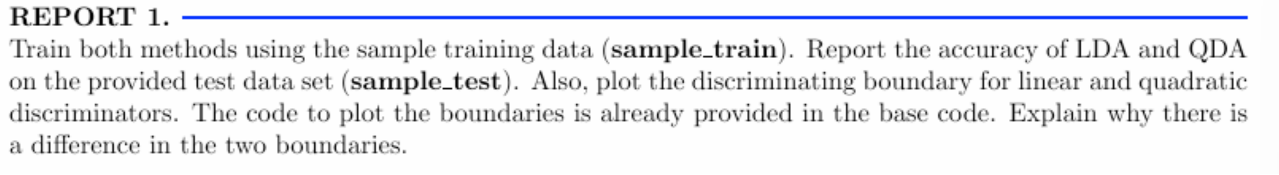
INTRO

Team Name : #GoBlue
Team Members :

S.No.	Team Members	UB Number	Mail ID
1	Sharanya Nallapeddi	50593866	snallape@buffalo.edu
2	Keerthana Vangala	50604773	kvangala@buffalo.edu
3	Aravind Mohan	50611294	amohan22@buffalo.edu



REPORT 1.

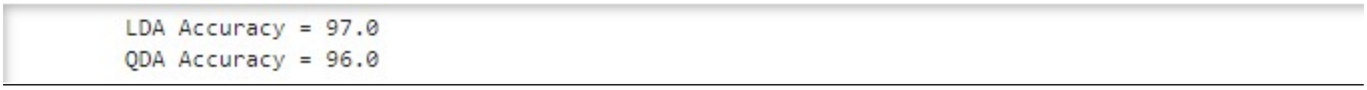
Train both methods using the sample training data (**sample_train**). Report the accuracy of LDA and QDA on the provided test data set (**sample_test**). Also, plot the discriminating boundary for linear and quadratic discriminators. The code to plot the boundaries is already provided in the base code.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

Report 1:
Accuracy of LDA and QDA:
LDA:

Hence the Accuracy of LDA is 97%

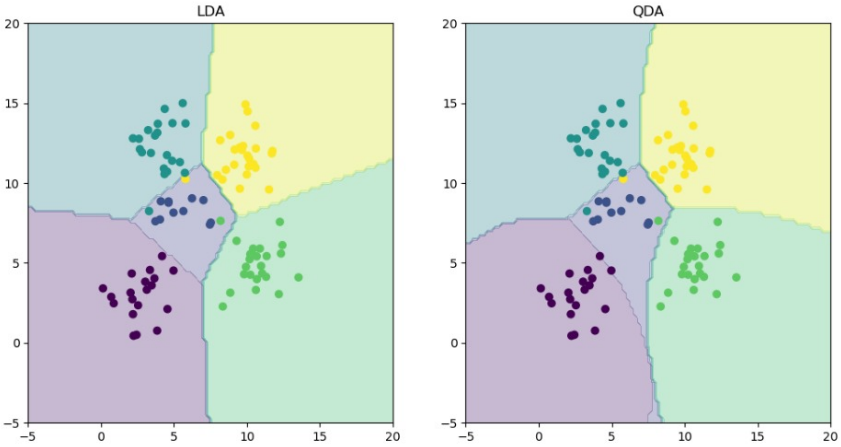
QDA:

Hence the accuracy of QDA is 96%.



```
LDA Accuracy = 97.0
QDA Accuracy = 96.0
```

Boundaries for LDA and QDA:



Reason for difference in two boundaries:

LDA: LDA assumes that every class has the same covariance matrix, which makes the model simpler but also restricts its adaptability. According to this assumption, each class's data have the same form and distribution (elliptical in nature), with the mean vectors serving as the only source of variation.
Because the distinction between classes is based on the same covariance structure after projecting the data onto the new space, the decision boundaries in LDA are linear. Given that the covariance matrix remains constant for every class

INTRO

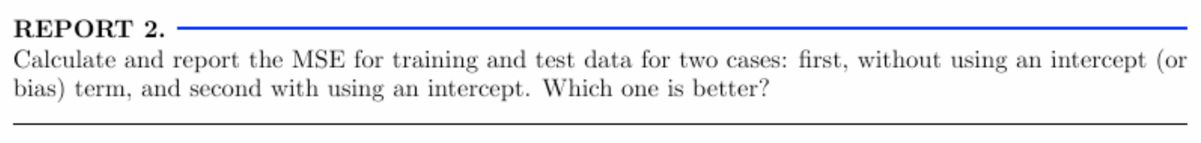
Even in cases when the class distributions are not completely aligned, LDA assumes the same covariance structure for all classes, resulting in straight lines connecting the classes and revealing the linear decision boundaries.

QDA:

QDA permits each class to have its own covariance matrix, in contrast to LDA. Since each class's data distribution can have a varied shape and direction, this leads to greater flexibility.

Because the covariance matrices in QDA vary throughout classes, the decision boundaries are more intricate and quadratic. More flexible decision regions result from the boundaries' ability to curve and change in accordance with the distribution and dispersion of data for each class.

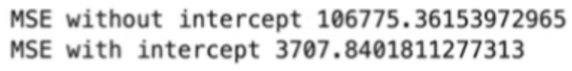
Since QDA adjusts the decision boundaries to the unique shapes and orientations of the class distributions, the more flexible boundaries permit curved separations between the classes.



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?

Report 2:



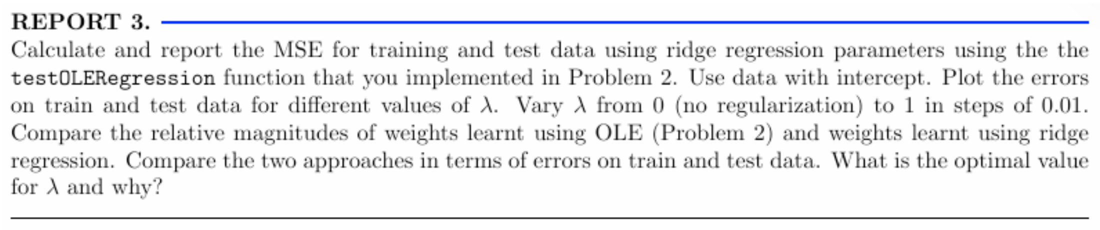
```
MSE without intercept 106775.36153972965
MSE with intercept 3707.8401811277313
```

The Measure of MSE with Intercept is much smaller than the MSE without intercept, hence MSE with intercept is better and closer to real representation of Data.

Reasoning:

Since all independent variables can be zero in the model with intercept, an intercept means that the line or plane can have a value. Usually, this results in a lower MSE by offering a more precise and flexible fit to the data.

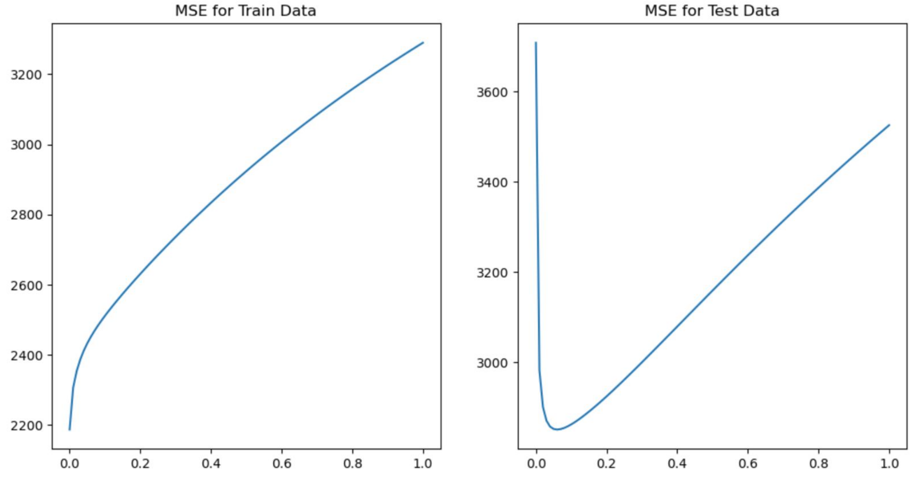
Whereas, the origin, or the point at which all independent variables are zero, is assumed to be passed through by the line or plane in the model without intercept. Given that the genuine relationship between the dependent and independent variables does not always flow through the origin, this constraint might not be suitable for many datasets. Consequently, the error (MSE) is significantly higher.



REPORT 3.

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?

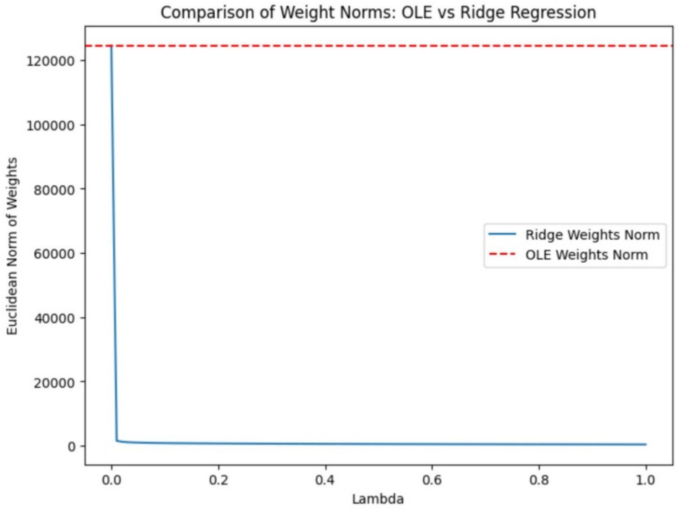
Report 3: Plot for errors on train and test for different values of λ

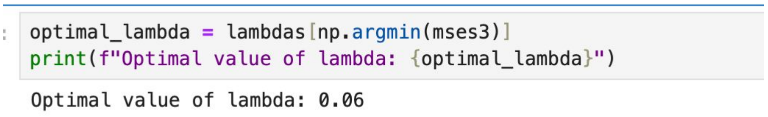


INTRO

Comparison of relative magnitudes of weights:

- Weights learnt using OLE: The weight are determined based on the data without punishing large coefficients. This leads to larger weights especially if data set has lot of collinearity or outliers.
- Weights using ridge regression: Ridge regression introduces a regularization term (L2 regularization) that punishes the magnitudes of weights, thereby shrinking the coefficients. The larger the value of λ





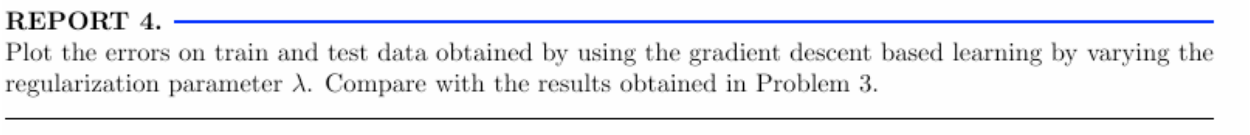
```
optimal_lambda = lambdas[np.argmin(mses3)]
print(f"Optimal value of lambda: {optimal_lambda}")
```

```
Optimal value of lambda: 0.06
```

Value of λ =0.06 optimal because, we observed minimum MSE on test set, which means this value of λ provides best trade of between bias and variance.

For $\lambda < 0.06$ model mostly overfits as seen by increasing test MSE.

For $\lambda > 0.06$ model mostly underfits as regularization term is stronger leading to high test MSE.

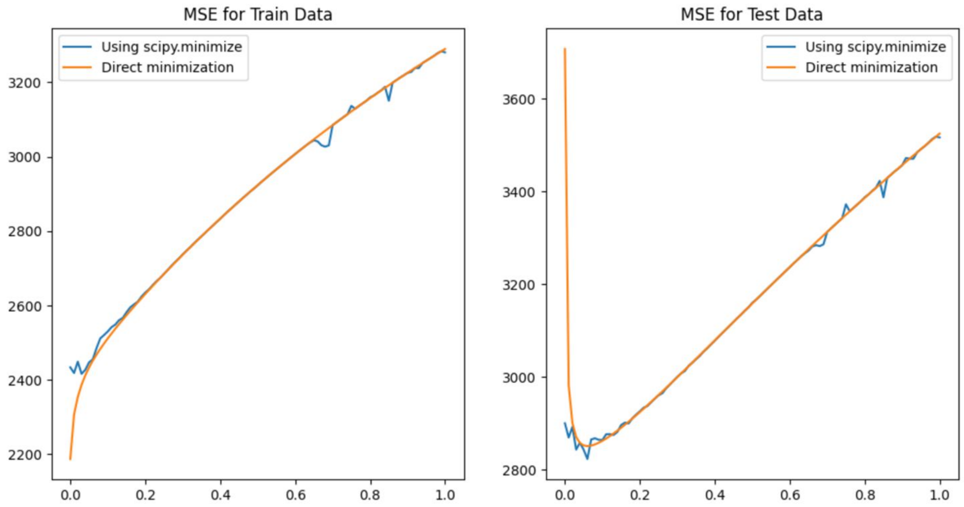


REPORT 4.

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

Report 4:

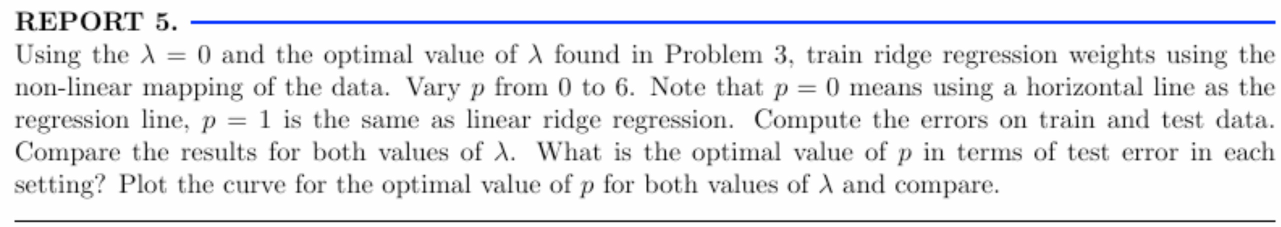
Plot for errors on train and test obtained by gradient descent:



For accuracy, for test and train data, the MSE values produced by both approaches are remarkably close. But the gradient descent approach (scipy.minimize) exhibits some minor variances or oscillations, especially for smaller values of λ. In this regard, direct minimization is more stable. For test and train data, the MSE values produced by both approaches are remarkably close. But the gradient descent approach (scipy.minimize) exhibits some minor variances or oscillations, especially for smaller values of λ. In this regard, direct minimization is more stable.

Considering efficiency, Due to the matrix inversion stage, direct minimization may struggle with very big datasets but is faster and more accurate for smaller datasets. Large datasets scale better with gradient descent, but it needs more careful calibration and can produce small deviations, as the second plot illustrates.

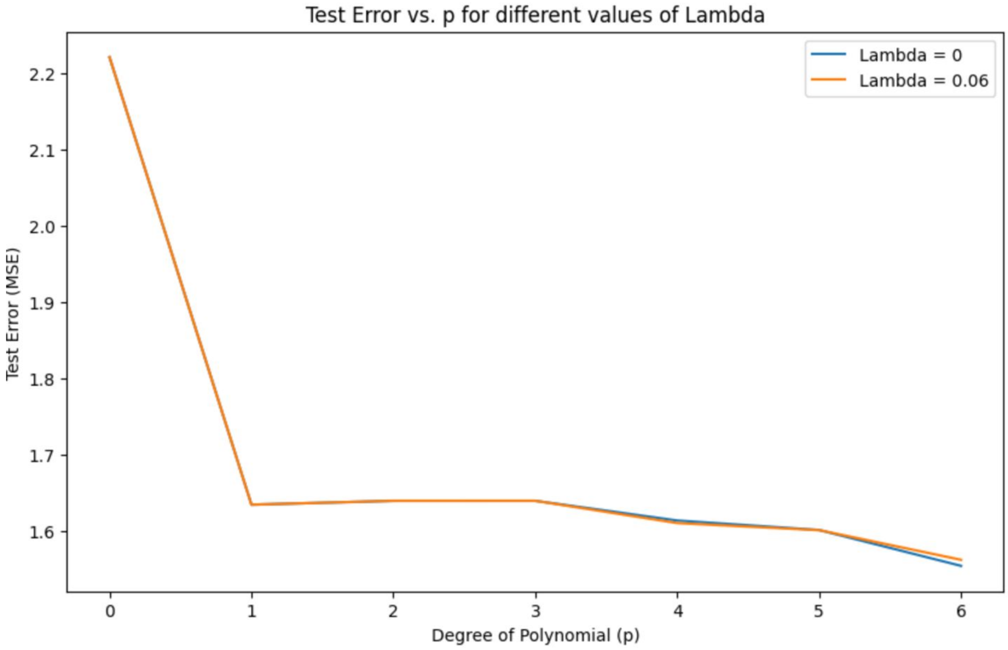
The test MSE is minimized at a similar optimal λ (about 0.06), which is appropriately identified by both methods. This demonstrates that both methods are successful in determining the ideal regularization parameter, despite the variations in their optimization processes.

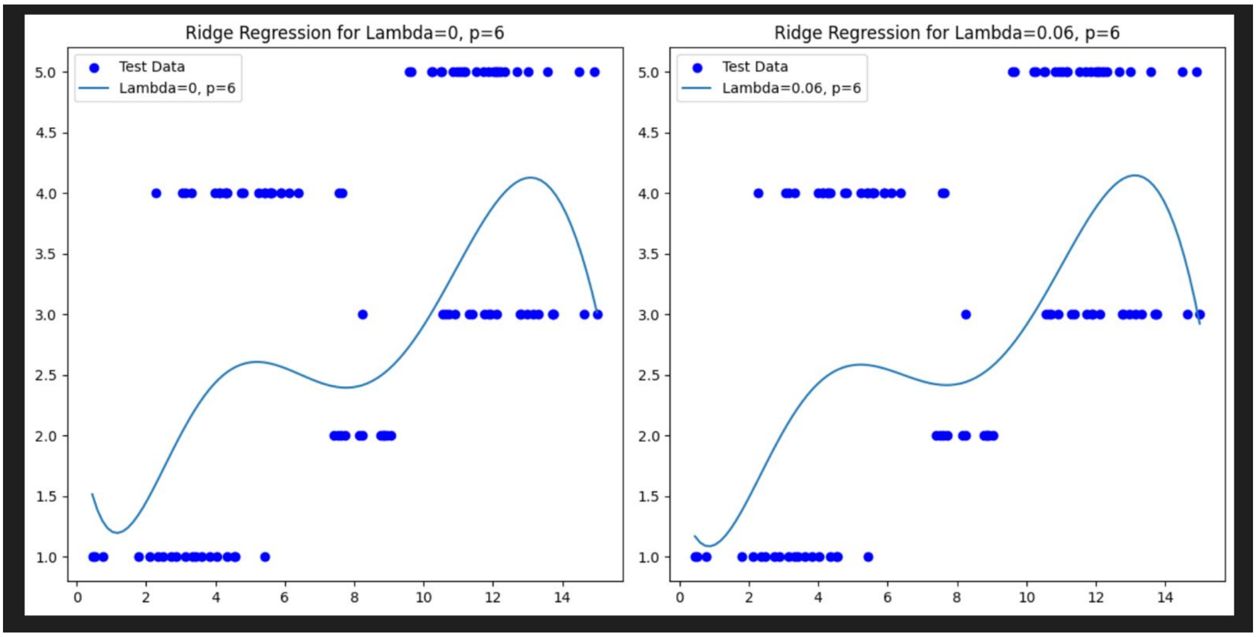


REPORT 5.

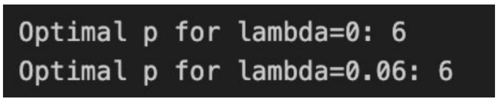
Using the $\lambda = 0$ and the optimal value of λ found in Problem 3, train ridge regression weights using the non-linear mapping of the data. Vary p from 0 to 6. Note that $p = 0$ means using a horizontal line as the regression line, $p = 1$ is the same as linear ridge regression. Compute the errors 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.

Errors on train and test data:

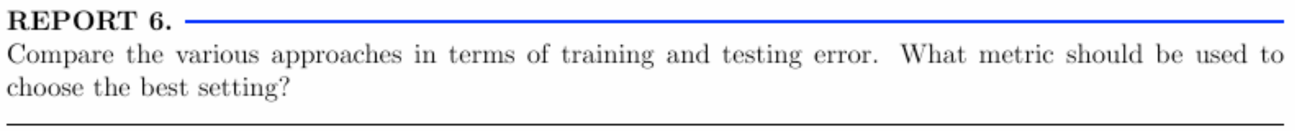




INTRO



```
Optimal p for lambda=0: 6
Optimal p for lambda=0.06: 6
```



REPORT 6.

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

1. OLE:

Model with intercept lead to a lower MSE value compared to the one without intercept.
Concluding that including an intercept will lead to optimal solution, as it can include relationships where the independent variables (features) can be equal to 0.

2.Ridge regression for MSE Comparison:

The optimal lambda was found to be 0.06. This value increases efficiency of MSE value on test set by maintaining balance between bias and variance.

If $\lambda = 0.06 \rightarrow optimal\ (no\ overfitting\ or\ underfitting)$

If $\lambda < 0.06$ →overfitting, increased MSE

If $\lambda > 0.06$ →underfitting due to strong regularization.

3.Ridge regression using Gradient Descent:

The model was trained using gradient descent, and it displayed consistent MSE values for both training and test datasets.
Overall, MSE values from gradient descent and direct minimization were similar, however gradient descent showed some minor oscillations for smaller λ values.
It was evident that gradient descent scales better for bigger datasets, although direct minimization is quicker and more accurate for smaller datasets.

4. Non-Linear Regression

- Error in Training and Validation for p = 6 and $\lambda = 0$ λ = 0 and $\lambda = 0.06$ λ = 0.06:
  The model with polynomial features of degree 6 ($p = 6$ p=6) tends to overfit the training data when employing $\lambda = 0$ λ=0 (no regularization), leading to a very low training error but a considerable rise in test error. This is due to the fact that a high-degree polynomial adds a great deal of complexity, capturing noise in the data as well as the main trend.

- Adding regularization helps to alleviate the overfitting problem for $\lambda = 0.06$ λ = 0.06. Large coefficients are penalized by regularization, which reduces the impact of higher-degree terms in the polynomial. This leads to a more balanced training error and test error, indicating better generalization.
  In comparison to $\lambda = 0$ λ = 0, the test MSE for $\lambda = 0.06$ λ = 0.06 was found to be much lower, suggesting superior generalization to unobserved data.

- The model overfits the training set without regularization ($\lambda = 0$ λ = 0), even if raising p to 6 gives the model more flexibility in fitting the data. When the model incorporates noise, the test error rises rapidly, particularly when the degree polynomial is high.

- Regularization manages the model complexity for $\lambda = 0.06$ λ = 0.06, enabling a better balance between fitting the training data and preserving respectable performance on the test data. Compared to $\lambda = 0$ λ=0, this results in less overfitting and a smaller test error.

Model Recommendation

INTRO

The model with $\lambda = 0.06$ λ = 0.06 and $p = 6$ p = 6 has the best generalization performance, according to the analysis. This configuration is appropriate for forecasting diabetes levels because it reduces overfitting and yields the lowest test MSE.

Model Selection measure: The Mean Squared Error (MSE) is still the suitable model selection measure. The test MSE shows how effectively the model generalizes to new data, and it represents the average squared difference between the predicted and actual values for diabetes.

Comparison of Training and Test Errors:
The model with $\lambda = 0$ λ = 0 has a little larger training error than the one with **G** = 0.06 λ = 0.06 and $p = 6$ p = 6, but the test error is significantly lower, indicating improved generalization.
With $p = 6$ p=6 and $\lambda = 0.06$ λ = 0.06, bias and variance are balanced, yielding a well-generalized model that stays clear of overfitting traps.