Assignment Report 1 - Machine Learning COL774

Piyush Kaul - 2015EEY7544

February 5, 2016

1 ANS-1 Linear Regression

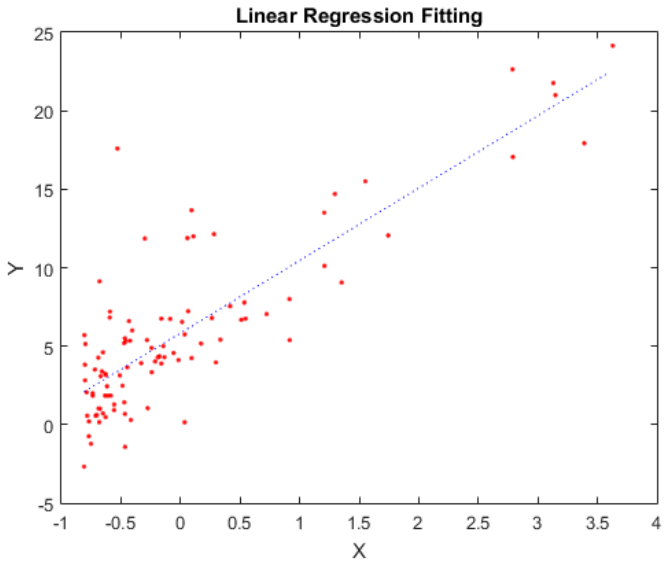
1.1 PART A

Parameters

- Learning rate = 0.1
- Stopping Criteria : $\sum_{j=1}^{n}(\theta_j(t+1)-\theta_j(t))^2 \leq 0.0001$
- θ = [4.616310 5.838457]

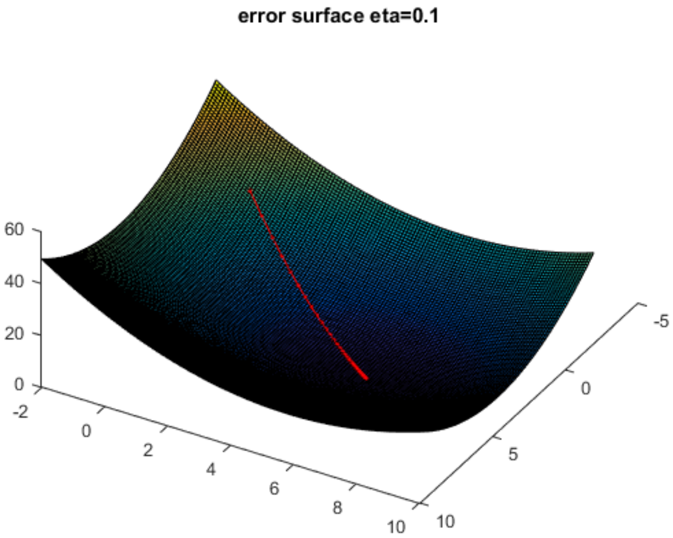
1.2 PART B

Figure 1.1: Linear Regression



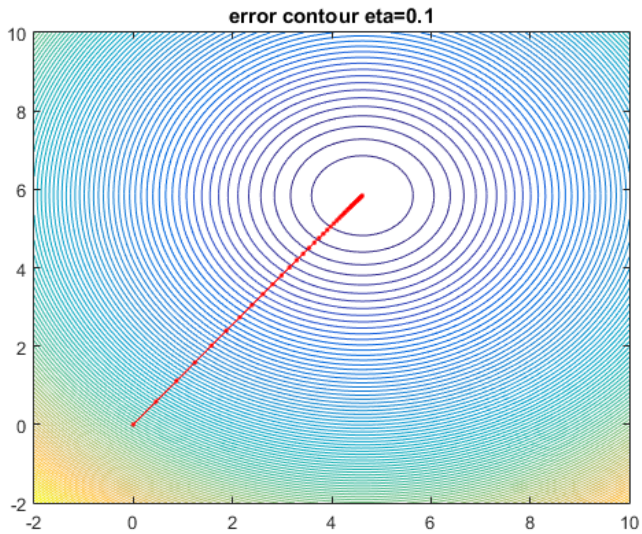
1.3 PART C

Figure 1.2: Error Surface with Eta=0.1



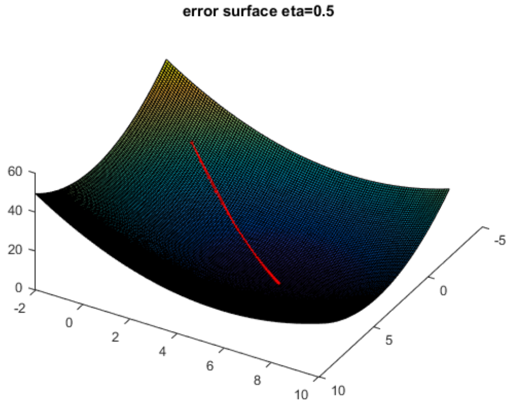
1.4 PART D

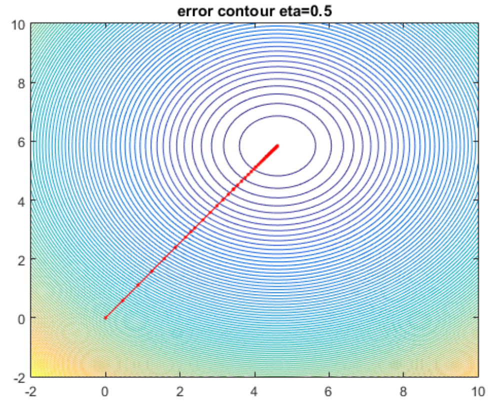
Figure 1.3: Error Contour with Eta=0.1

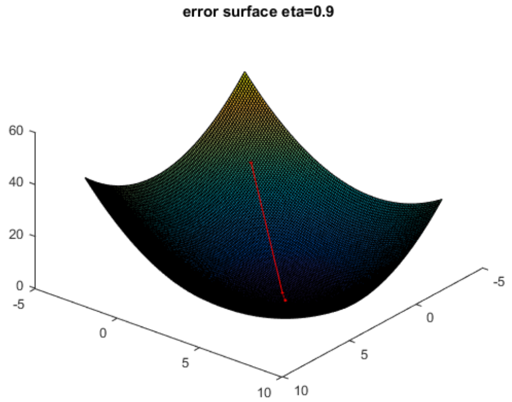


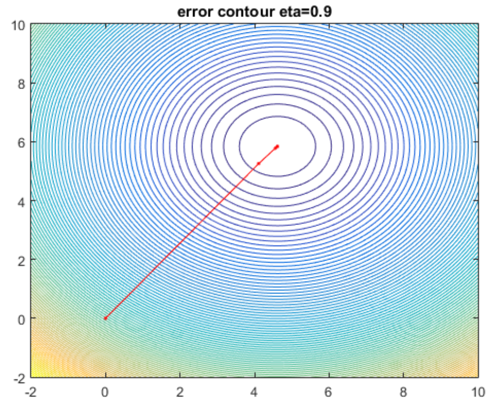
1.5 PART E

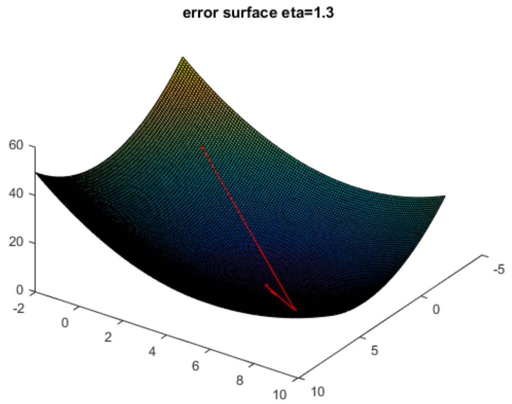
As visible in the below graphs, the number of iterations for convergence reduces with increased η till 0.9. For $\eta = 1.3$ gradient descent starts to overshoot the minimum but still returns towards the minimum. However for $\eta >= 2.1$ the gradient descent starts to diverge.

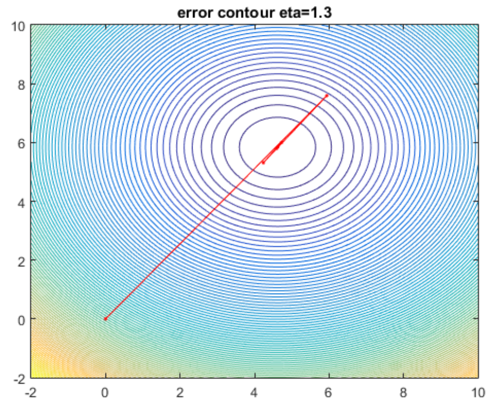

error surface eta=0.5
60
40
20
0
-2
0
2
4
6
8
10
10
5
0
-5

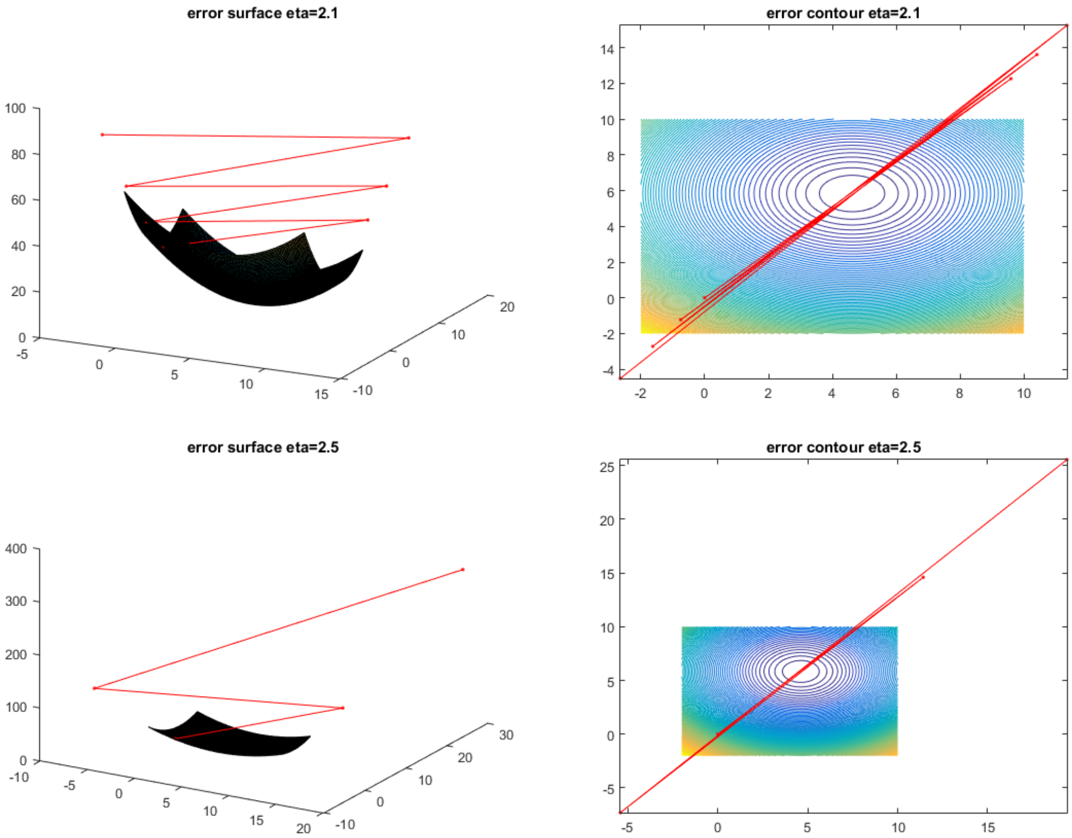

error contour eta=0.5
10
8
6
4
2
0
-2
-2
0
2
4
6
8
10


error surface eta=0.9
60
40
20
0
-5
0
5
10
10
5
0
-5


error contour eta=0.9
10
8
6
4
2
0
-2
-2
0
2
4
6
8
10


error surface eta=1.3
60
40
20
0
-2
0
2
4
6
8
10
10
5
0
-5


error contour eta=1.3
10
8
6
4
2
0
-2
-2
0
2
4
6
8
10

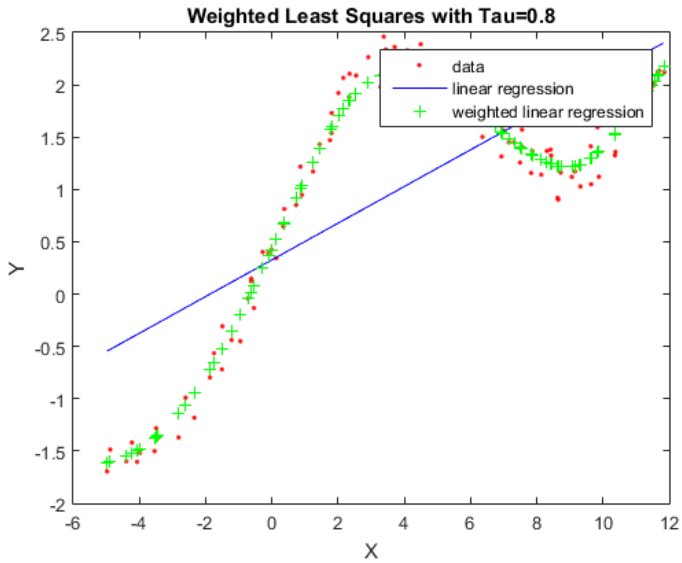


2 ANS-2 Locally Weighted Linear Regression

$$
\begin{aligned}
&J(\theta) = 1/2(X\Theta - Y)^2 W(X\Theta - Y)\\
&\Delta_\Theta J(\theta) = \Delta_\Theta tr(\Theta^T X^T W X\Theta - \Theta^T X^T W Y - Y^T W X\Theta + Y^T W Y)\\
&\Delta_\Theta J(\theta) = \Delta_\Theta tr(\Theta^T X^T W X\Theta - 2Y^T W X\Theta\\
&Since\\
&\Delta_A tr ABA^T C = CAB + C^T AB^T\\
&\Delta_{A^T} tr ABA^T C = B^T A^T C^T + BA^T C\\
&Hence\\
&\Delta_\Theta J(\theta) = (X^T W X)^T\Theta + (X^T W X)\Theta - 2(Y^T W X\Theta)\\
&= 2X^T W X\Theta - 2X^T W Y\\
&Equating\\
&\Delta_\Theta J(\theta) = 0\\
&\Theta = (X^T W X)^{-1} X^T W Y
\end{aligned}
\tag{2.1}
$$

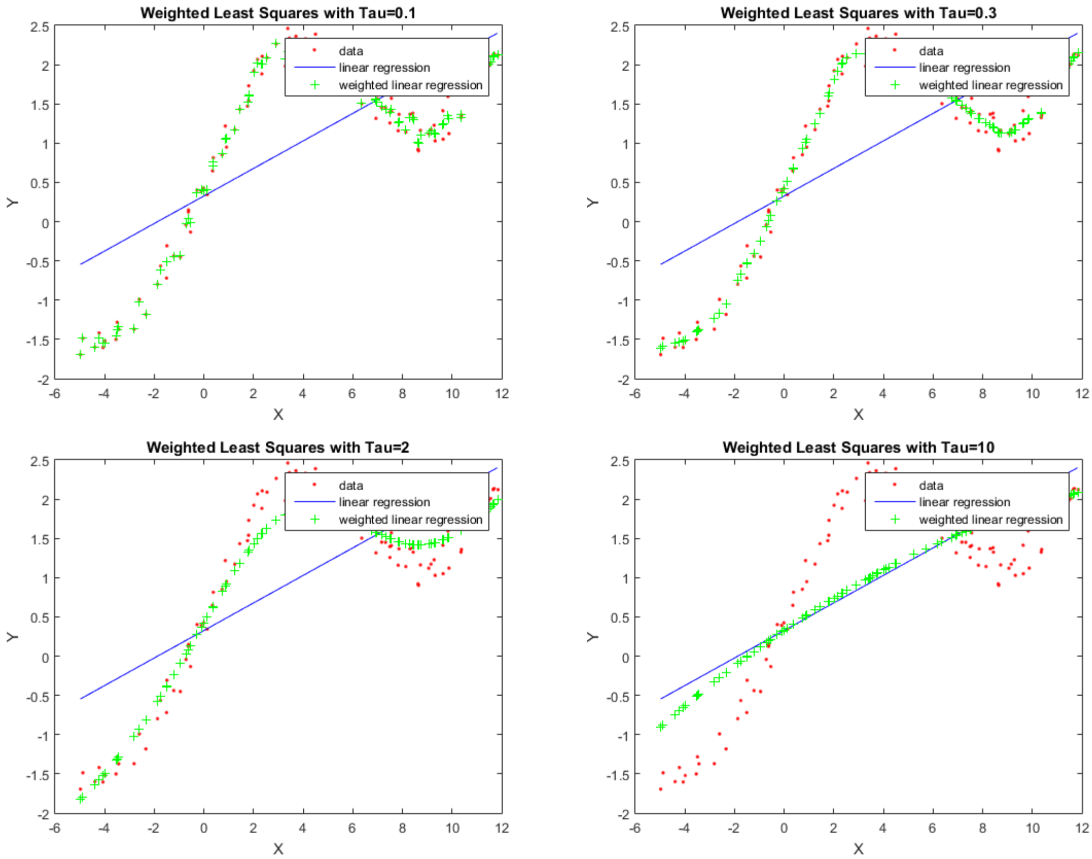
2.1 PART A & B

Figure 2.1: with tau = 0.8



2.2 PART C

As visible from the graph plots below, we get smooth fit to the curve for Tau=0.8. For very low Tau=0.1, we start to fit the noise in the the datapoints and smoothness is lost. For very high Tau=10, Locally Weighted Least Squares becomes same as standard Least Squares and we start fitting a line to the curve.



3 ANS3 - Logistic Regression

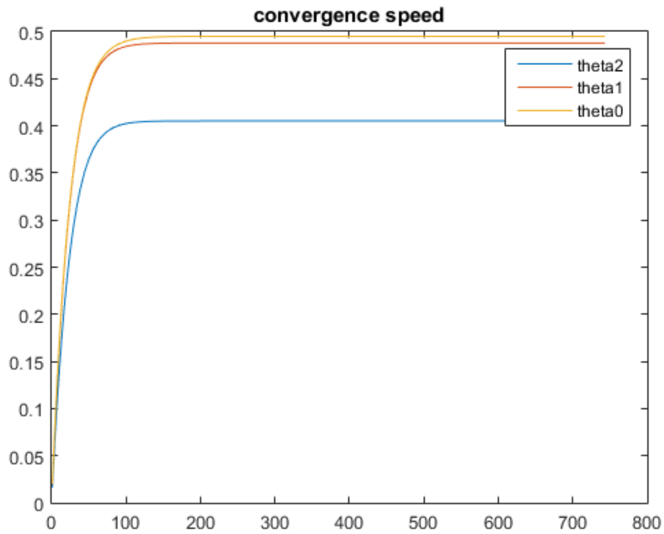
Calculation of Hessian can be done as below

$$
\begin{aligned}
LL(\theta) &= \sum_{i=1}^{m} y^{(i)} log h(x^{(i)}) + (1-y^{(i)}) log(1-h(x^{(i)}) \\
\frac{\partial}{\partial\theta_k\partial\theta_j} &= \frac{\partial}{\partial\theta_k}\sum_{i=1}^{m} y^{(i)}(1-g(\theta^T x^{(i)}) + (1-y^{(i)})(g(\theta^T x^{(i)}) x_j^{(i)} \\
&= -\frac{\partial}{\partial\theta_k}\sum_{i=1}^{m} y^{(i)}x_j^{(i)} - y^{(i)}g(\theta^T x(i))x_j^{(i)} - g(\theta^T x(i)x_j^{(i)} + y^{(i)}g(\theta^T x(i)x_j^{(i)} \\
&= -\sum_{i=1}^{m} g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))x_k^{(i)}x_j^{(i)}
\end{aligned} \tag{3.1}
$$

3.1 Part A

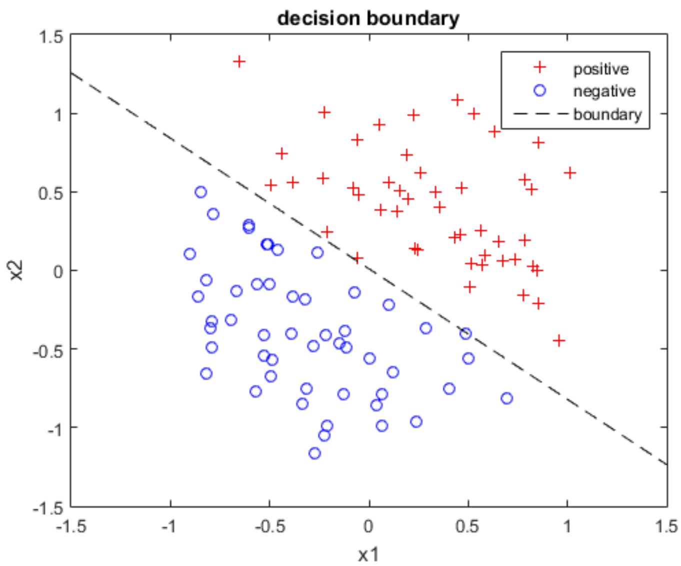
$$
\theta = \begin{bmatrix} 0.405158 & 0.487819 & 0.494949 \end{bmatrix} \tag{3.2}
$$

Figure 3.1: Convergence.



3.2 PART B

Figure 3.2: Decision Boundary.



4 ANS4 - Gaussian Discriminant Analysis

4.1 Part A

Covariance Matrix Σ =

$$\begin{bmatrix} 287.5 & -26.7 \\ -26.7 & 1123 \end{bmatrix} \tag{4.1}$$

Means $\mu 0$ =

$$\begin{bmatrix} 98.3800 & 429.6600 \end{bmatrix} \tag{4.2}$$

Mean $\mu 1$ =

$$\begin{bmatrix} 137.4600 & 366.6200 \end{bmatrix} \tag{4.3}$$

4.2 Part B and Part C

For case of Equal Covariance Matrix, the equation for boundary is given by

$$\log \frac{P(x|y=0)}{P(x|y=1)} = 0 \tag{4.4}$$

$$=> (\mu_0 - \mu_1)\Sigma^{-1} X = 1/2(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1) \tag{4.5}$$

$$Let \tag{4.6}$$

$$c = 1/2(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1) \tag{4.7}$$

$$P = [p_0 \quad p_1]^T = (\mu_0 - \mu_1)\Sigma^{-1} \tag{4.8}$$

$$x = [x_0 \quad x_1]^T \tag{4.9}$$

$$c \in R, P \in R^2 and x \in R^2 \tag{4.10}$$

$$Hence \tag{4.11}$$

$$P^T x = c \tag{4.12}$$

$$p_0 x_0 + p_1 x_1 = c \tag{4.13}$$

$$x_1 = \frac{c - p_0 x_0}{p1} \tag{4.14}$$

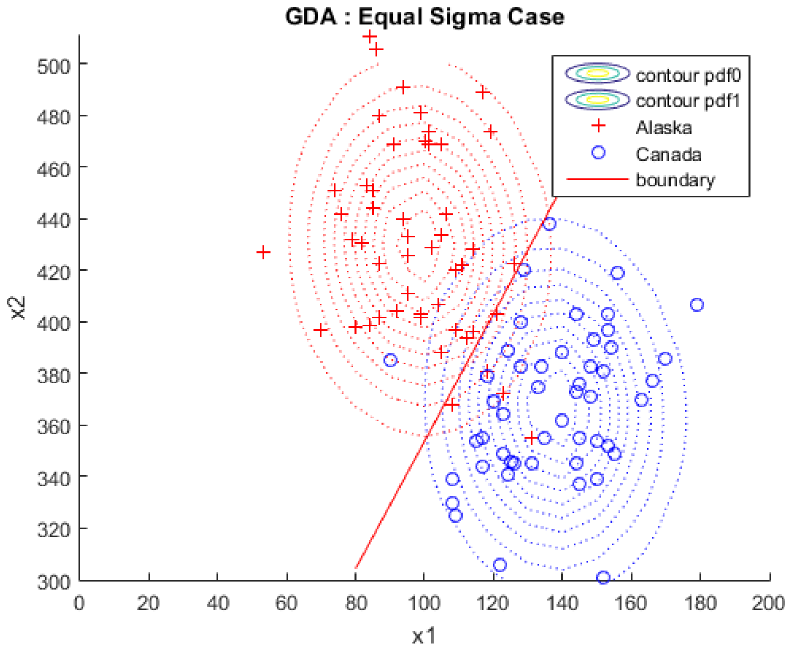


Figure 4.1: Linear Discriminant Analysis.

4.3 Part D

Covariance Matrix Σ_0 =

$$\begin{bmatrix} 63.8489 & -46.0827 \\ -46.0827 & 342.7761 \end{bmatrix} \tag{4.15}$$

Covariance Matrix Σ_1 =

$$\begin{bmatrix} 79.8921 & 32.7087 \\ 32.7087 & 218.8489 \end{bmatrix} \tag{4.16}$$

Mean $\mu 0$ =

$$\begin{bmatrix} 98.3800 & 429.6600 \end{bmatrix} \tag{4.17}$$

Mean $\mu 1$ =

$$\begin{bmatrix} 137.4600 & 366.6200 \end{bmatrix} \tag{4.18}$$

4.4 Part E

For case of unequal covariance the boundary is give by

$$\log\frac{P(x|y=0)P(y=0)}{P(x|y=1)P(y=1)}=0 \tag{4.19}$$

$$=> log(\frac{1-\phi}{\phi}-1/2(x^T(\Sigma_0^{-1}-\Sigma_1^{-1})x-2x^T(\Sigma_0^{-1}\mu_0-\Sigma_1^{-1}\mu_1)+(\mu_0^T\Sigma_0^{-1}\mu_0-\mu_1^T\Sigma_1^{-1}\mu_1)) \tag{4.20}$$

$$Let \tag{4.21}$$

$$A=(\Sigma_0^{-1}-\Sigma_1^{-1}) \tag{4.22}$$

$$B=(\Sigma_0^{-1}\mu_0-\Sigma_1^{-1}\mu_1) \tag{4.23}$$

$$C=(\mu_0^T\Sigma_0^{-1}\mu_0-\mu_1^T\Sigma_1^{-1}\mu_1)) \tag{4.24}$$

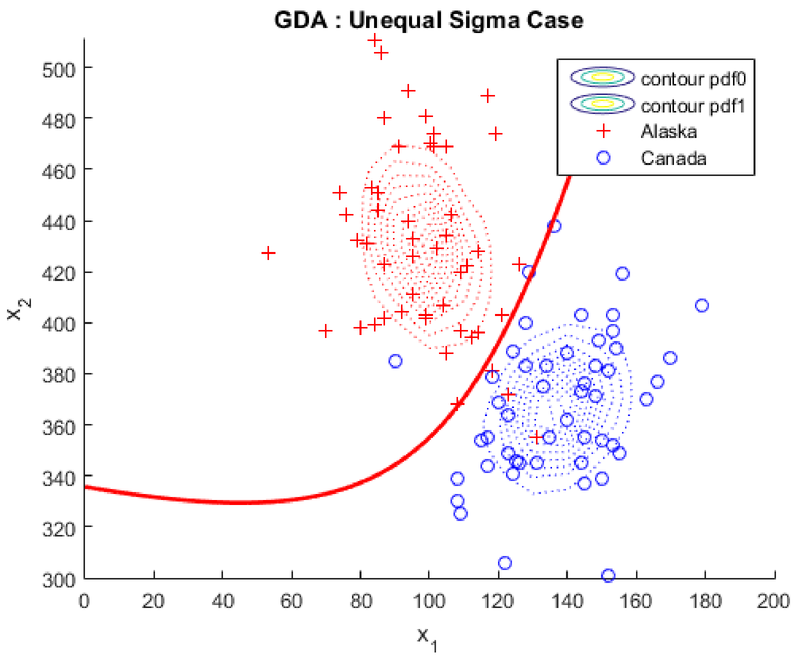
$$Hence \tag{4.25}$$

$$x^TAx-2x^TB+C=0 \tag{4.26}$$

$$x_1^2*A_{11}+x_1x_2(A_{21}+A_{12})+x_2^2A_{22}=2x_1b_1-2x_2b_2+C=0 \tag{4.27}$$

$$\tag{4.28}$$

Figure 4.2: Quadratic Discriminant Analysis.



4.5 Part F

With Equal Covariance Matrices we get aligned elliptical contours around for the two cases. Also the boundary seperating the two cases is a linear. Hence this method is called Linear

Discriminant Analysis (LDA). With Separate Covariance Matrices, we get elliptical contours for the two cases which are not-aligned with each other. Also the boundary separating the two cases is quadratic . Hence this is called Quadratic Discriminant Analysis (QDA). Since the data does have inherently different Covariances for the two cases, the Quadratic Discriminant Analysis is likely to perform better. For the training data itself, 5 errors can be counted for QDA and 6 erros for LDA.