ANKUSH PHULIA
2014CS50279

Assignment 1 – Machine Learning

LINEAR REGRESSION

BATCH GRADIENT DESCENT

Learning Rate α : 0.02

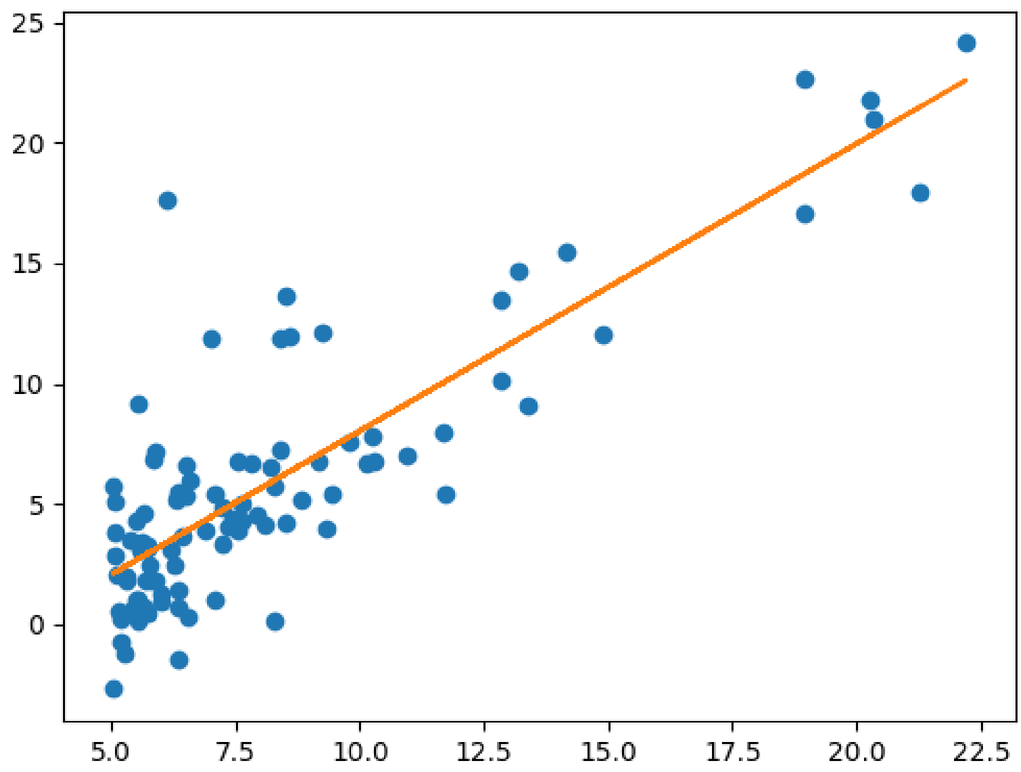
Stopping Criteria : The difference between error function $J(\theta)$ of previous and current iteration becomes less than ϵ

Stopping Parameter ϵ : 0.0000000001

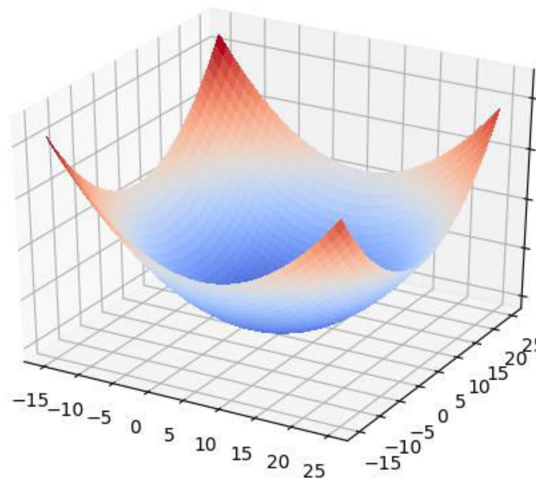
Final Coefficients θ : <-3.89565, 1.19302>

Final Coefficients θ (after normalization) : <5.839118, 4.593028>

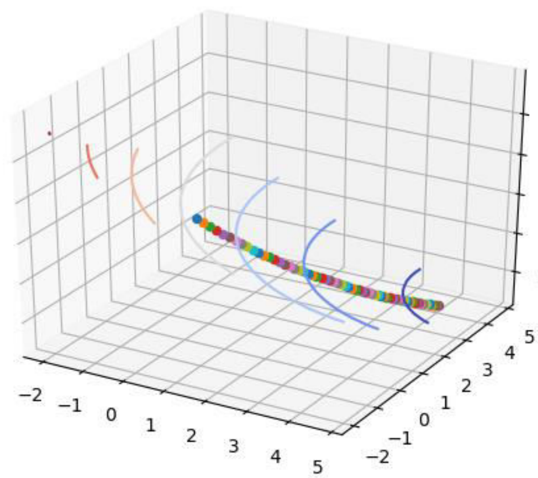
GRAPH



Red Line indicates the hypothesis function learnt, and blue points are data points

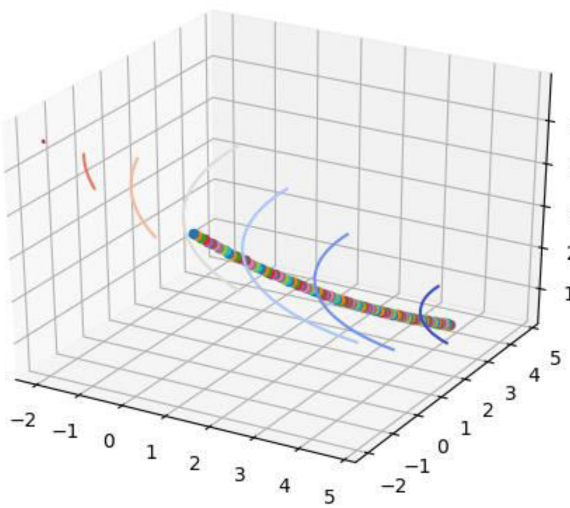


Mesh of J (θ), α = 0.02

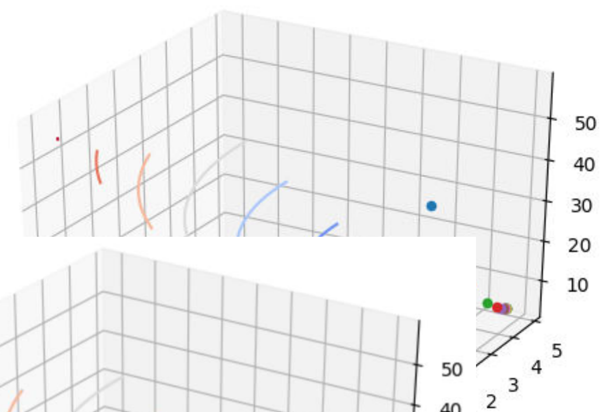


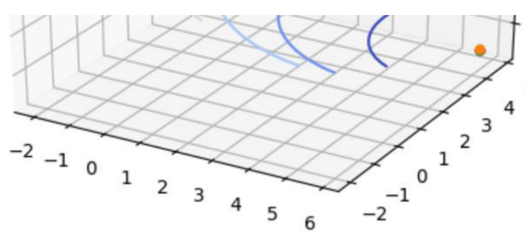
Contours α = 0.02

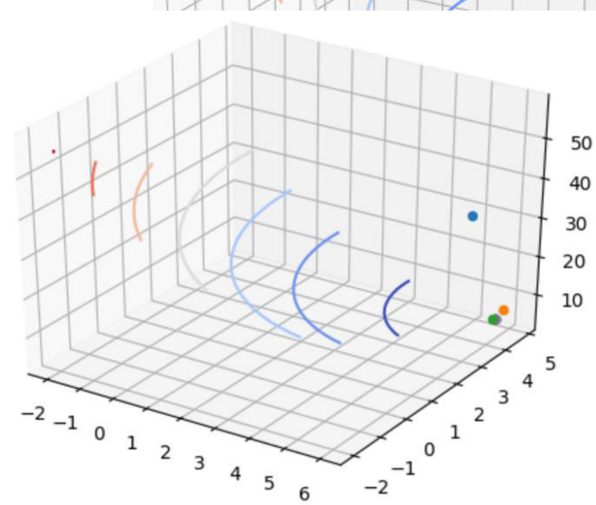
α = 0.01



α = 0.05



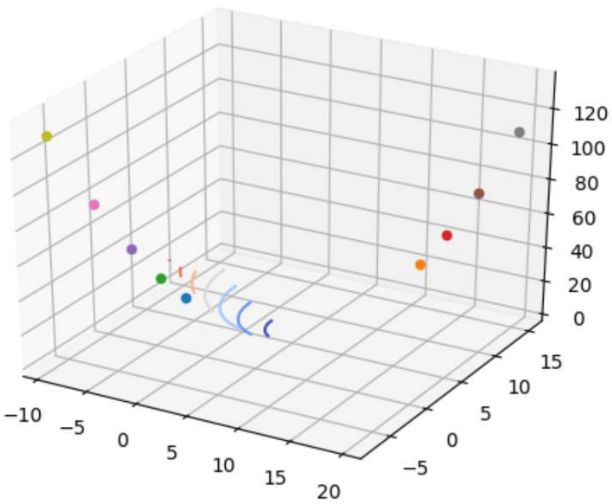




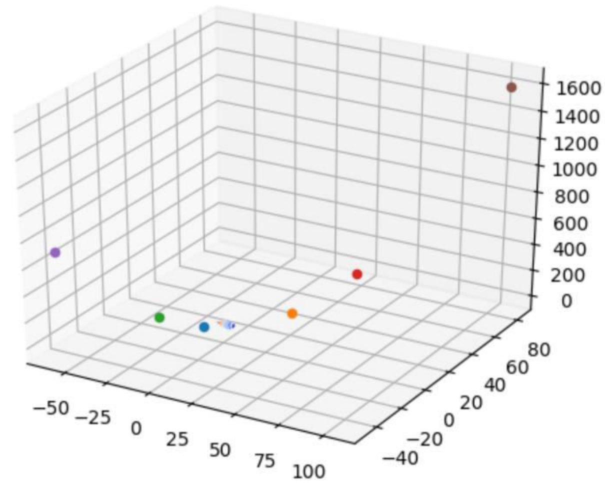
α = 0.9

α = 1.3

Note: Left axis is for θ_0 and Right is for θ_1



α = 2.1



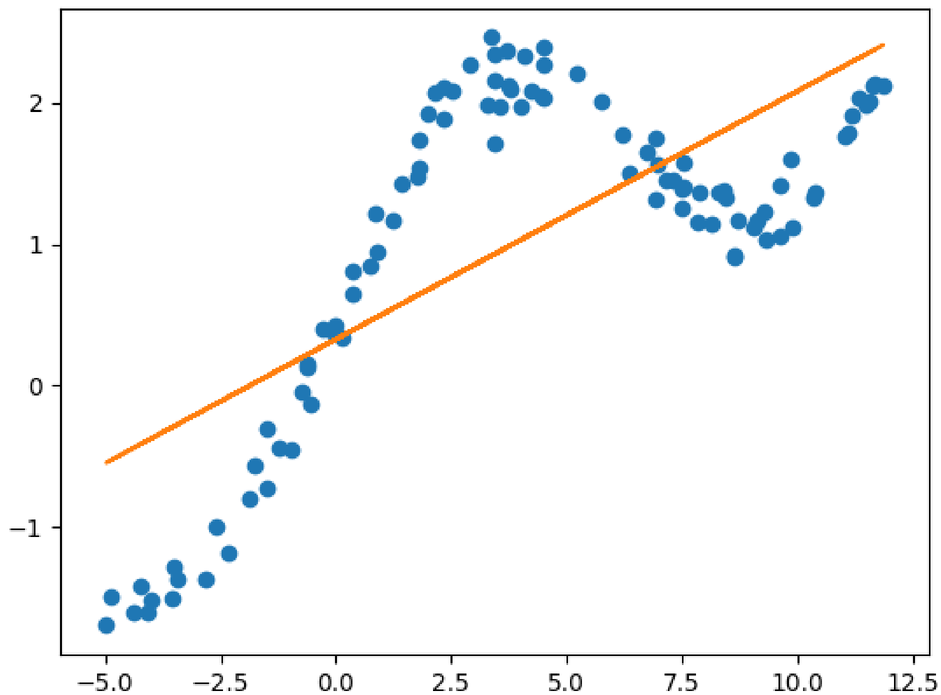
α = 2.5

Note: Left axis is for θ_0 and Right is for θ_1

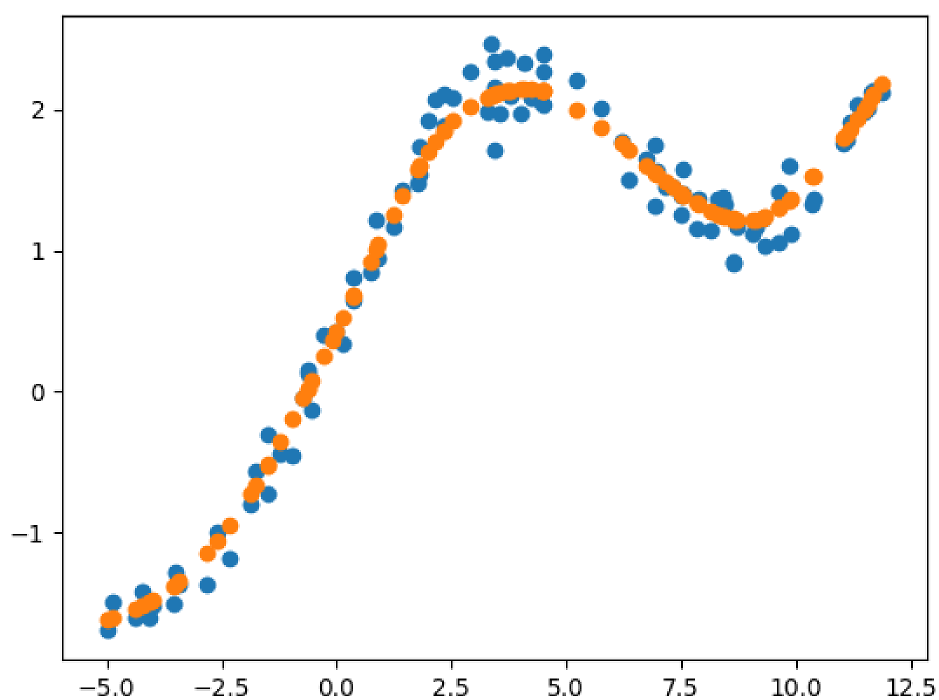
The graphs for different α are as above. For small values of α, the algorithm converges to the minimum, the rate of which increases with α. However, for α = 1.3 the algorithm had overshot the minimum, ending up on the other side, but this overshoot was small enough that the algorithm still converged. However, for α = 2.1 and 2.5, the algorithm diverges as the overshoot was so large that the algorithm actually ended up at a location with higher cost than before every time.

LOCALLY WEIGHTED LINEAR REGRESSION

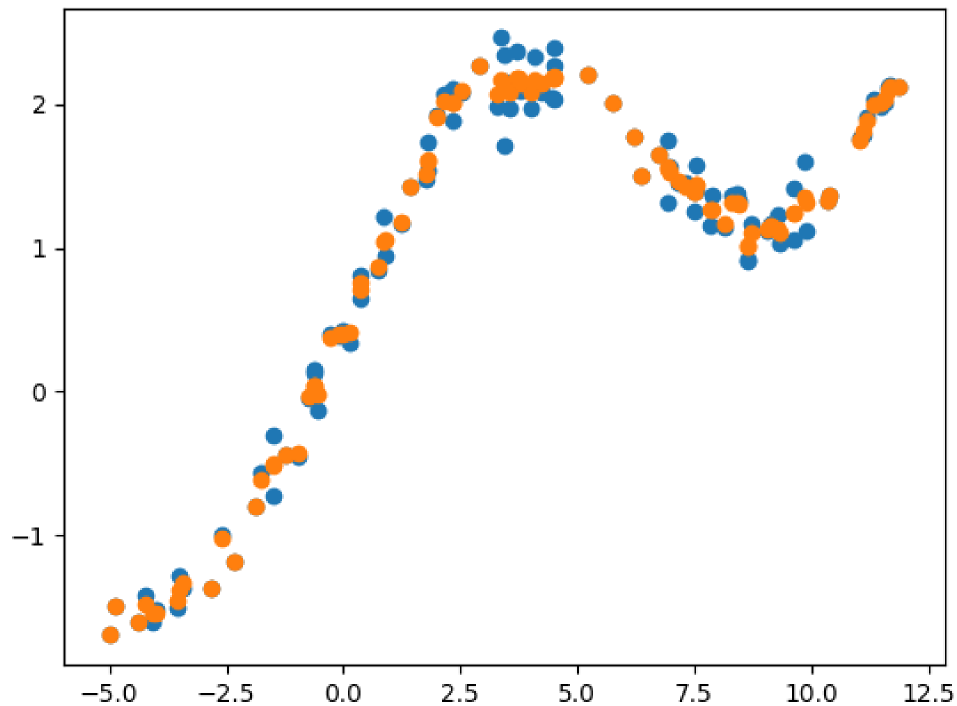
UNWEIGHTED LINEAR REGRESSION



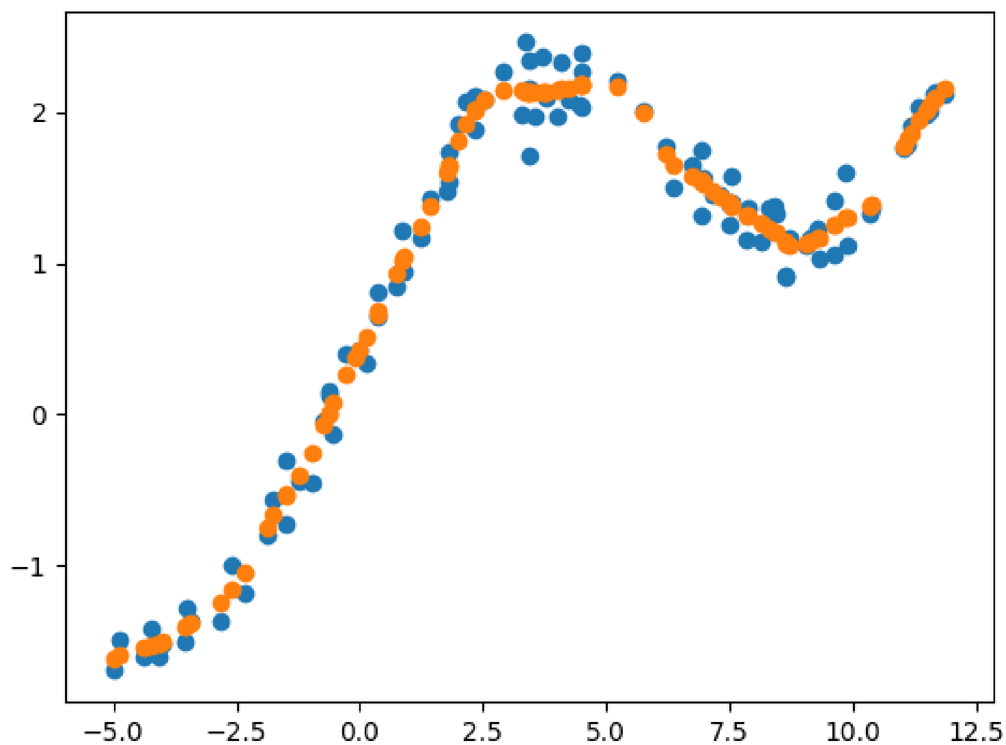
WEIGHTED LINEAR REGRESSION



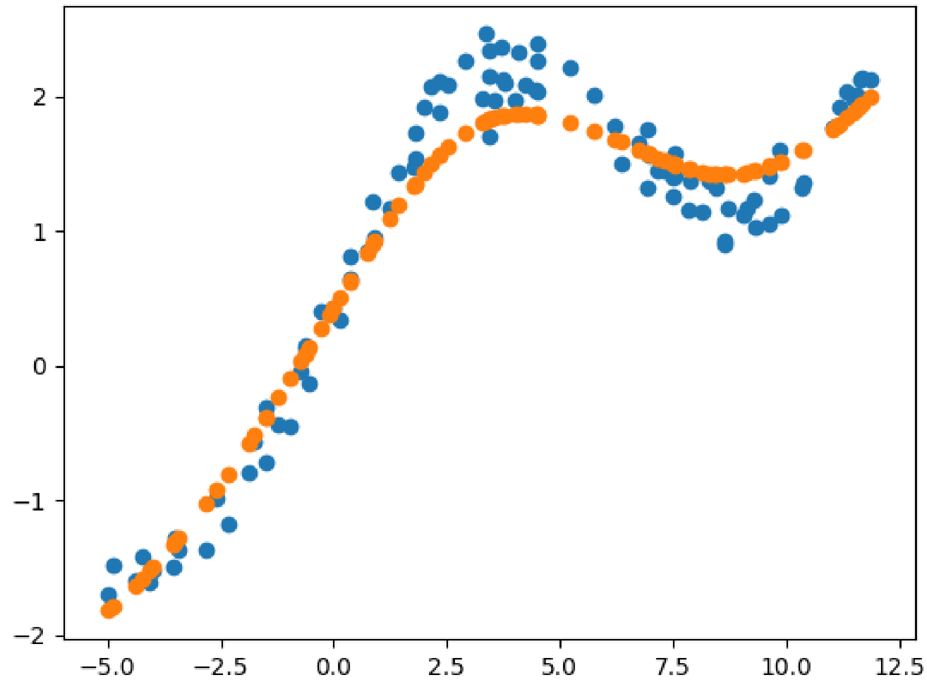
For τ = 0.8, the hypothesis plots the red points (the blue points are training data points)



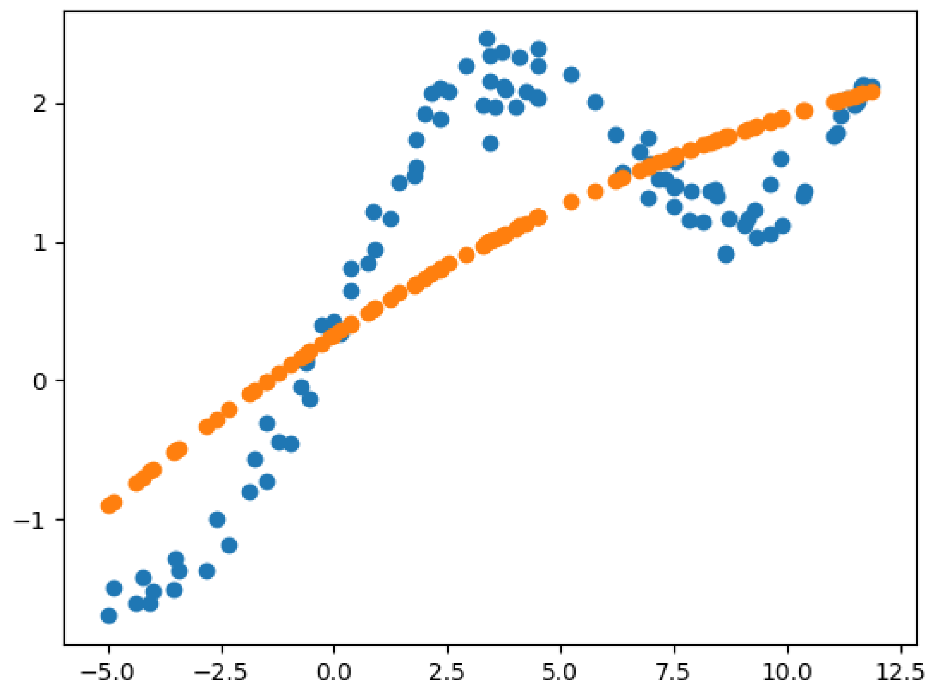
τ = 0.1



τ = 0.3



τ = 2



τ = 10

As τ increases, the points closer to the test point carry weight that isn't too much higher than those far away, whereas on decreasing τ, points closer affect the fitting of the curve more strongly. Thus for very high τ we will tend to under-fit, or as τ tends to infinity, weights tend to one and the resulting curve becomes a line. As τ tends to zero, closest points have very high weight and can cause overfitting of learned hypothesis to training data

τ = 0.3 seems to fit the data best

LOGISTIC REGRESSION

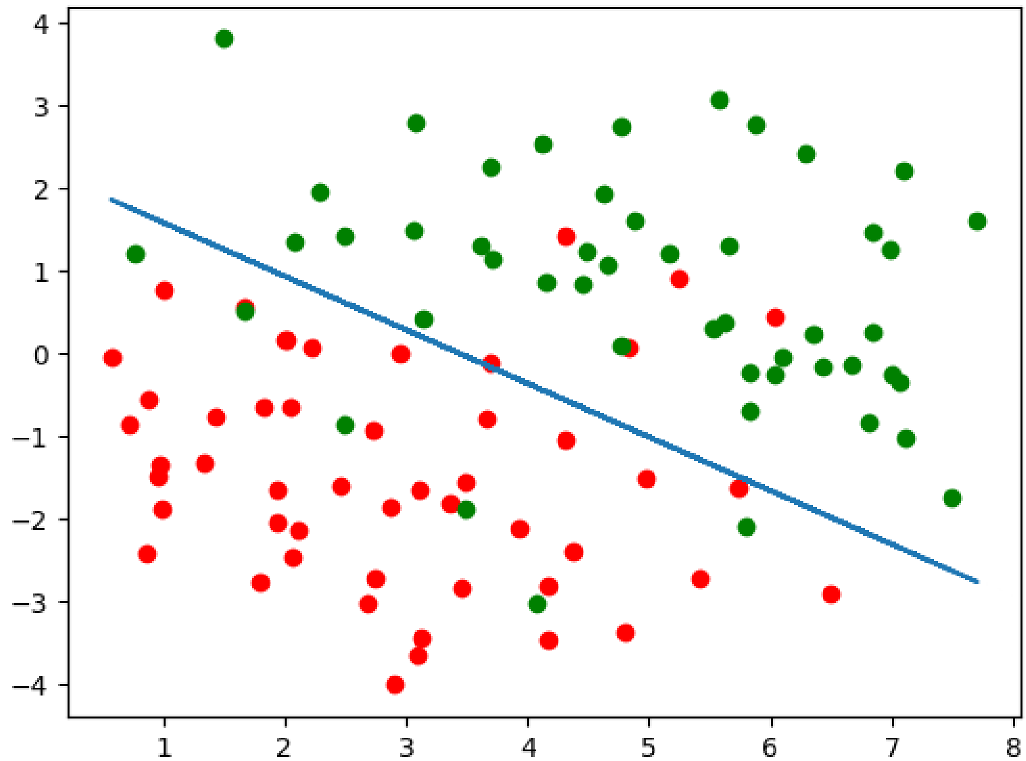
NEWTON-RAPHSON METHOD

Stopping Criteria : The difference between error change in θ, summed, of previous and current iteration becomes less than ϵ

Stopping Parameter ϵ : 0.0000000001

Final Coefficients θ : <-2.62049, 0.76036, 1.17194>

GRAPH



Red Points are those with y = 0, Green have y = 1 and the Blue line is the decision boundary

Note: X axis is for X_1 and Y is for X_2

PARAMETERS (SAME COVARIANCE)

$\Phi = 0.5$

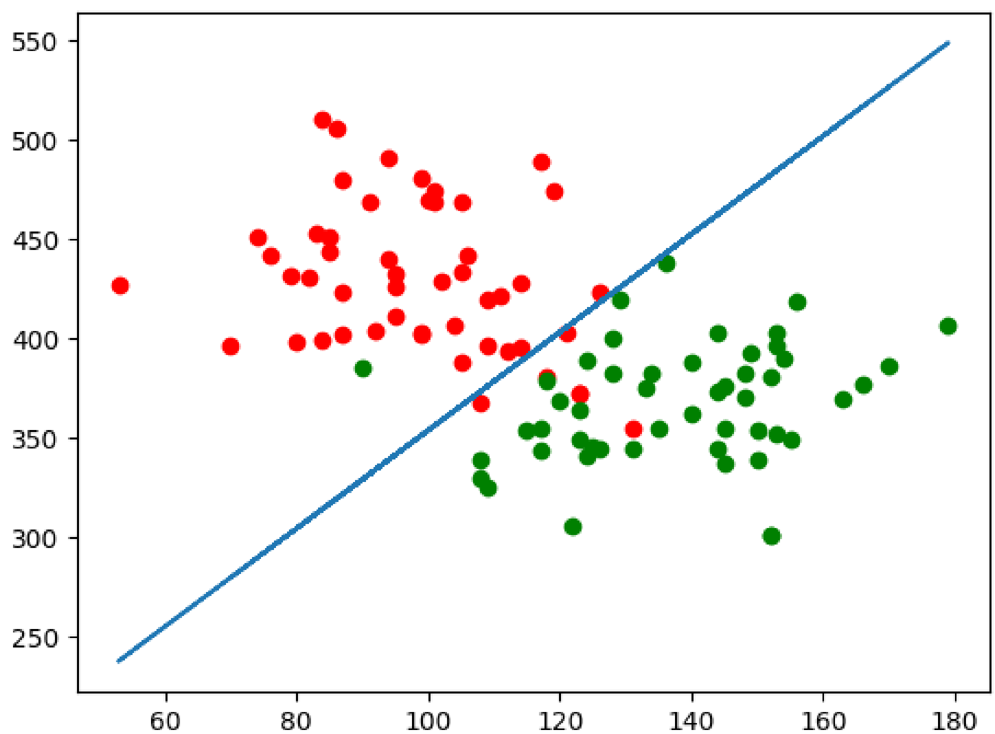
μ_0 = <98.38, 429.66>

μ_1 = <137.46, 366.62>

Σ = <<287.482, -26.748>

<-26.748, 1123.25> >

GRAPH (SAME COVARIANCE)



Note: X axis is for X_1(Fresh Water) and Y is for X_2(Marine Water)

Green is for Alaska and red is for Canada, and blue line is the linear decision boundary misclassifying around 5 red and 1 green point

The decision boundary is given by the equation : (represented by x)

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

PARAMETERS (DIFFERENT COVARIANCE)

Φ = 0.5

μ_0 = <98.38, 429.66>

μ_1 = <137.46, 366.62>

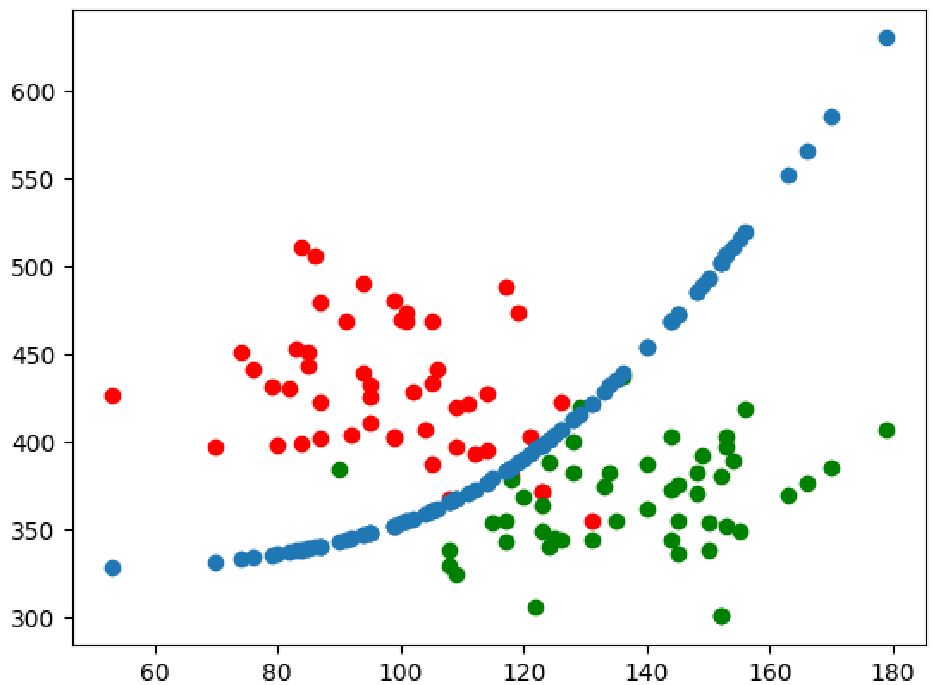
Σ_0 = <<255.395, -184.330>

<-184.330, 1371.1044> >

Σ_1 = <<319.568, -130.834>

<-130.834, 875.395> >

GRAPH (SAME COVARIANCE)



Note: X axis is for X_1(Fresh Water) and Y is for X_2(Marine Water)

Green is for Alaska and red is for Canada, and blue set of points is the quadratic decision boundary misclassifying 2 red and 2 blue points

The decision boundary is given by the equation : (represented by x)

$$x^T(\Sigma_0^{-1} - \Sigma_1^{-1})x + (2(\mu_1^T\Sigma_1^{-1} - \mu_0^T\Sigma_0^{-1}))x + (\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) + \log\frac{||\Sigma_0||^{0.5}}{||\Sigma_1||^{0.5}} = 0$$

In this general equation if both covariance matrices are the same, then the first term would vanish, which leads to the quadratic term in the boundary. This decision boundary is actually a hyperbola, observed if the graph is large enough. The difference in the boundaries leads to a different classification of data points very close to the boundary, and the quadratic boundary seems to partition the data better (4 vs 6 misclassifications)