COL774 - Machine Learning
Assignment 1

Ritesh Baldva
2012CS50711

February 10, 2017

1 Question 1

1.1 Batch Gradient Descent

I was very pessimistic while choosing the learning rate. The answer I get seems to be correct from the minima obtained from the plot. The problem with the parameters chosen here, is that it took 1 million iterations to converge which is a lot. So, initially,

$$\eta = 0.00001$$
$$J(\Theta^{(t+1)}) - J(\Theta^{(t)}) = 10^{-12}$$
$$\Theta = [5.83888639, 4.59284554]^T$$

Interestingly, I also tried to understand how the number of iterations change with the change in learning step size and the convergence condition. It can be seen from the table that, step size affects the number of iterations more compared to the limit for convergence. I finally chose the combination of step size to be 0.75 and convergence limit to be 1e-3 for my plots, since it is has less than 1% error of the million iterations answer (supposed to be optimal) and converges very quickly too.

Step Size	Convergence Condition	Number of iterations	θ_0	$\theta 1$
1e-5	1e-12	1.08e06	5.83888639	4.59284554
1e-5	1e-9	6.61e05	5.8312753	4.58685868
1e-3	1e-9	8910	5.83835009	4.59242369
1e-3	1e-6	5458	5.81431611	4.57351865
1e-2	1e-6	659	5.83137362	4.58693602
1e-2	1e-3	315	5.59285183	4.3993157
1e-1	1e-3	42	5.76922592	4.53805088
1e-1	1e-2	31	5.61636012	4.41780723
0.75	1e-3	5	5.83343277	4.58855574

Thus, final values are:

$$\eta = 0.75$$
$$J(\Theta^{(t+1)}) - J(\Theta^{(t)}) = 10^{-3}$$
$$\Theta = [5.83343277, 4.58855574]^T$$

1.2 Data and Hypothesis Plot

The Data plot and hypothesis function can be seen in Q1 Figure (b)

1.3 3D Error function plot and Iteration Plot

The 3D error function plot can be seen in Q1 Figure (c1) and the iteration plot can be seen in Q1 Figure (c2).

1.4 Contour and Iteration Plot

The contour plot can be seen in Q1 Figure (d).

1.5 Different Learning Rates

The contour plots with different etas can be seen in Q1 Figure (e1=0.1, e2=0.3, e3=0.9, e4=1.3, e5 = 2.1, e6 = 2.5). You can see total iterations by seeing the number of dots on the plots. Like the table observed in the first part, initially, the number of iterations decreases, with the increase in the step size but for e4, the step size overshoots and thus, we can see a bit of oscillation from the start. Note with higher learning rates (e5 = 2.1, e6 = 2.5), the gradient descent diverges quickly.

2 Question 2

2.1 Unweighted Linear Regression

The plot of Unweighted regression can be seen in Figure Q2 (a).

2.2 Locally Weighted Regression

The plot of weighted regression, with bandwidth = 0.8, can be seen in Figure Q2 (b).

2.3 Different Bandwidth Parameters

The plot of weighted regression, with bandwidth = 0.1, 0.3, 2, 10, can be seen in Figure Q2 (c1), Figure Q2 (c2), Figure Q2 (c2) and Figure Q2 (c4) respectively. It can be seen that bandwidth = 0.3 suits the best to data. Notice that for bandwidth (10), it mimics the linear regression fit. That is because all the training points are given a very high weight since all lie within twice of bandwidth for any query point in the range of training set. To see the effects of how fit changes when bandwidth parameter gets smaller (to get the idea of over-fitting), I also tried bandwidth=0.025. For this, as we can see from Figure Q2 (c5), we see how the fit is trying to follow each point individually, and there's sort of a sinusoidal pattern emerging in the fit.

3 Question 3

3.1 Newton's Method Implementation & Logistic Parameters

The equations for Logisitic regression can also be solved using matrices making it easier to implement. The values of parameters obtained are:

$$\Theta = [-0.04717577, 1.46005896, 2.06586134]^T$$

3.2 Logistic Discriminant Plot

The data and discriminator can be seen in Figure Q3 (b).

4 Question 4

From the practice questions, we can see the values of the ML parameters, obtained by maximizing the log-likelihood equation:

$$LL(\phi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^{m} log(p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma)p(y^{(i)}; \phi))$$

4.1 Gaussian Discriminant Analysis

Note the $0, 1$ mapping to the outcomes $Alaska, Canada$ is maintained. Thus, the following values can be simply calculated from the values obtained:

$$\phi = \frac{1}{m}\sum_{i=1}^{m} 1(y^{(i)} = 1) = 0.5$$

$$\mu_0 = \frac{\sum_{i=1}^{m} 1(y^{(i)} = 0) * x^{(i)}}{\sum_{i=1}^{m} 1(y^{(i)} = 0)} = [-0.75529433, 0.68509431]^T$$

$$\mu_1 = \frac{\sum_{i=1}^{m} 1(y^{(i)} = 1) * x^{(i)}}{\sum_{i=1}^{m} 1(y^{(i)} = 1)} = [0.75529433, -0.68509431]^T$$

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

4.2 Original Data Plot

The graph can be seen in Figure Q4(b).

4.3 Linear Discriminant Plot

The Linear Discriminator can be seen in Figure Q4(c).

4.4 Different Covariance Matrices

Values of parameters obtained when covariance matrices are different.

$$\phi = 0.5$$

$$\mu_0 = [-0.75529433, 0.68509431]^T$$

$$\mu_1 = [0.75529433, -0.68509431]^T$$

$$\Sigma_0 = \begin{bmatrix} 0.38158978 & -0.15486516 \\ -0.15486516 & 0.64773717 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 0.47747117 & 0.1099206 \\ 0.1099206 & 0.41355441 \end{bmatrix}$$

4.5 Quadratic and Linear Discriminant Plot

The Linear and Quadratic Discriminator in Figure Q4(e).

4.6 Observations

The quadratic separator observed is the upper lobe of a hyperbola. While choosing the points, in the code, I chose only the positive y values. To see the other lobe, we just had to choose the negative values of y. After solving the boundary condition problem, we get the following equation:

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

This equation in two variables can thus be reduced to:

$$Ax_0^2 + Bx_0x_1 + Cx_1^2 + Dx_0 + Ex_1 + F = 0$$

And thus, we can find the eccentricity of the curve here,

$$e = \sqrt[2]{\frac{2\sqrt[2]{(A-C)^2 + B^2}}{\eta(A+C) + \sqrt[2]{(A-C)^2 + B^2}}}$$

For this case, the eccentricity comes out to be $e = 1.37$, suggesting that the quadratic separator is a hyperbola.

From the plot, we can see that the quadratic classifier, separates the Canadian fish better than the linear separator.

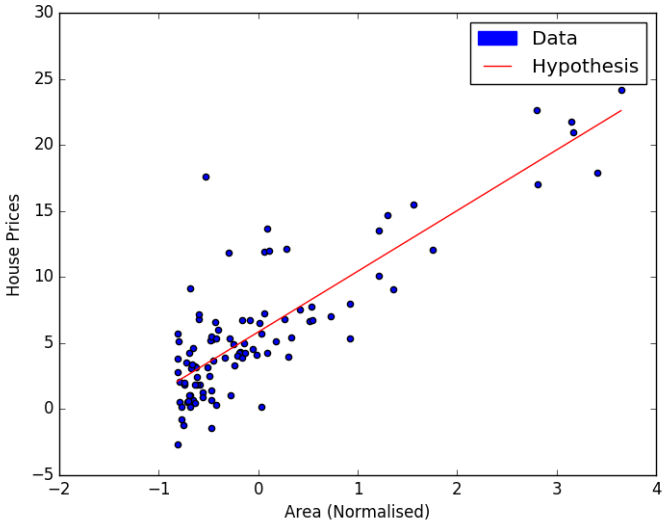

Figure 1: Q1(b) Data and Hypothesis

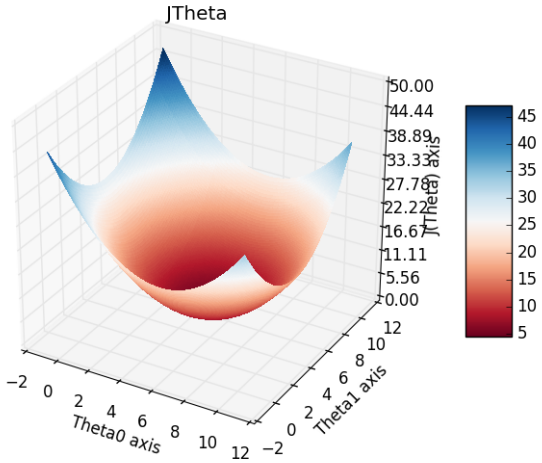

Figure 2: Q1(c1) 3D Error Function Plot

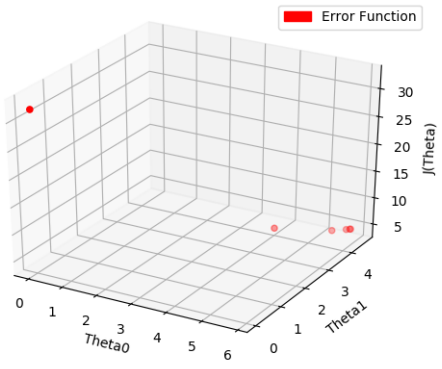

Figure 3: Q1(c2) 3D Iteration Plot

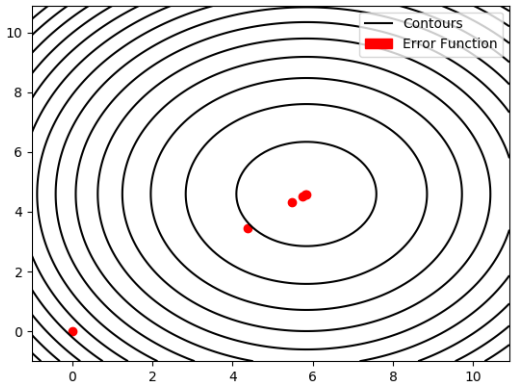

Figure 4: Q1(d) 3D Iteration Plot

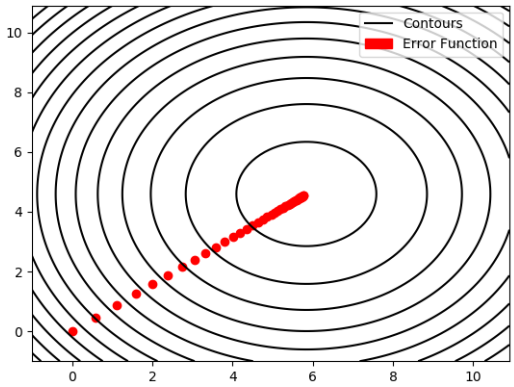

Figure 5: Q1(e1) Eta = 0.1

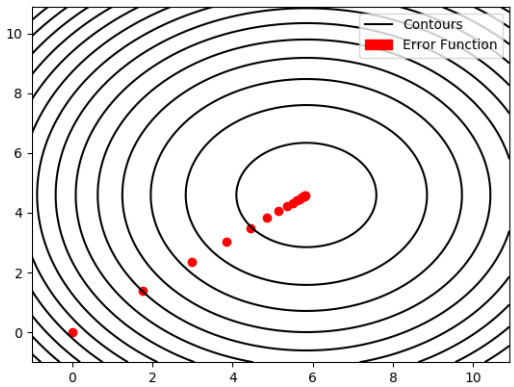

Figure 6: Q1(e2) Eta = 0.5

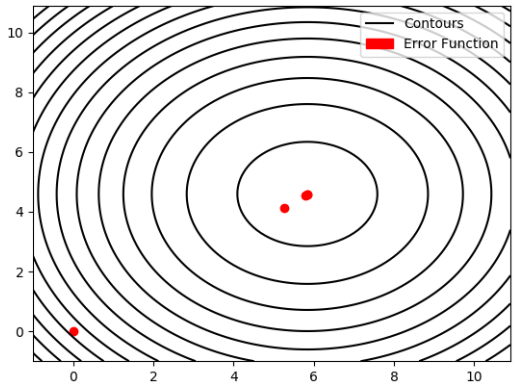

Figure 7: Q1(e3) Eta = 0.9

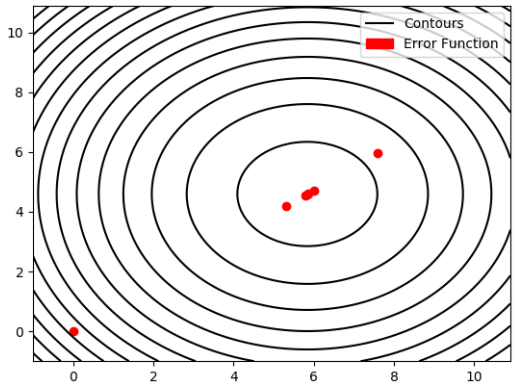

Figure 8: Q1(e4) Eta = 1.3

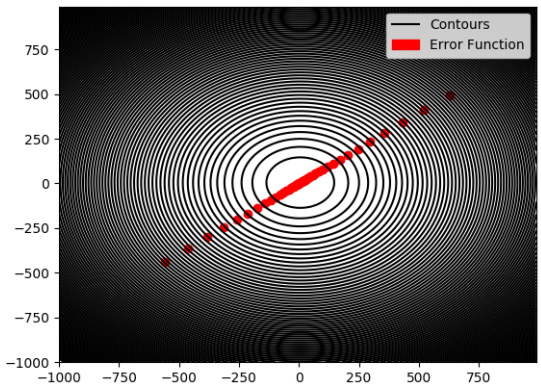

Figure 9: Q1(e5) Eta = 2.1

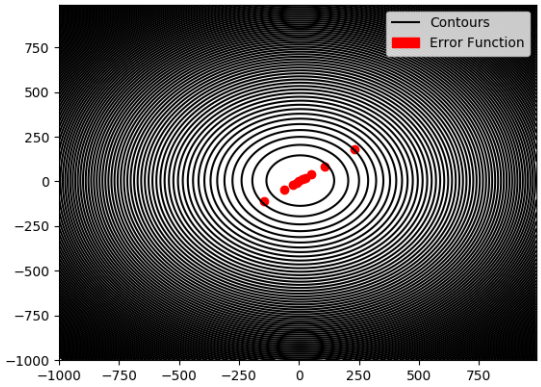

Figure 10: Q1(e6) Eta = 2.5

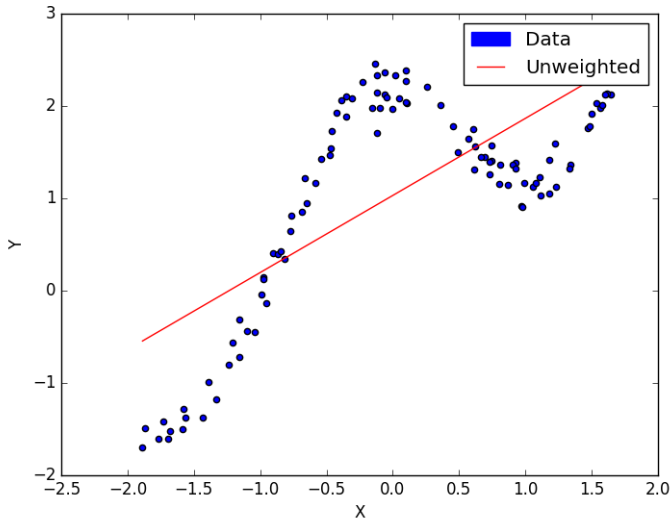

Figure 11: Q2(a) Unweighted linear regression

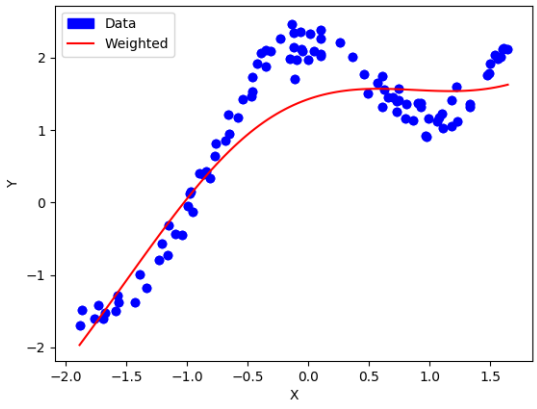

Figure 12: Q2(b) Weighted (Bandwidth = 0.8) linear regression

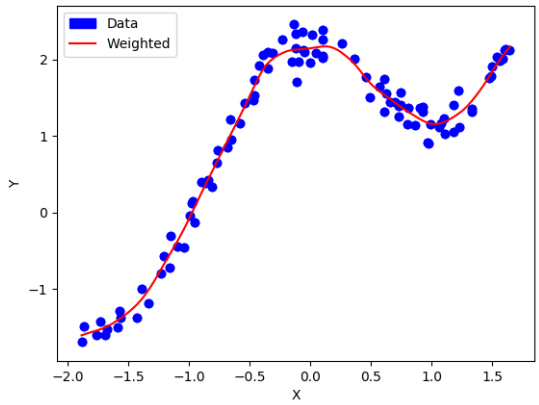

Figure 13: Q2(c1) Weighted (Bandwidth = 0.1) linear regression

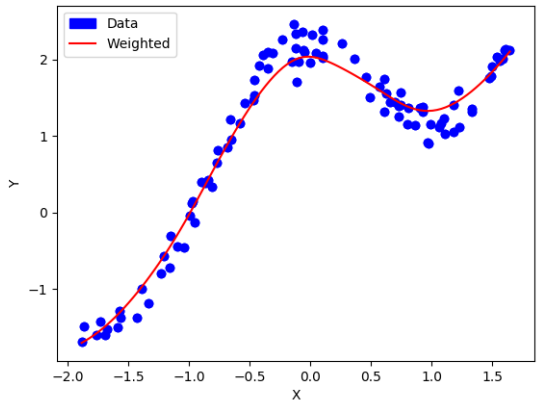

Figure 14: Q2(c2) Weighted (Bandwidth = 0.3) linear regression

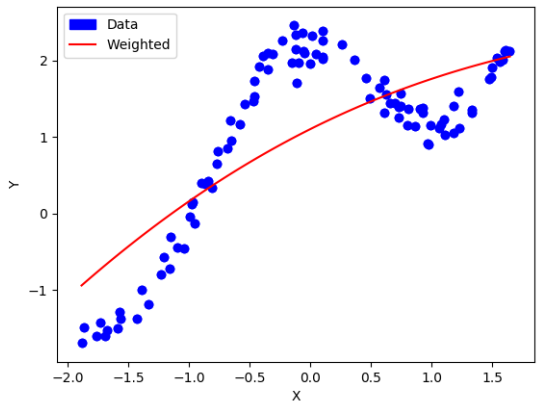

Figure 15: Q2(c3) Weighted (Bandwidth = 2) linear regression

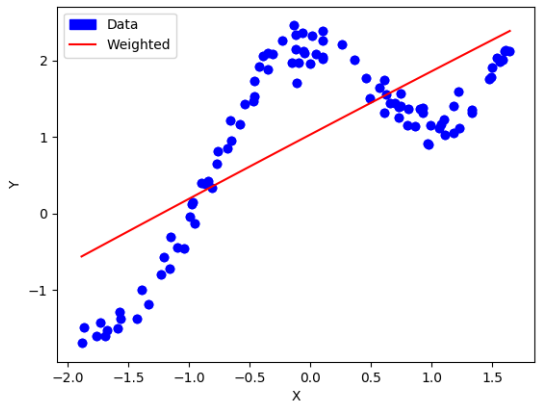

Figure 16: Q2(c4) Weighted (Bandwidth = 10) linear regression

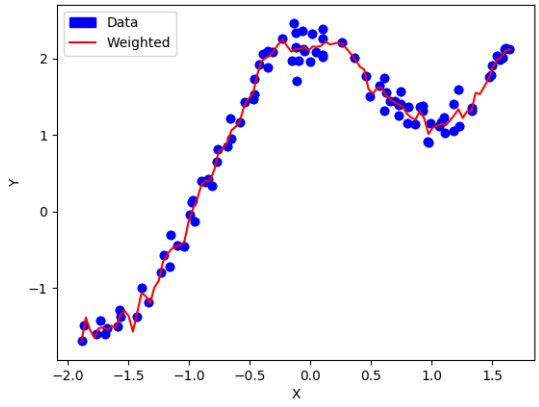

Figure 17: Q2(c4) Weighted (Bandwidth = 0.025) linear regression

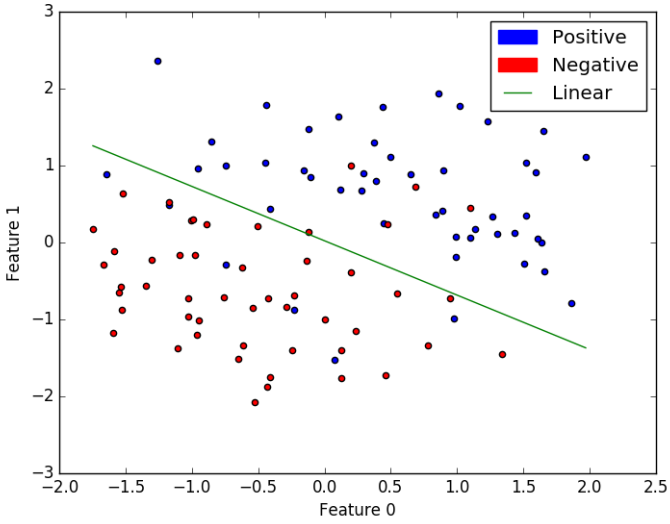

Figure 18: Q3(b) Logistic Discriminator and Data in the Feature Space

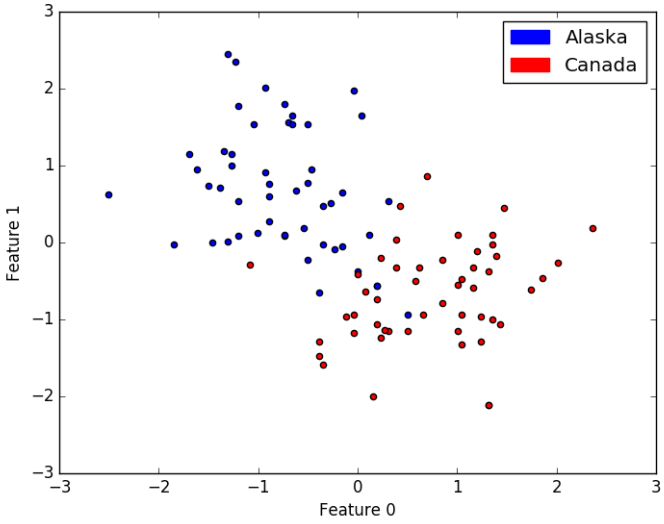

Figure 19: Q4(b) Data in the Feature Space

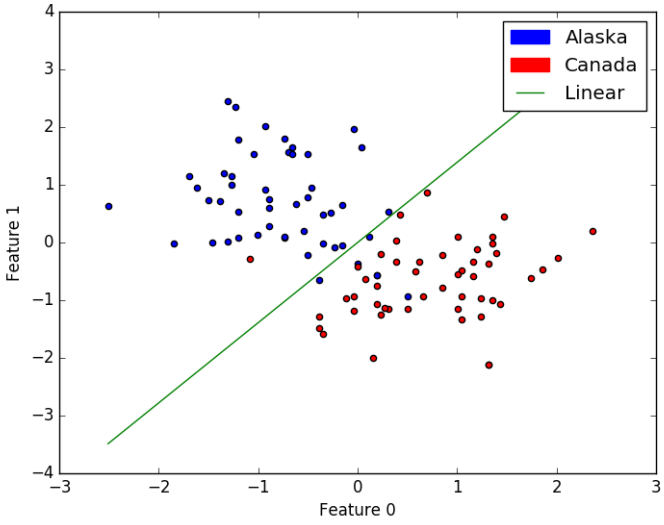

Figure 20: Q4(c) Linear Separator

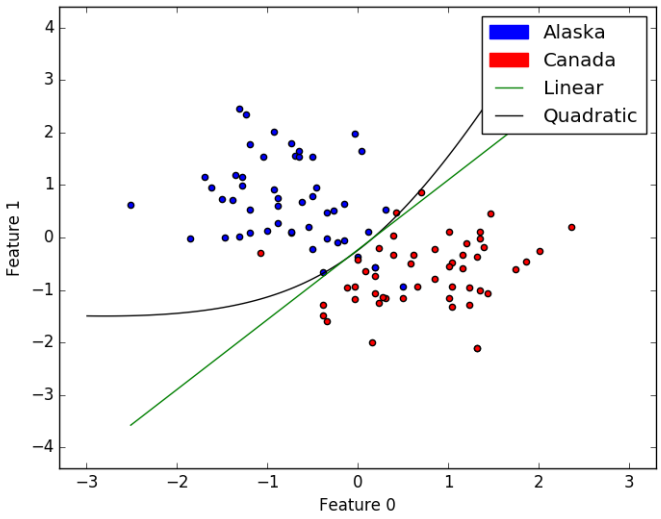

Figure 21: Q4(e) Quadratic and Linear Separator