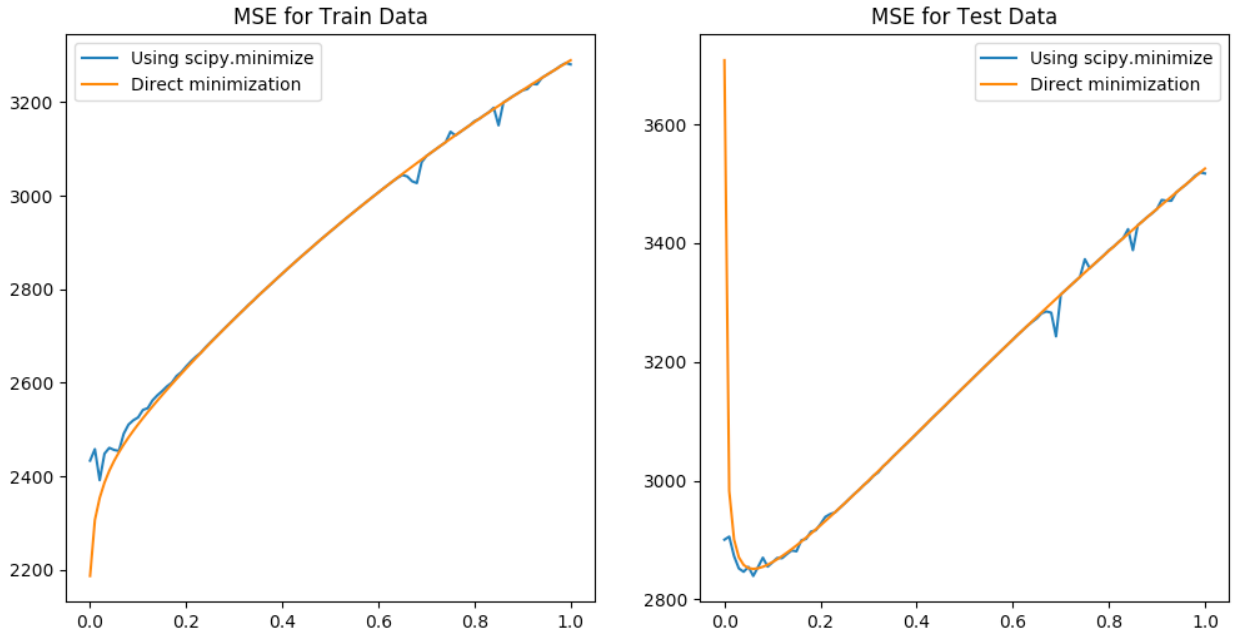
The best λ is 0.06, when the MSE for the test data reach the minimum point. We should choose the λ based on the minimum MSE from test data rather than the training data. Because if we choose $\lambda = 0$, the minimum MSE for training data will be zero, but the model would be

overfitting and the MSE for test data would intend to infinite. Thus Ridge Regression model could help us avoid overfitting by adding the regularization term.



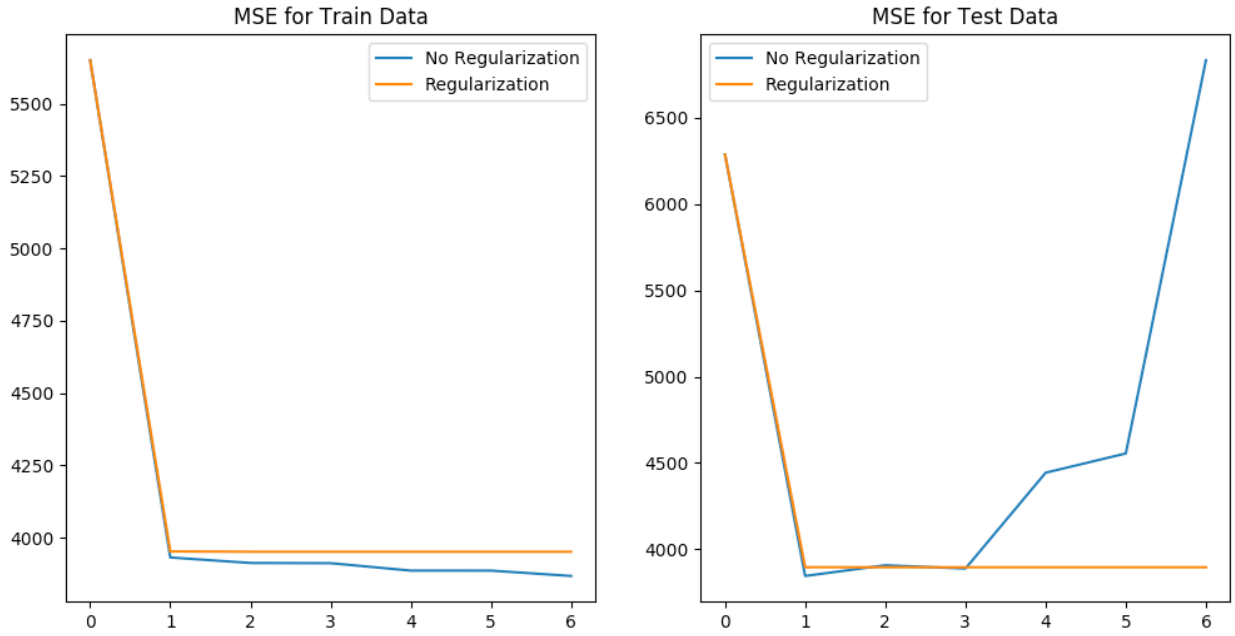
Problem 4: Using Gradient Descent for Ridge Regression Learning

The MSE from the analytical equation and the MSE from the gradient descent are pretty close both for training data and test data. The optimal λ's are 0.06 for both two methods, however the gradient descent could avoid to calculate the matrix inverse of $(XX^T)^{-1}$, which not only time costly as O(d^3) but also might be singular.



Problem 5: Non-linear Regression

For non regularization model, with the increase of the order p the model would be more accurate for training data, however the MSE for test data will increase very fast because of the overfitting. After adding the regularization term, when p is equal or larger than 1, the MSE for both training data and test data become to almost same. Thus the optimal value of p is one with the regularization, because there is no need to use higher order regression model in this case.



Problem 6: Interpreting Results

We should use regularized squared loss function with the regularization term ($\lambda = 0.06$) to quantize the accuracy of the regression model to avoid overfitting. The linear regression model (p=1) with intercept is recommended based on the comparison of MSE between with intercept and without intercept. Compared with Ridge Regression, minimizing the MSE we should use gradient descent method in order to avoid calculating the matrix inverse of $(XX^T)^{-1}$ which not only time costly as $O(d^3)$ but also might be singular. Another drawback of the Ridge Regression is that if the λ goes to zero, the MSE will goes to infinite for test data, however the gradient descent would not drive to the model into insane overfitting even when λ goes to zero.

In sum, a linear regression model (p=1) with intercept and regularization ($\lambda = 0.06$) should be used to predict the level of diabetes. Minimizing the regularized squared loss function should use gradient descent.