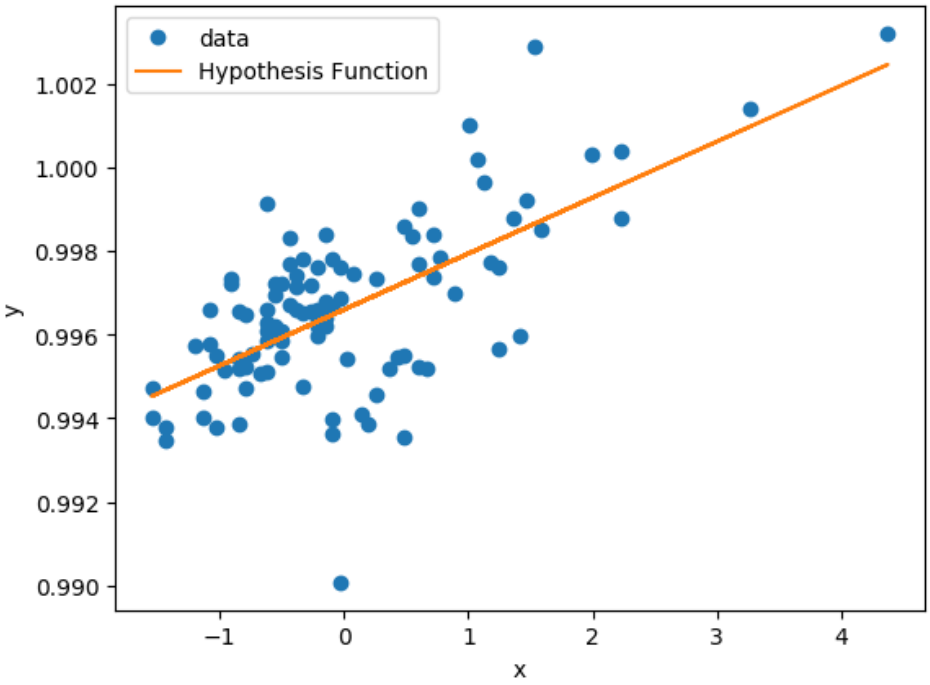
COL 774: Machine Learning. Assignment 1

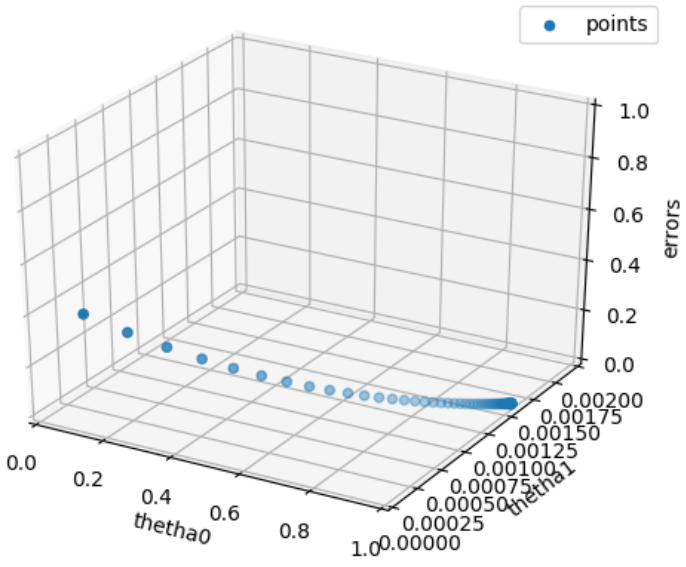
Q1. Linear Regression:

(a) The learning rate is 0.1 and the convergence condition here was that the difference between the log likelihood in the consecutive steps should be less than 10^{-10} .
It converged in 100 steps.
θ=[0.9965936283759845 0.0013401604210299 45]

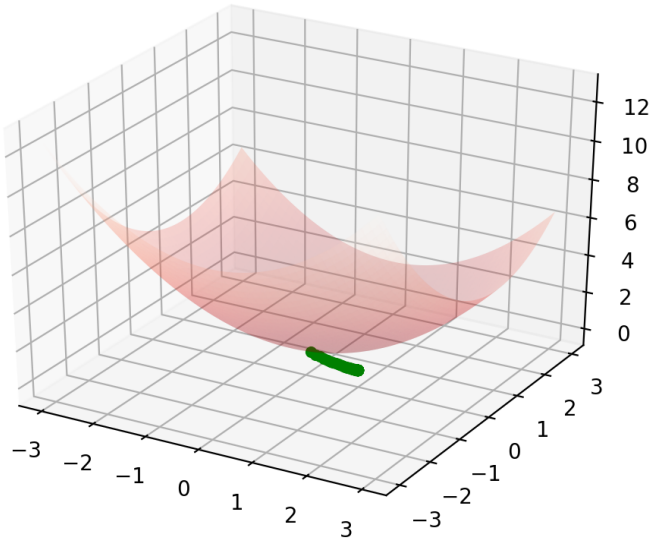
(b)



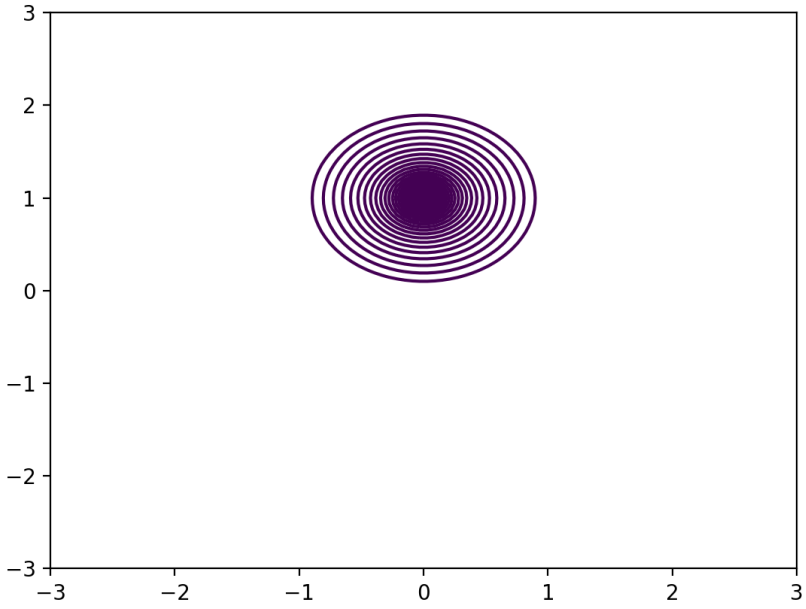
(c) error vs theta:



3-dimensional mesh :

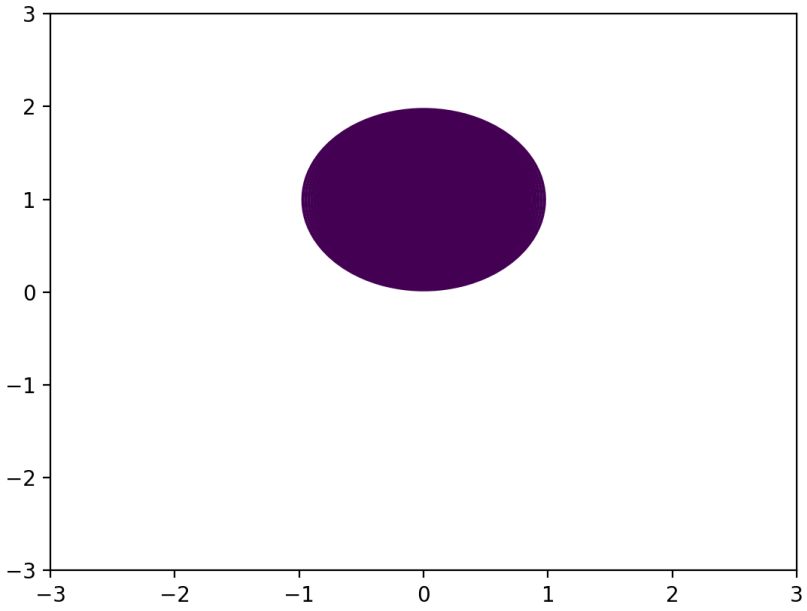


(d) Contours of the error function:

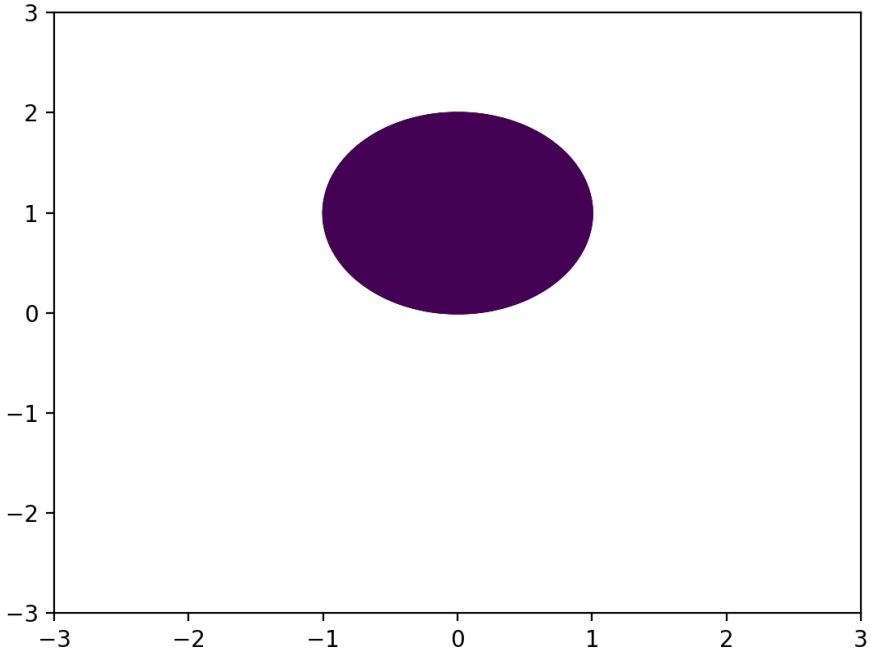


(e) The number of iterations were larger in η = 0.025(383 steps) than 0.1(100 steps) and even much larger in 0.001(8053 steps). Hence, due to larger number of iterations, the contours were closer for these.

For η = 0.025:



For η = 0.001:



Q2. Sampling and Stochastic Gradient Descent:

(b) For r=1: θ = [2.97450607 1.00845608 1.95809396]

The convergence criterion is 20000 iterations.

For r=100: θ = [2.9961305 0.9960019 1.99807078]

The convergence criterion is 5 epoch.

For r=10000: θ = [2.90907832 1.02020454 1.99349441]

The convergence criterion is 250 epoch.

For r=1000000: The convergence didn't occur in some reasonably long time

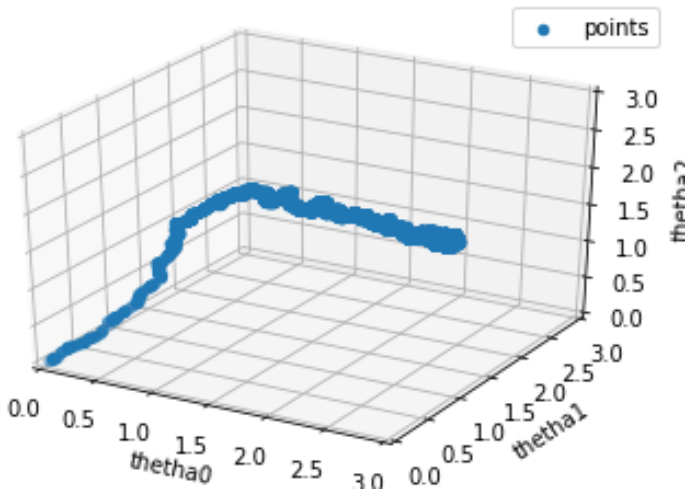
(c) For r=1: error now is 1.0809882347696573, earlier it was 0.6510995532136361

For r=100: error now is 0.984489183816584, earlier it was 0.8027950692626882

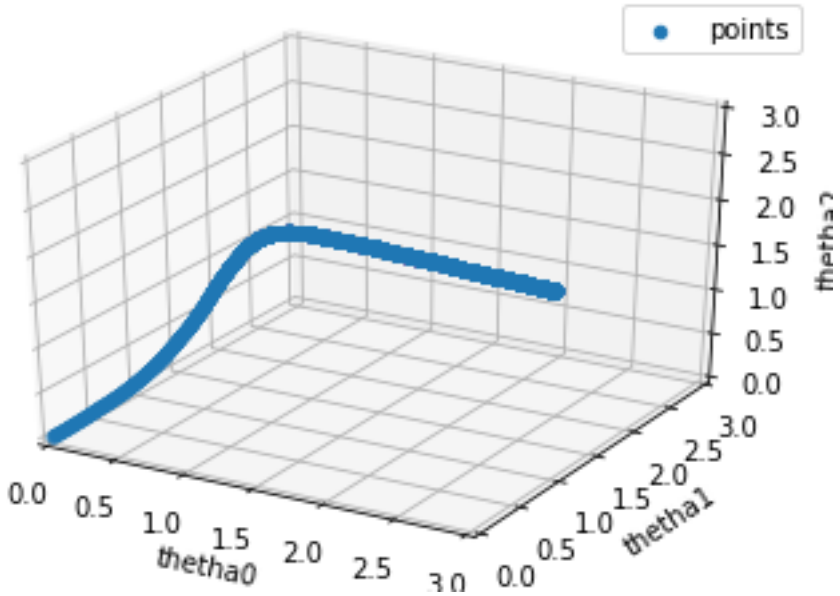
For r=10000: error now is 1.007566055306812, earlier it was 1.0026676297186212

(d) As we increase the value of the batch size(r), we see less deviation in the path to the convergence point, i.e the curve becomes smoother, This is because in very small batch size, the average error we'll get is more at each step. In a larger batch size, the error is normalised over more examples and hence we get a smooth path from origin to the convergence point.

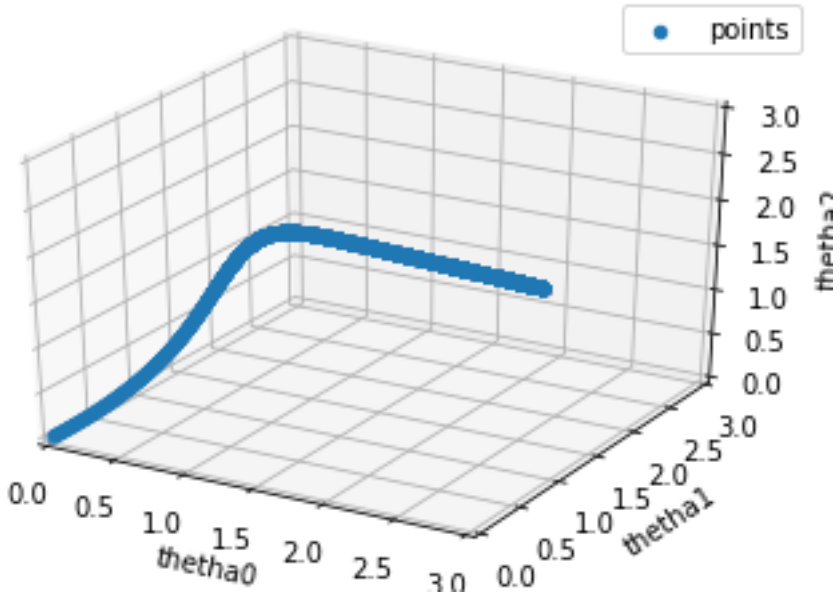
r=1:



r=100:



r=10000: (every 100th value sampled)



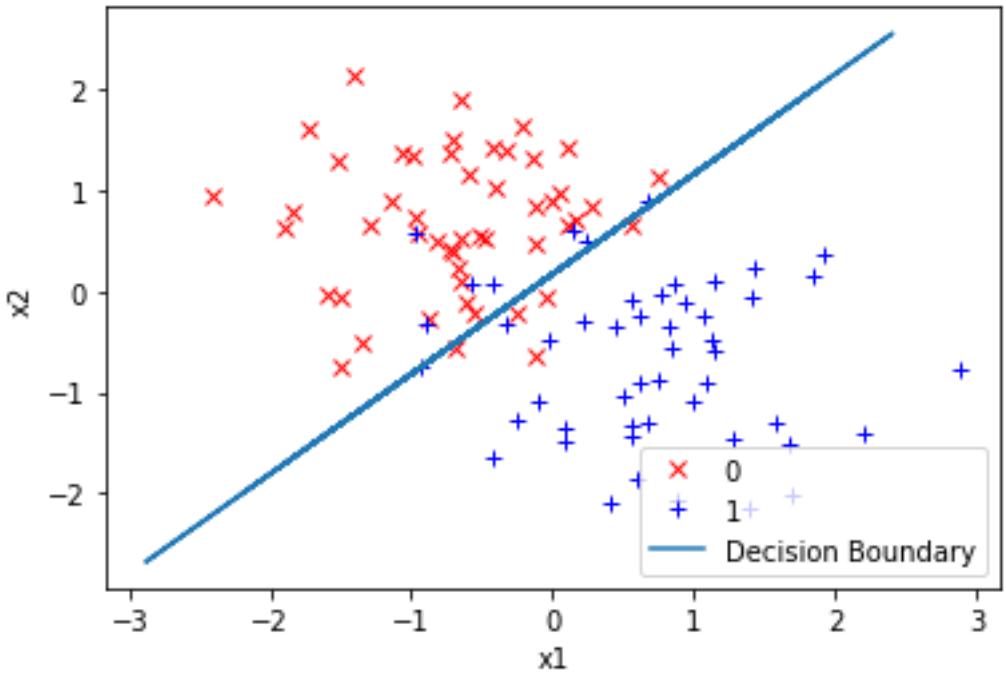
Q3. Logistic Regression:

(a) θ =[-0.11982439, -0.70571821, 0.69828642]

Hypothesis function = $x^T\theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

The convergence condition here was that the difference between the log likelihood in the consecutive steps should be less than 10^{-6}. It converged in 62 steps.

(b)



Q4. Gaussian Discrmimant Analysis:

(a),(d) μ_0 : [-0.75529433 0.68509431]

μ_1 : [0.75529433 -0.68509431]

Σ_0 : [[0.38158978 -0.15486516]

[-0.15486516 0.64773717]]

Σ_0 : [[0.47747117 0.1099206]

[0.1099206 0.41355441]]

Σ : [[0.42953048 -0.02247228]

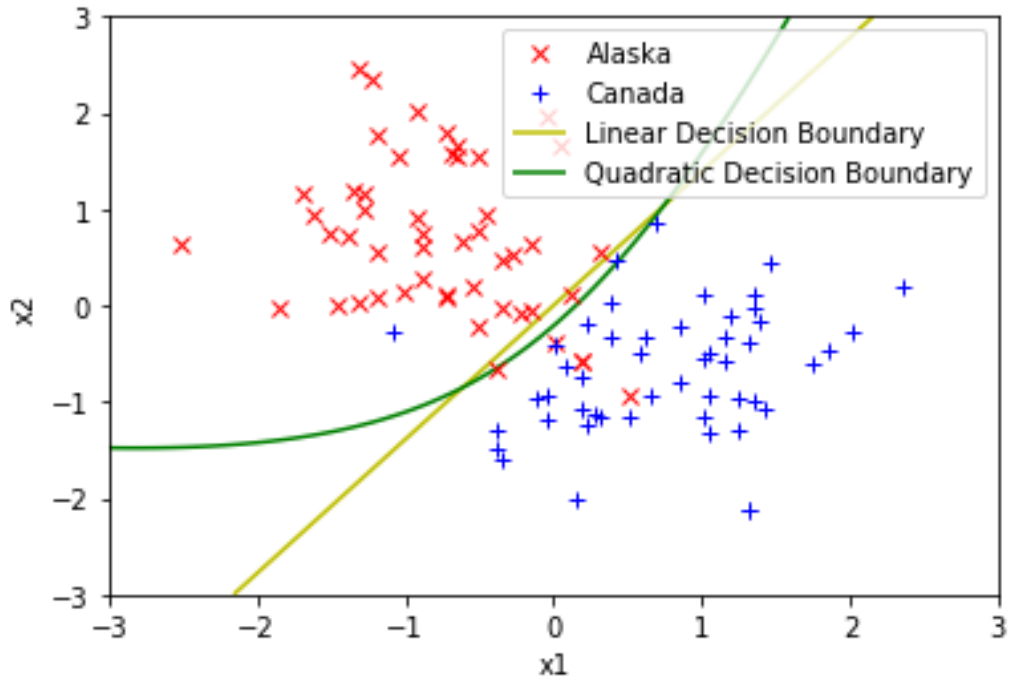
[-0.02247228 0.53064579]]

(c) The equation foe the linear boundary:

$$(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)-(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)=2\log\varphi-2\log(1-\varphi)$$

(e)The equation for the quadratic boundary:

$$(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)-(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)=\log|\Sigma_0|-\log|\Sigma_0|+2\log\varphi-2\log(1-\varphi)$$



(f) Linear Decision Boundary almost passes through the origin in the normalised data. The quadratic boundary is a hyperbola.
Both the boundaries are good classifiers. However, the quadratic boundary is more accurate by a few examples, and hence we can conclude that the underlying distribution is gaussian.