COL774: Assignment 1

Nikhil Goyal
2015CS50287

February 13, 2018

1 Linear Regression

Note: All results in this section are based on normalized training data

1.a

Learning rate (η) = 0.017

Stopping Criteria = $|J(\theta^{(t+1)}) - J(\theta^t)| < 1 \times 10^{-10}$ or iterations ≥ 1000

$\theta = \begin{bmatrix} 0.99661881 \\ 0.00134019 \end{bmatrix}$ Number of iterations = 38

1.b

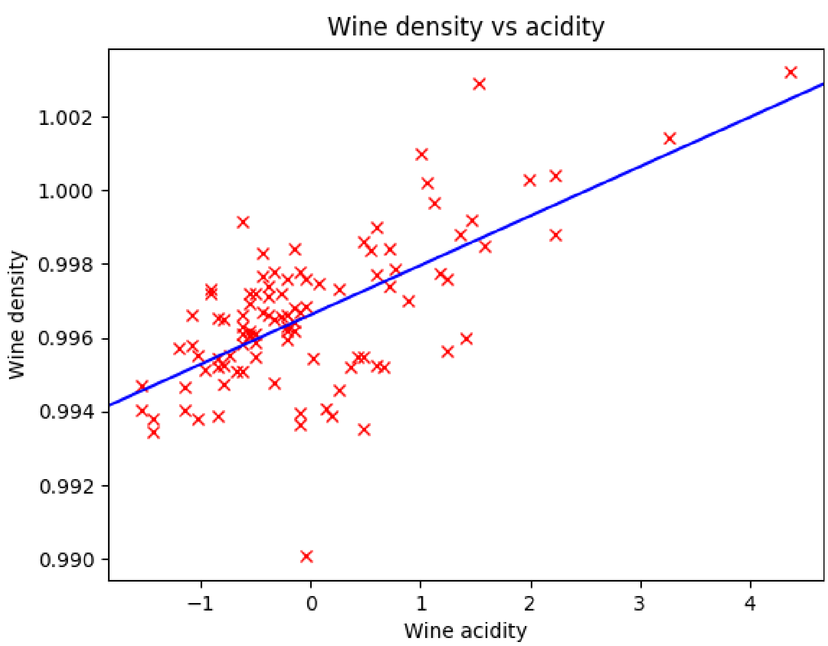

Figure 1: Wine Density vs Wine Acidity

1.c

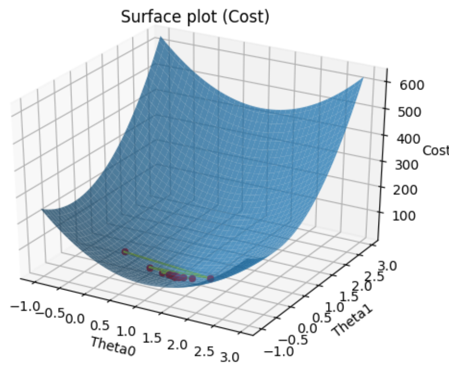

Figure 2: Path taken by gradient descent on surface plot of $J(\theta)$

1.d

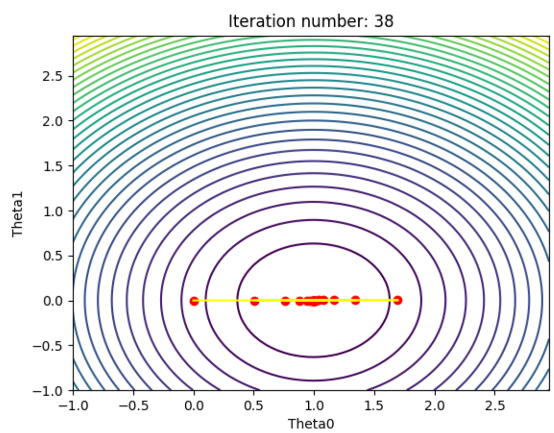

Figure 3: Path taken by gradient descent on contour plot of $J(\theta)$

1.e

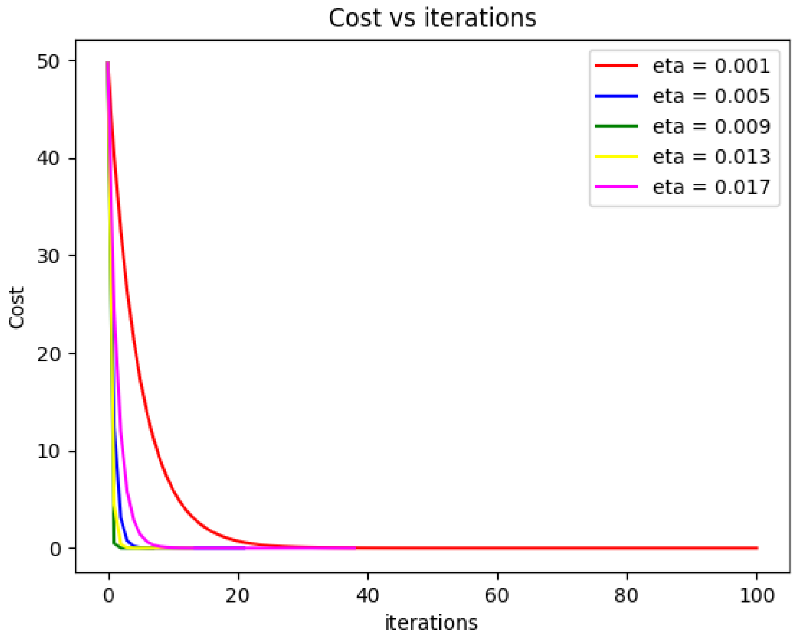

Figure 4: $J(\theta)$ vs iterations for various η

Gradient descent diverges for $\eta = 0.021$ and $\eta = 0.025$. With a small learning rate, we find that gradient descent takes a very long time to converge to the optimal value. Conversely, with a large learning rate, gradient descent might not converge or might even diverge.

2 Locally Weighted Linear Regression

2.a

$\theta = \begin{bmatrix} 0.32767322 \\ 0.17531247 \end{bmatrix}$

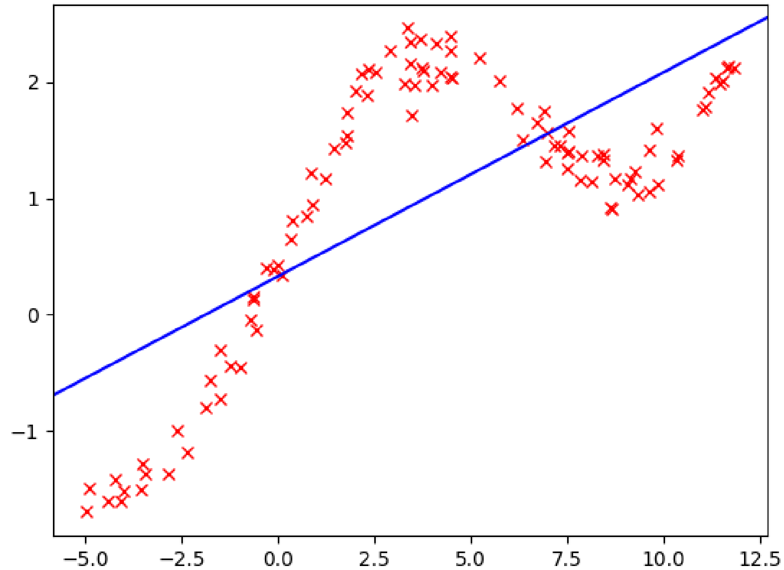

Figure 5: Linear Regression

2.b

In locally weighted linear regression

$$\theta = (X^T W X)^{-1} X^T W Y$$

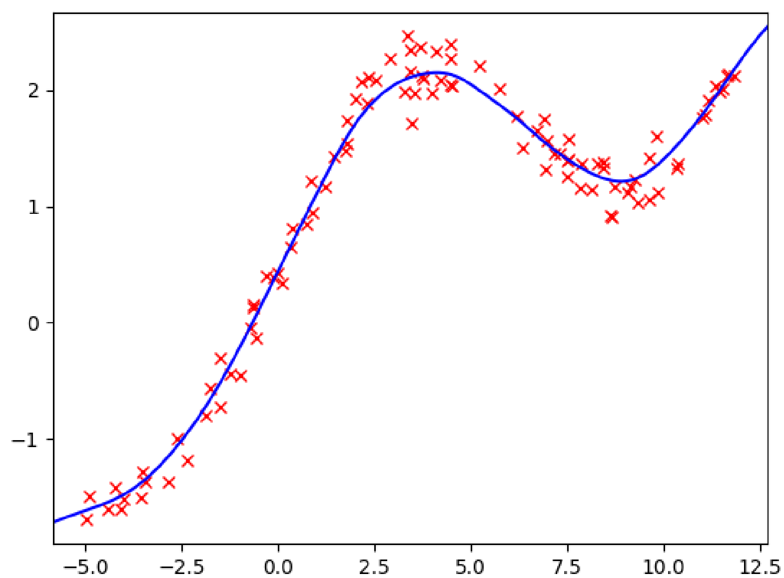

Figure 6: Locally Weighted Linear Regression ($\tau = 0.8$)

2.c

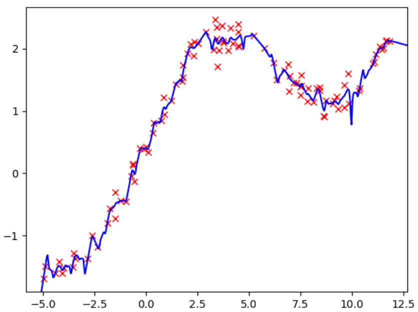

Figure 7: $\tau = 0.1$

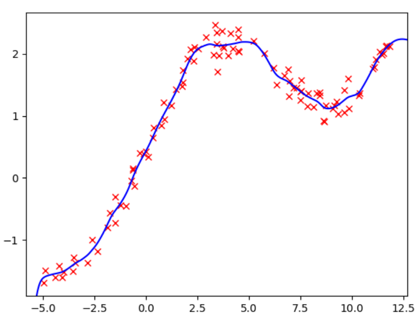

Figure 8: $\tau = 0.3$

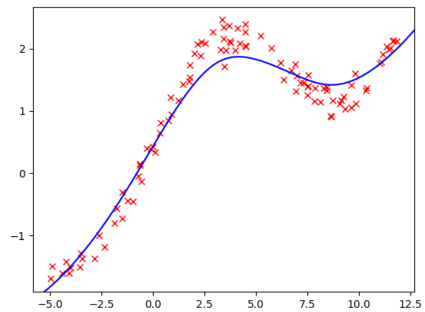

Figure 9: $\tau = 2$

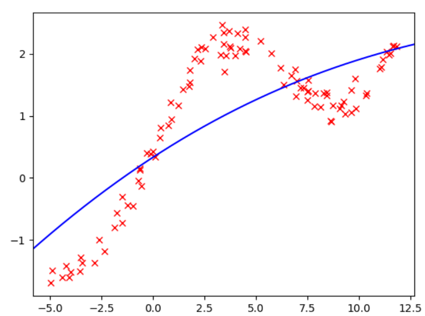

Figure 10: $\tau = 10$

$\tau = 0.8$ works the best as it caputures the data structure correctly and generalizes it well.

When τ goes to zero, the curve becomes more and more like a perfect polynomial fitting all points correctly. Conversely, when τ is very large, hyothesis become almost same as that of (un-weighted) linear regression.

3 Logistic Regression

Note: All results in this section are based on normalized training data

3.a

$\theta = \begin{bmatrix} 0.40125316 \\ 2.5885477 \\ -2.72558849 \end{bmatrix}$ Number of iterations = 8

3.b

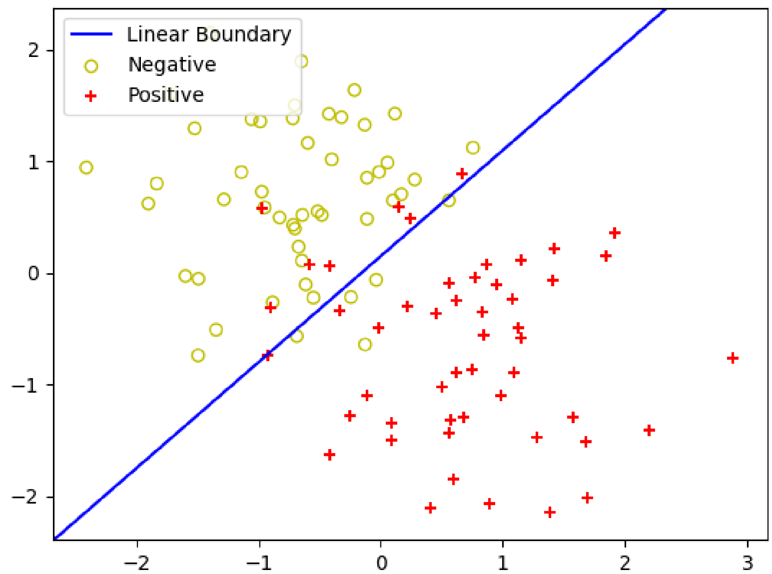

Figure 11: Linear decision boundary for Logistic regression

4 Gaussian Discrimant Analysis

4.a

$\phi = 0.5$

$$\mu_0 = \begin{bmatrix} 98.38 \\ 429.66 \end{bmatrix}$$

$$\mu_1 = \begin{bmatrix} 137.46 \\ 366.62 \end{bmatrix}$$

$$\Sigma_0 = \Sigma_1 = \Sigma = \begin{bmatrix} 287.482 & 26.748 \\ -26.748 & 1123.25 \end{bmatrix}$$

4.b

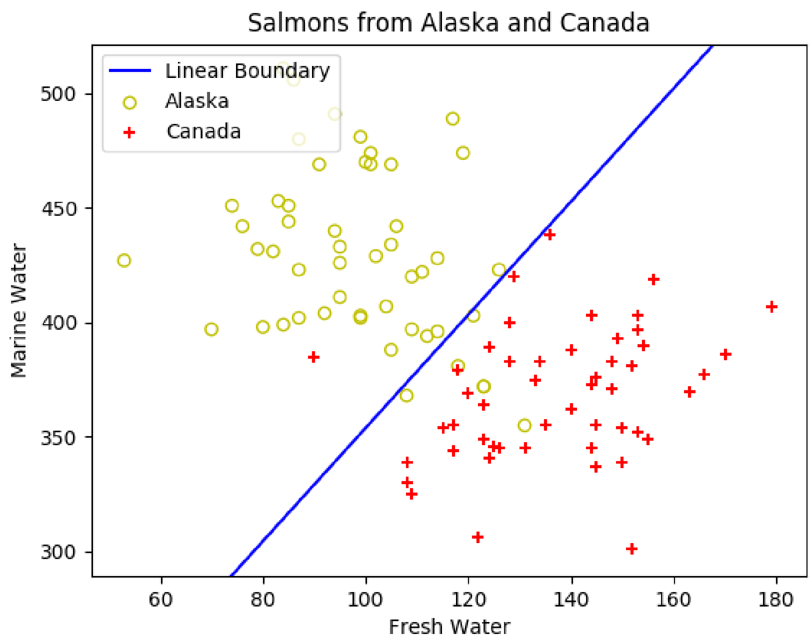

Figure 12: GDA decision boundary when $\Sigma_0 = \Sigma_1$

4.c

Decision boundary is linear when $\Sigma_0 = \Sigma_1$ and the equations is given by

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

Or in other form as

$$\theta^T x' = 0, \quad where$$

$$\theta = \begin{bmatrix} \frac{-1}{2}(\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0) - log(\frac{1-\phi}{\phi}) \\ \Sigma^{-1}(\mu_0 - \mu_1) \end{bmatrix}$$

$$x' = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

4.d

$\phi = 0.5$

$\mu_0 = \begin{bmatrix} 98.38 \\ 429.66 \end{bmatrix}$

$\mu_1 = \begin{bmatrix} 137.46 \\ 366.62 \end{bmatrix}$

$\Sigma_0 = \begin{bmatrix} 255.3956 & -184.3308 \\ -184.3308 & 1371.1044 \end{bmatrix}$

$\Sigma_1 = \begin{bmatrix} 319.5684 & 130.8348 \\ 130.8348 & 875.3956 \end{bmatrix}$

4.e

Decision boundary is general case is given by

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

which can be written as

$$x^TAx + B^Tx + C = 0, \quad where$$
$$A = \Sigma_1^{-1} - \Sigma_0^{-1}$$
$$B = -2(\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0)$$
$$C = \mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0 - 2log(\frac{\phi}{1-\phi}) + log(\frac{|\Sigma_1|}{|\Sigma_0|})$$

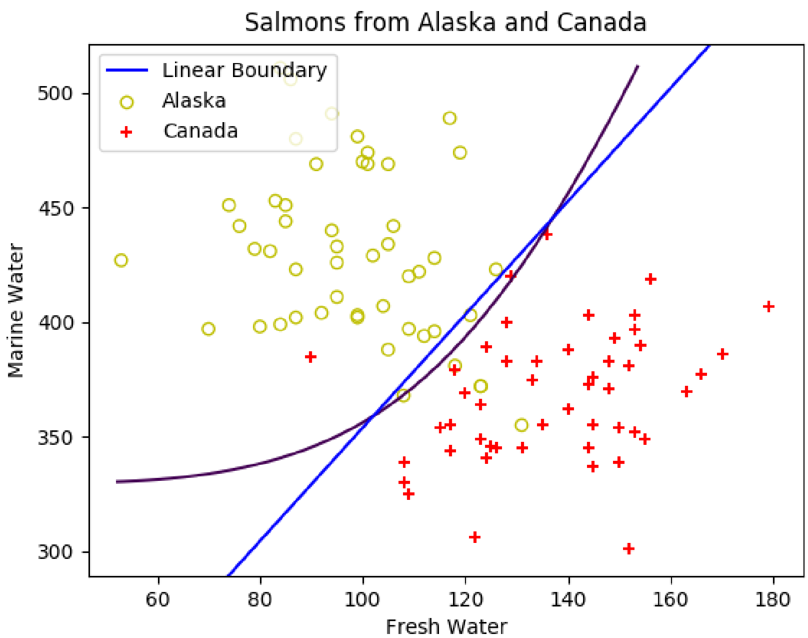

Figure 13: GDA decision boundary in general setting

4.f

From the quadratic equation written above we have in our case

$$A = \begin{bmatrix} -5.8321 & -4.9822 \\ -4.9822 & 12.1672 \end{bmatrix} \times 10^{-4}$$

The conic is degenerate and $A_{10}A_{01} - A_{00}A_{11} = 9.57836 \times 10^{-7} > 0$, so the quadratic boundary obtained is a hyperbola. The mirror curve of hyperbola is just a mathematical artifact and has no relation to the classification boundary. The hyperbola does a better job in classification as we can see from the graph. This is due to the fact general GDA works with weaker assumptions than the case when GDA is applied with $\Sigma_0 = \Sigma_1$.