INDIAN INSTITUTE OF TECHNOLOGY, DELHI

ASSIGNMENT REPORT

Assignment 1

APAR AHUJA | ENTRY NO. 2019CS10465

Course - COL774 | Prof. Parag Singla

Compiled on September 12, 2021

Contents

Chapter 1

Linear Regression

1.1 *a*

Learning Rate: 1
Number of iterations: 2
Final Cost: 1.1947898 x 10^{-6}
Theta: $(\theta_1, \theta_0) = (0.0013402, 0.9966201)$
Stopping Criteria: $|J(\theta_t) - J(\theta_{t-1})) \leq 10^{-20}$

1.2 *b*

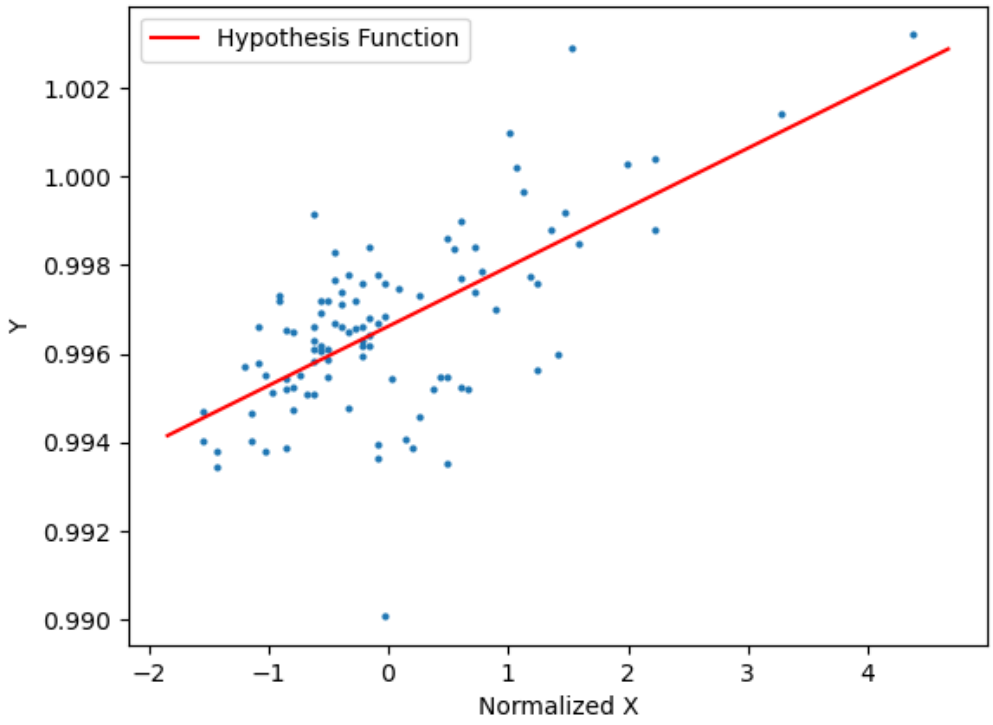

Figure 1.1: 1b

1.3 *c*

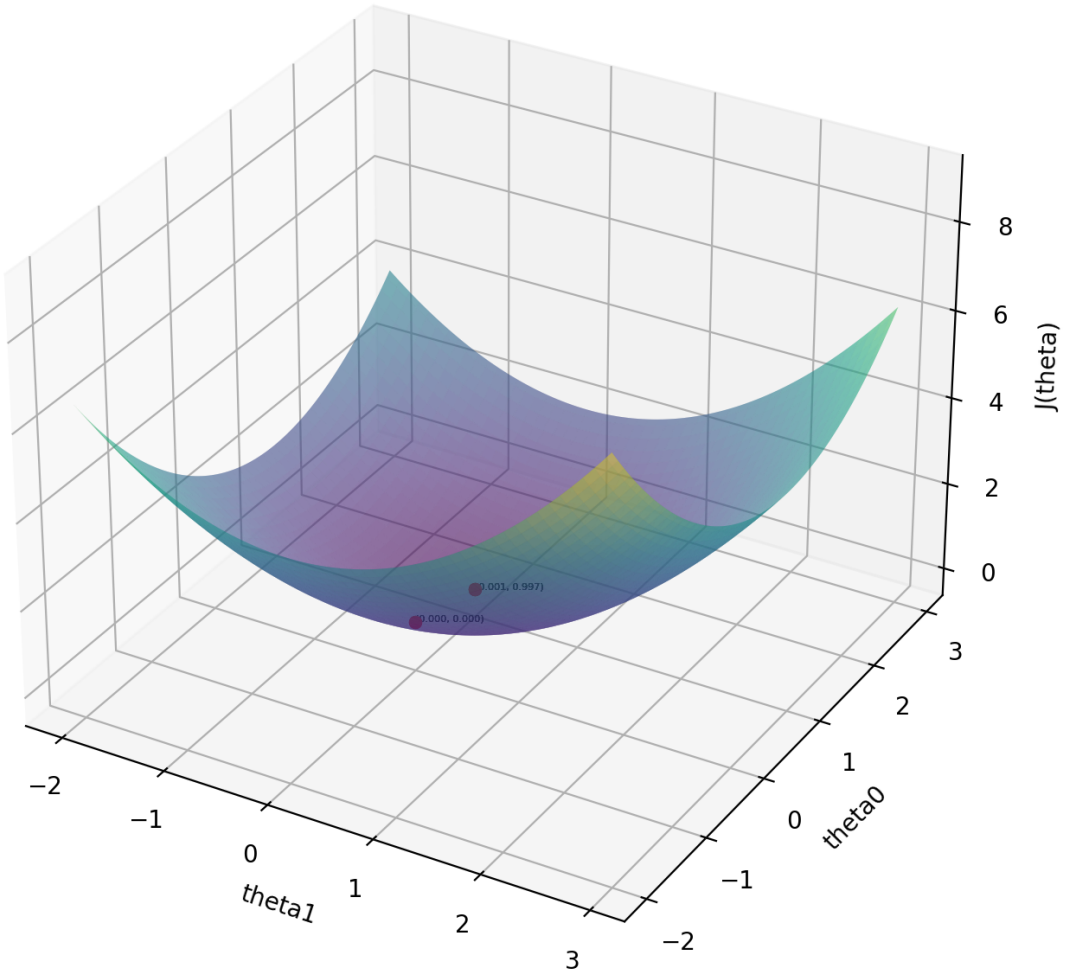

Figure 1.2: 1c

1.4 *d*

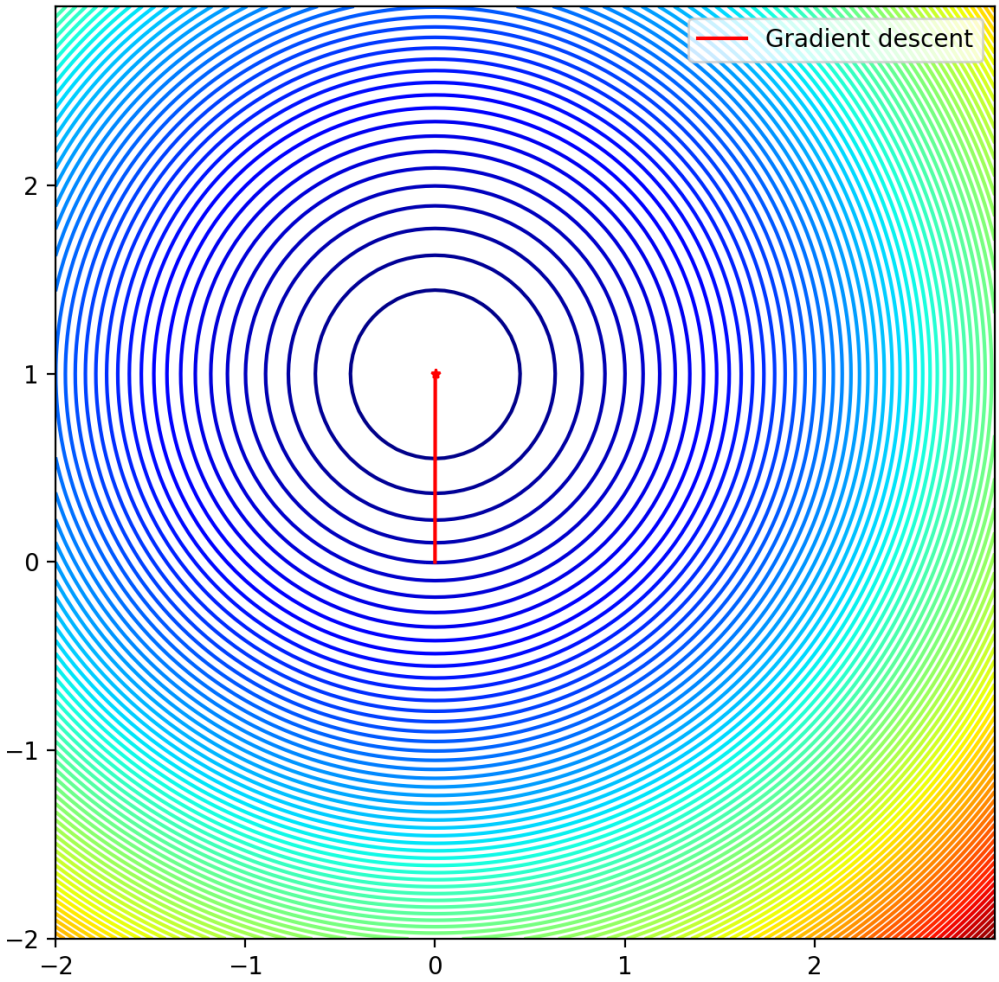

Figure 1.3: 1d

1.5 *e*

Animations for different values of eta look very similar and are attached with the code. I observe that number of iterations for eta = 0.1 are 198, eta = 0.025 are 793 and for eta = 0.001 are 18341. Hence, rate of convergence increases with eta for the given values.

Chapter 2

Sampling and Stochastic Gradient Descent

2.1 *a b c*

Convergence Criteria - The absolute difference of the average cost in last k = 2000 iterations and average cost in k iterations before these should be less than ϵ.

$\theta_{original} = (\theta_2, \theta_1, \theta_0) = [2, 1, 3]$
Learning rate $\eta = 0.001$
Test Error($\theta_{original}$) = 0.9829469215

r	$\theta_{learnt} = (\theta_2, \theta_1, \theta_0)$	ϵ	Iteration	Time (in s)	Test Error
1	[2.0029740, 0.9892885, 2.980094]	10^{-7}	30109	5.7630	0.9901652224
100	[1.9956010, 1.0022010, 2.9764057]	10^{-7}	17421	3.0024	0.9849137412
10000	[1.99983221, 0.99856521, 3.00359158]	10^{-7}	27997	9.1022	0.9831435038
1000000	[1.99906421, 1.00170245, 2.99147114]	$*10^{-6}$	23713	118.0960	0.9832273683

*ϵ was set to a higher value as the program for this large batch size was particularly slow.

Comments -
1. All the batches converge to values very close to the original hypothesis.
2. The values are within 0.2% of original theta value.
3. Larger batch sizes lead to faster convergence.
4. The number of iterations are lower for larger batch sizes as the gradient is more precise.
5. Test errors are fairly close to the original hypothesis error.
6. All models converged fairly close to the original values.

2.2 *d*

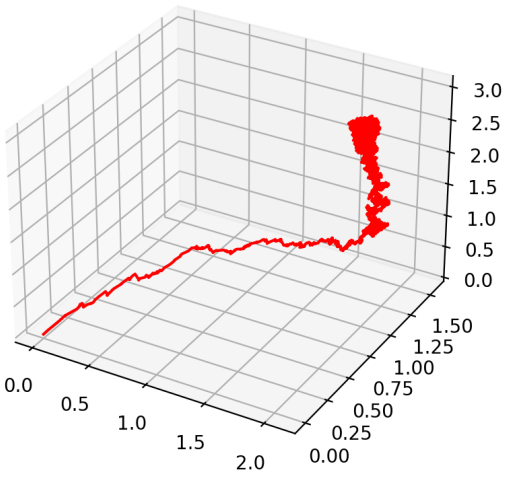

Figure 2.1: Batch Size = 1

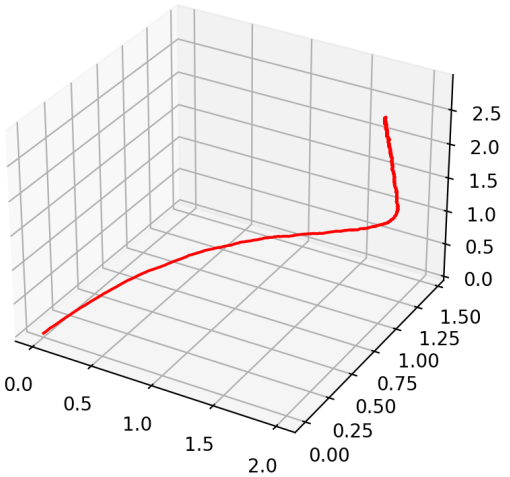

Figure 2.2: Batch Size = 100

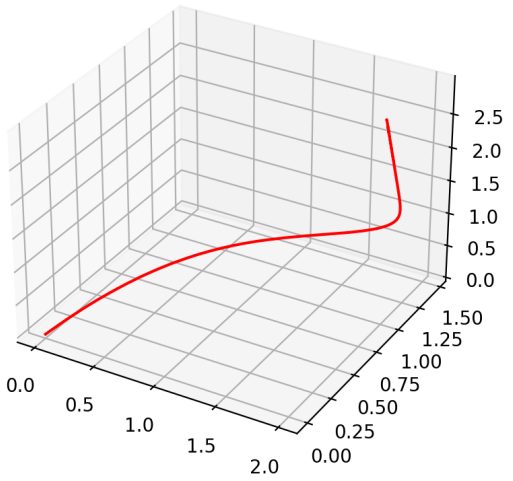

Figure 2.3: Batch Size = 10000

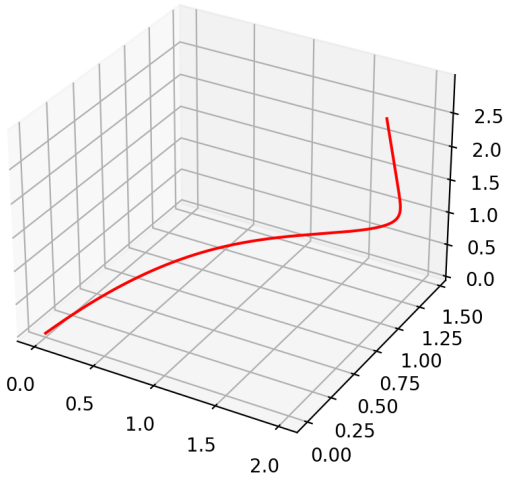

Figure 2.4: Batch Size = 1000000

The graph for batch size 1 is very noisy due the small batch size. As the batch size increases the gradient direction is more fine tuned for each iteration leading to a smoother trajectory to the optimal solution. Although, this smoother trajectory comes with a significantly higher runtime.

Chapter 3

Logistic Regression

We have $\mathrm{h}_\theta(\mathrm{x}^{(i)}) = \frac{1}{1+e^{\theta^T x^{(i)}}}$ then gradient $= \frac{X^T(Y-h_\theta(X))}{m}$ and the hessian $= -\frac{X^T(X*h_\theta(X)*(1-h_\theta(X))}{m}$

Here $h_\theta(X)$ denotes the vector h where $\mathrm{h}_i = \mathrm{h}_\theta(\mathrm{x}^{(i)})$. A∗B denotes array broadcast in numpy. Also note that the cost function is the log likelihood and not the mean square error function here. I assumed the first column to be X_2 and the second column as X_1.

3.1 a

Number of iterations: 8
Final Cost: − 0.22834144984472393
Theta: (θ_2, θ_1, θ_0) = (2.5885477, −2.72558849, 0.40125316)
Stopping Criteria: $|\mathrm{J}(\theta_t) - \mathrm{J}(\theta_{t-1})) \leq 10^{-15}$

3.2 b

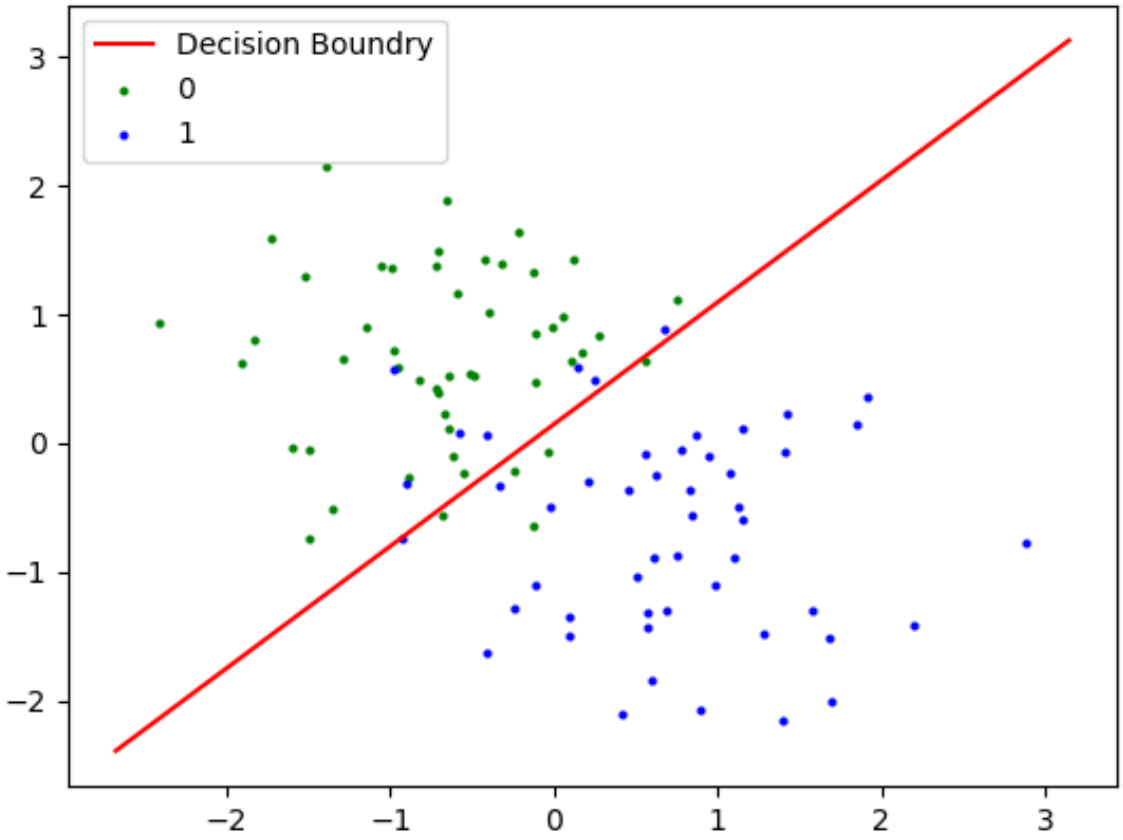

Figure 3.1: Decision Boundary and Data Plot

Chapter 4

Gaussian Discriminant Analysis

Assume Canada as class - 1 and Alaska as class - 0.

4.1 a

$\phi = 0.5$

$$\mu_0 = \begin{bmatrix} -0.75529433 \\ 0.68509431 \end{bmatrix}$$

$$\mu_1 = \begin{bmatrix} 0.75529433 \\ -0.68509431 \end{bmatrix}$$

$$\sum = \begin{bmatrix} 0.42953048 & -0.02247228 \\ -0.02247228 & 0.53064579 \end{bmatrix}$$

4.2 b

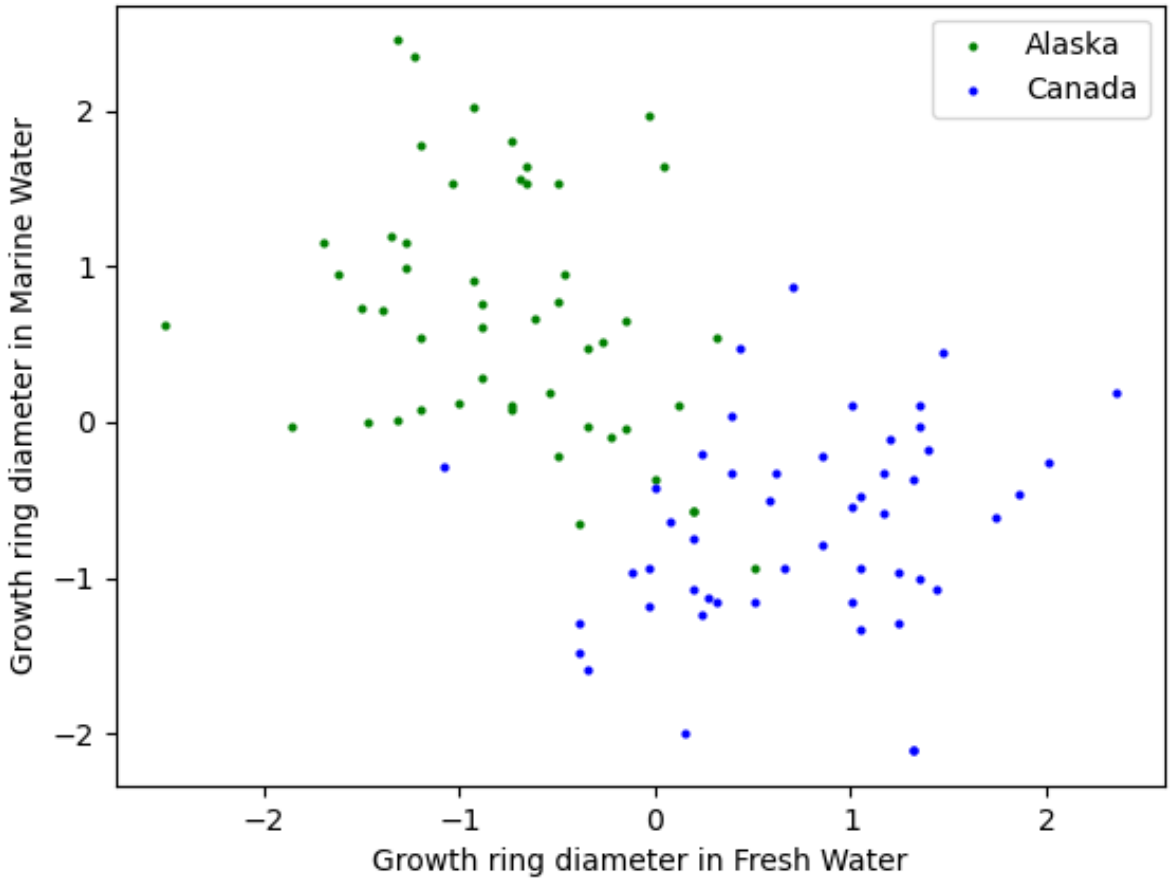

Figure 4.1: Data Plot

4.3 c

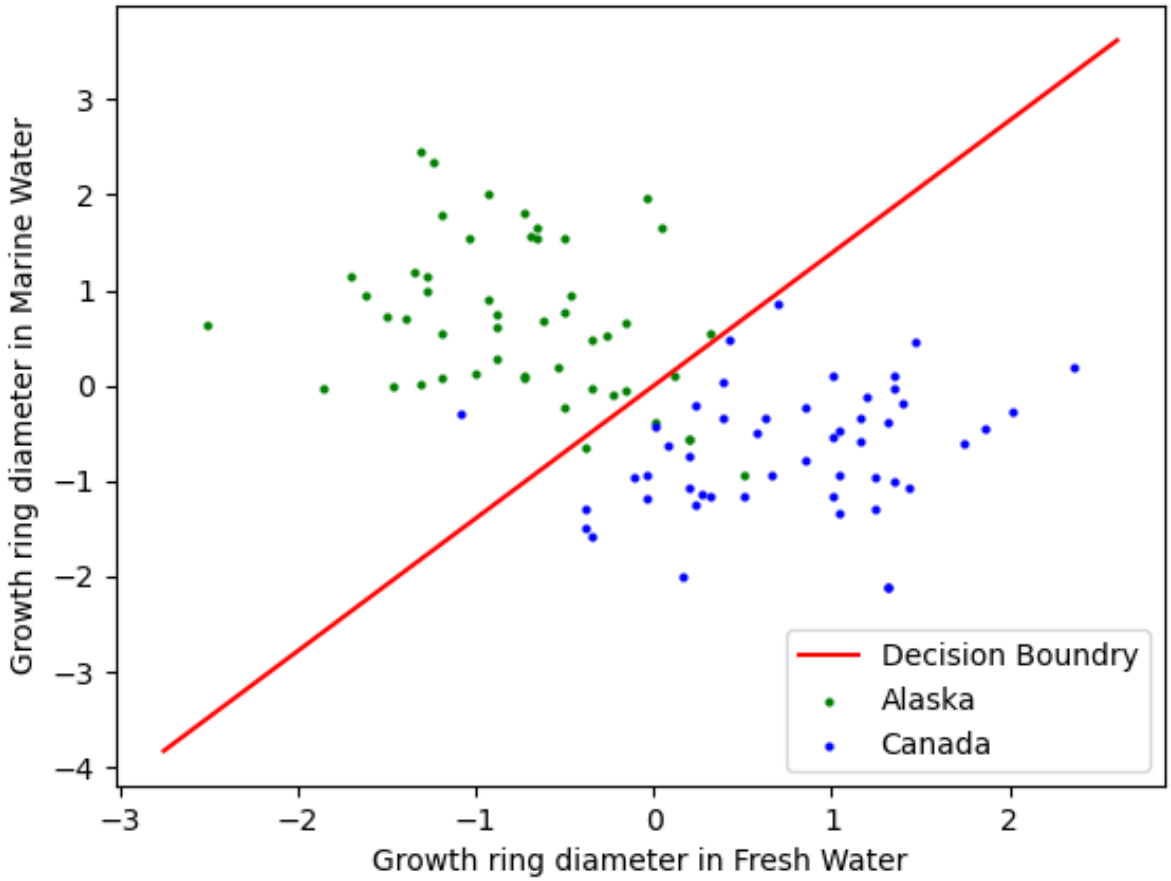

Figure 4.2: Linear Decision Boundary and Data Plot

4.4 d

$\phi = 0.5$

$$\mu_0 = \begin{bmatrix} -0.75529433 \\ 0.68509431 \end{bmatrix}$$

$$\mu_1 = \begin{bmatrix} 0.75529433 \\ -0.68509431 \end{bmatrix}$$

$$\sum_0 = \begin{bmatrix} 0.38158978 & -0.15486516 \\ -0.15486516 & 0.64773717 \end{bmatrix}$$

$$\sum_1 = \begin{bmatrix} 0.47747117 & 0.1099206 \\ 0.1099206 & 0.41355441 \end{bmatrix}$$

4.5 e

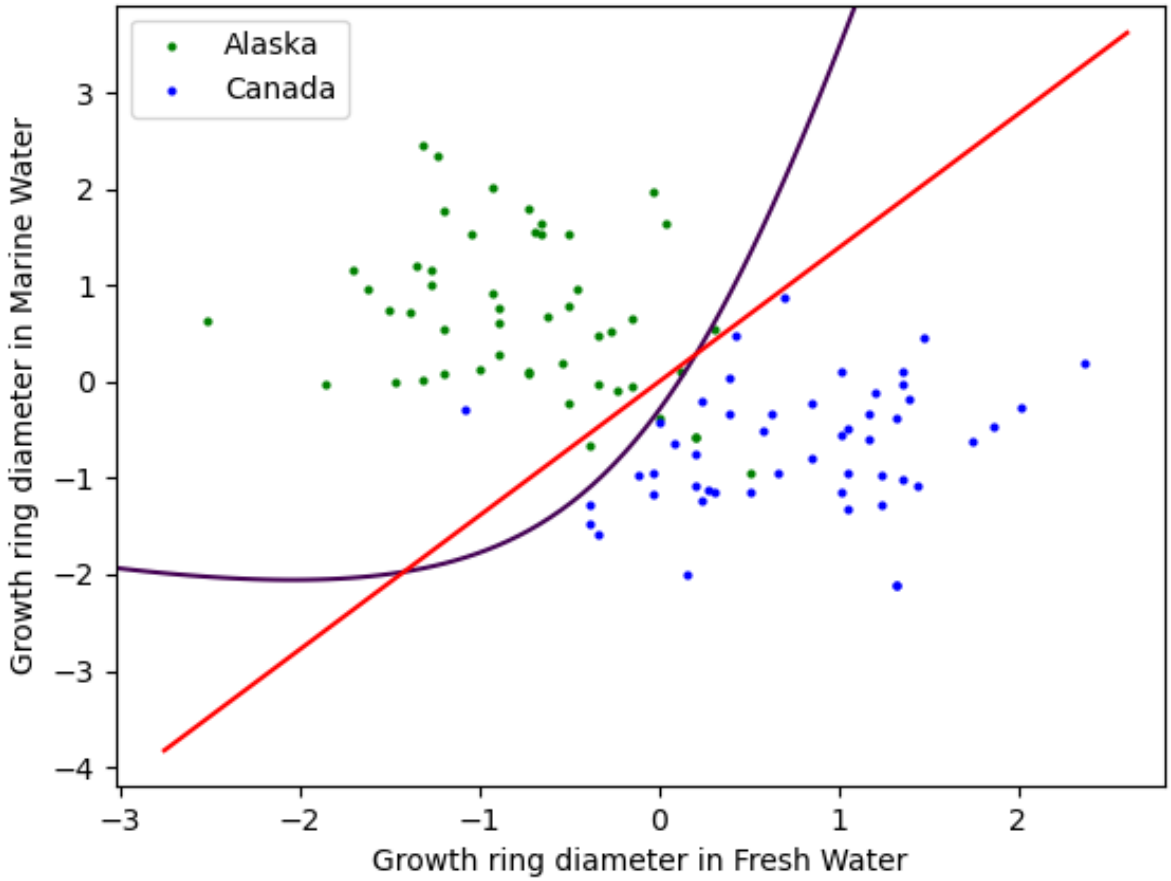

Figure 4.3: Linear and Quadratic Boundary Plot

4.6 f

1. We notice that in Figure 4.3 a few samples mislabelled by the linear classifier as Canada are correctly recognized by the quadratic classifier.
2. Both the classifiers seem to work very well on the small training data. The quadratic curve improved the classification but it's likely the curved boundary has overfitted the training data but we will need more data to confirm this.