Problem 1

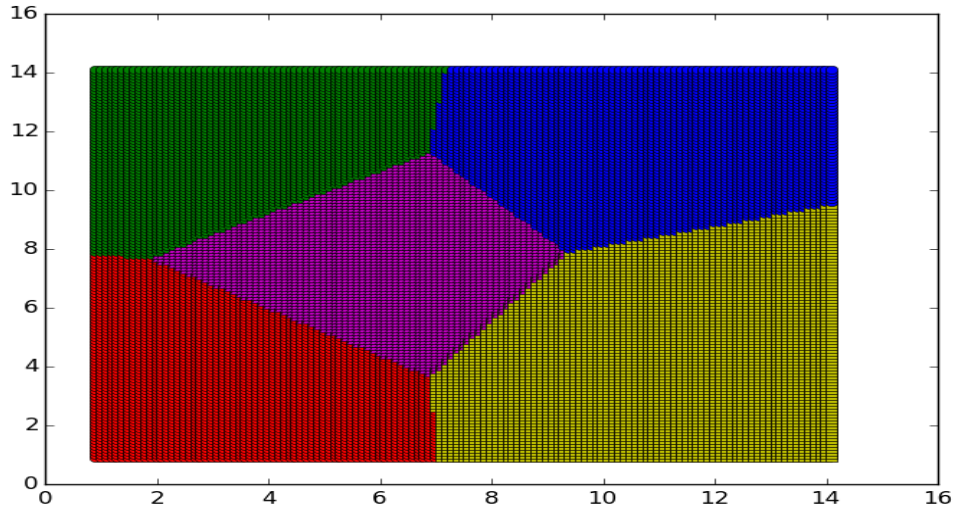
Accuracy for LDA was found to be **97%**
Accuracy for QDA was found to be **96%**

To plot the discriminating boundary we have created a meshgrid that goes from minimum values of the two dimensions of input to the maximum values. This was found to be (0.9,0.9) for the minimum values and (14.2,14.2) for the maximum values.

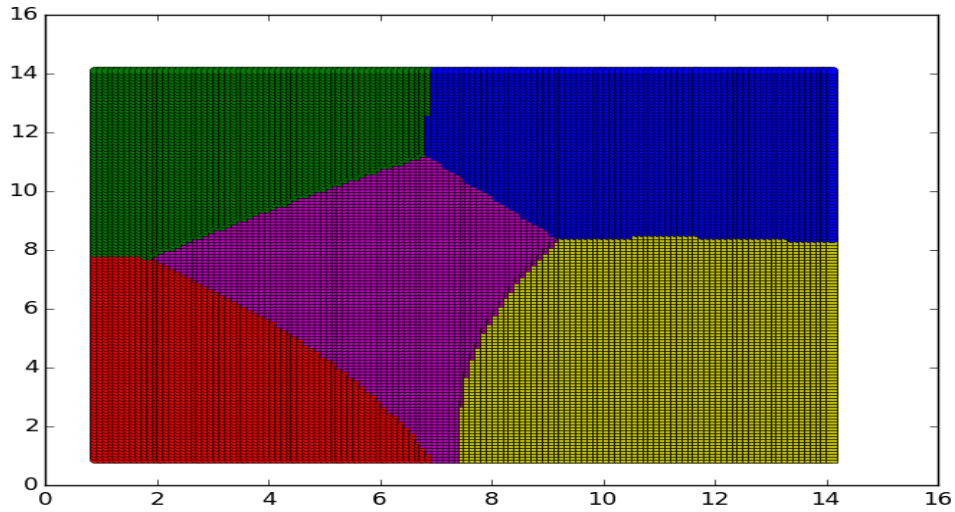
Legend for graph:

Red – Class 1
Magenta – Class 2
Green - Class 3
Yellow – Class 4
Blue – Class 5

LDA Discriminating Boundary



QDA Discriminating Boundary



$$p(y=c|\mathbf{x},\boldsymbol{\theta}) \propto \pi_c \exp\left[\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c\right] \quad (4.34)$$

$$= \exp\left[\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \log\pi_c\right]\exp[-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}] \quad (4.35)$$

As you can see for LDA in eq (4.35), the covariance element is independent of class. So, the quadratic term can be cancelled out from the numerator and denominator while finding the posterior density of y. The leads to a separating line which is linear in x when finding the equation of the separating line.

$$p(y=c|\mathbf{x},\boldsymbol{\theta}) = \frac{\pi_c|2\pi\boldsymbol{\Sigma}_c|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c^{-1}(\mathbf{x}-\boldsymbol{\mu}_c)\right]}{\sum_{c'}\pi_{c'}|2\pi\boldsymbol{\Sigma}_{c'}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{c'})^T\boldsymbol{\Sigma}_{c'}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{c'})\right]} \quad (4.33)$$

Whereas for QDA, eq (4.33) shows that the quadratic term varies for each class as it contains different covariance matrices for each class. This leads to the term not being cancelled out and hence non-linear mappings when we try to find the equation of the separating line between classes which comes out to be quadratic in x.

Problem 2

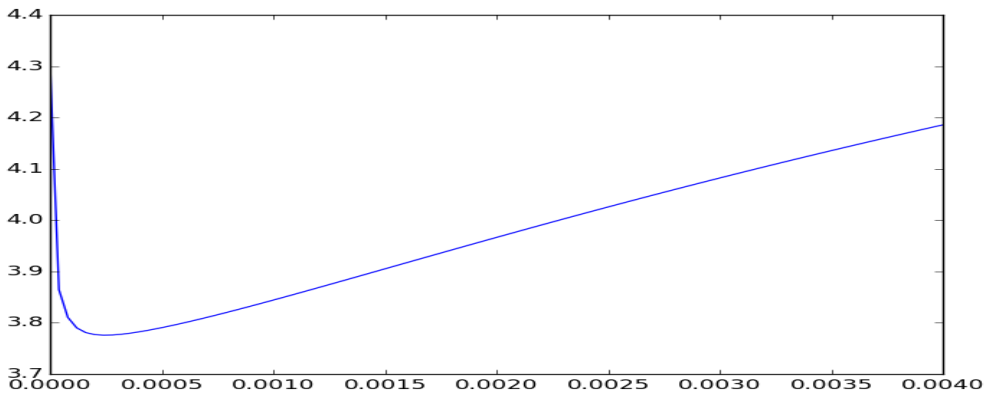
RMSE without intercept (bias) term:
For train data: 8.88388
For test data: 23.10577

RMSE with intercept (bias) term:
For train data: 3.0063
For test data: 4.305717

RMSE with intercept is better; this is because with a bias term, the line learn't need not pass through the origin. This results in a more precise line being learnt, with a lower RMSE. If we did not have an intercept, the learn't line will have to pass through origin and it will give correct results only if the mean of the data is centered , which is not usually the case.

Problem 3

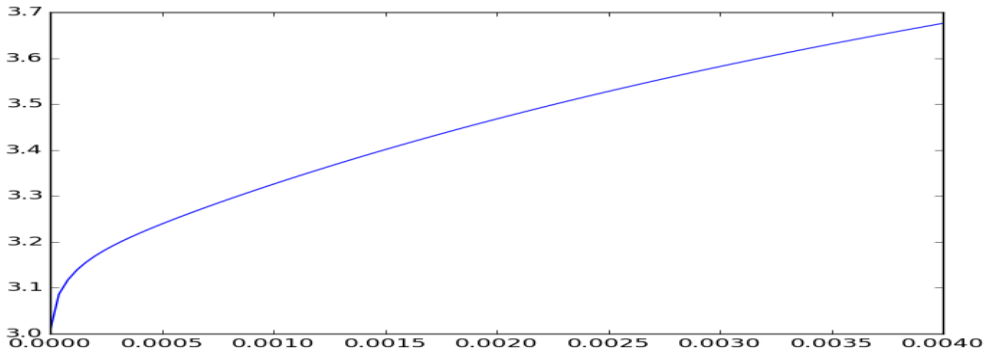
Error on Test Data (lambda from 0 to 0.004, steps = 101)



x axis – lambda y axis - RMSE

The optimal value of lambda was found to be 0.00024

Error on Train Data (lambda from 0 to 0.004, steps = 101)



x axis – lambda y axis - RMSE

RMSE for Linear Regression	RMSE for Ridge Regression (lambda in brackets)
4.305717	4.305717 (lambda = 0)
4.305717	3.77582332 (lambda = 0.00024)
4.305717	4.1856 (lambda = 0.004)

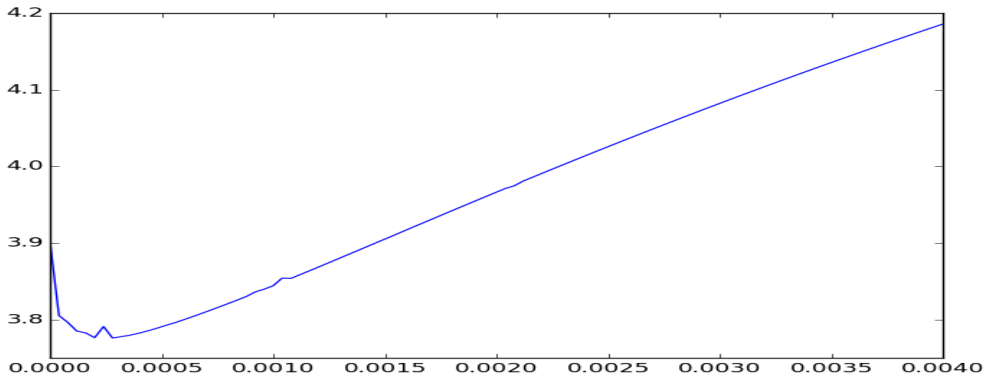
The use of regularization tends to avoid overfitting the training data and as you can see from the table above leads to a lesser RMSE at optimal lambda value than linear regression.

As compared to OLE, at an optimal lambda value, ridge regression will produce more error on training data than OLE but lesser error on test data due to the effect of regularization.

The magnitude (Eucledian norm) of weight vector learnt by OLE is 124531.52 and the magnitude of weight vector learnt by Ridge Regression is 435.83. Due to the regularization term in Ridge Regression, the magnitude of weights is suppressed to learn smooth separating lines and avoid overfitting , which is evident from the result.

Problem 4

Error on Test Data (lambda from 0 to 0.004, steps = 101)

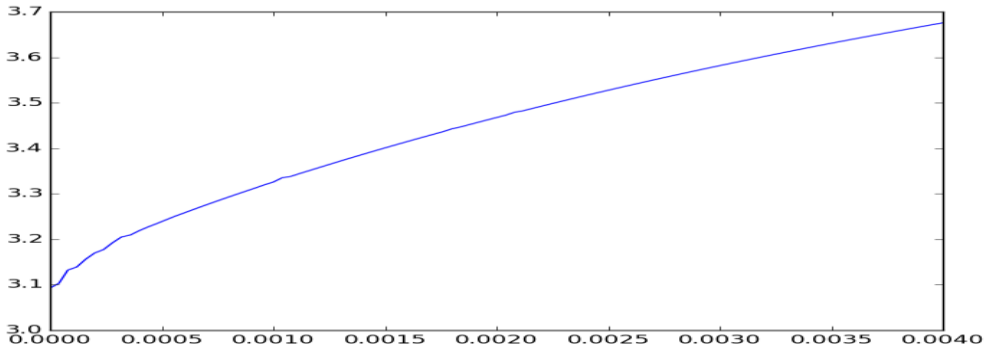


x axis – lambda y axis – RMSE

Optimal value of lambda using the gradient descent method with regularization was found to be **0.00028**

This is in sync with problem 3 where optimal value of lambda was found to be **0.00024**

Error on Train Data (lambda from 0 to 0.004, steps = 101)

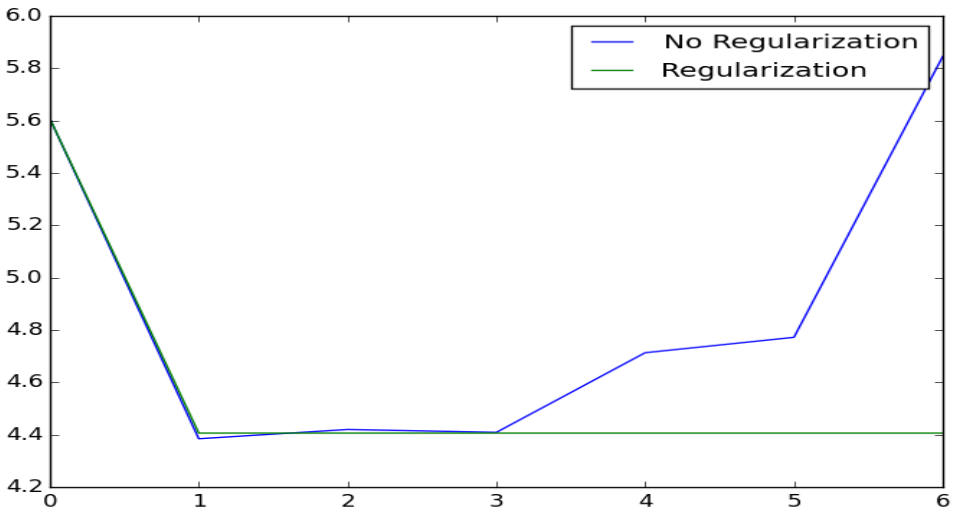


x axis – lambda y axis - RMSE

In both these cases (problem 3 and 4), the methods yield the same results although their approaches of getting there are different. Gradient Descent can be computationally efficient as it avoids the difficulties involved in calculating the inverse of a matrix.

Problem 5

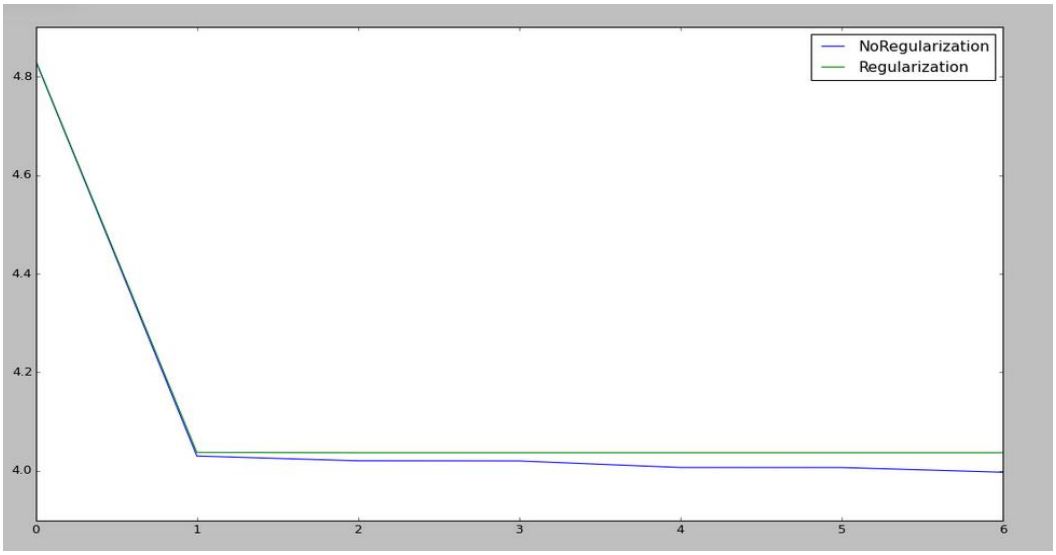
Plot of lambda = 0 and lambda = 0.00028 (optimal) v/s number of dimensions on test data



x axis – p y axis – RMSE

Beyond , p = 3, the RMSE for lambda = 0 keeps increasing as higher dimensions makes it fit the training data more tightly, thus increasing the RMSE on test data due to the model that is overfitted on train data. But, the RMSE for lambda = 0.00028 (optimal) avoids overfitting and hence remains around the same level even at p = 4, 5 and 6.

Plot of lambda = 0 and lambda = 0.00028 (optimal) v/s number of dimensions on train data



x axis – p y axis – RMSE

As we increase the number of dimensions , the RMSE on the train data goes down due to overfitting with train data. However , if we use the regularization term , the regularization term tries to lessen the impact of higher order terms by avoiding overfitting and result is as we see in the graph for the regularization plot.

The optimal dimension is linear (p=1 i.e. data with bias) as you can see from the plot above. For the data given to us, mapping non-linear curves is not going to be of any benefit as linear mapping is doing the most efficient job.

Problem 6

Train Data

RMSE	Approach
3.0063	Linear Regression with intercept
3.00630212	Ridge Regression
3.08852925	Ridge Regression (GD)
3.99	Non – Linear Regression (without regularization)

Test Data

RMSE	Approach
4.305	Linear Regression with intercept
3.77	Ridge Regression (lambda – 0.00024)
3.7758	Ridge Regression (GD) (lambda – 0.00028)
4.38	Non – Linear Regression (without regularization)

To choose the best setting , we should choose the approach which produces the least error on the test data. A low error on training data indicates overfitting which might lead to increased RMSE on test data.

From the above experiments we find that , ridge regression with lambda 0.00024 and ridge regression with gradient descent with lambda 0.00028 perform equally well on test data and hence are the final recommendations to predict diabetes level using input features.