COL774: Machine Learning	**Date: February 12, 2019**

Assignment 1 Report

Entry Number: 2016CS10363	**Name: Manish Tanwar**

1 Linear Regression:

Note: I have normalized the data and calculated the parameters back for original data.

(a):

- Learning Rate = 0.5
- **Stopping Criteria:**

$$\left|\theta_j^{(t+1)} - \theta_j^{(t)}\right| < \epsilon \qquad \forall j \in 1, 2...n \qquad \text{(for a sufficiently small } \epsilon \text{ (took } \epsilon = 10^{-4}))$$

- Parameters (Calculated for original(unnormalized) data):

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 0.990289 \\ 0.000778 \end{bmatrix}$$

(b):

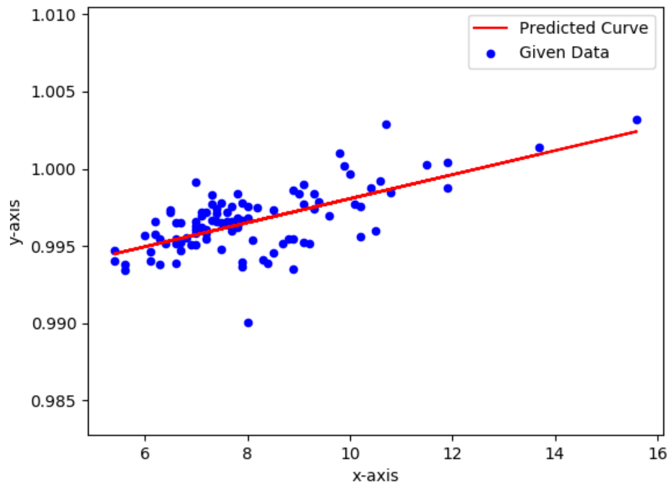

Figure 1: Linear Regression : Data Points and Hypothesis Function Plot

(c):

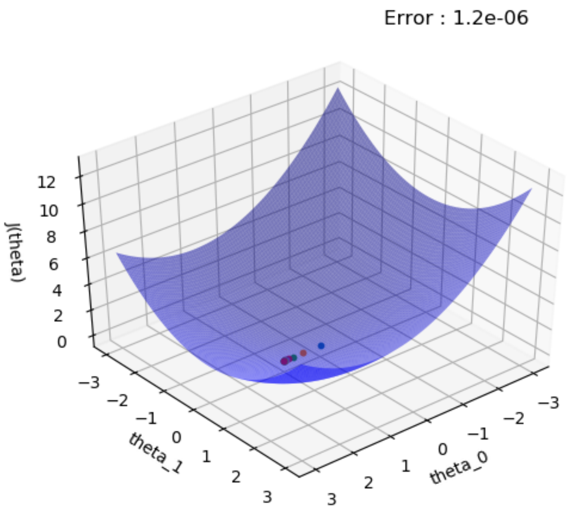

Figure 2: Linear Regression : 3D mesh

(d):

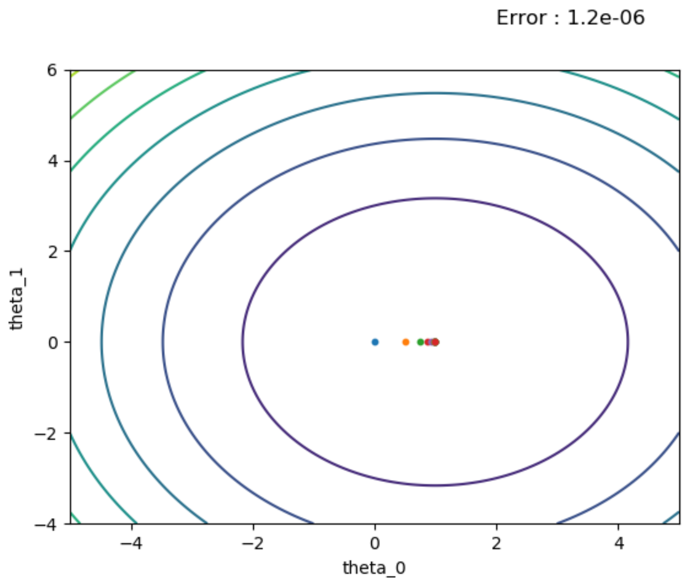

Figure 3: Linear Regression : Contours

(e) Observations:

On plotting the contours figures for various values of Learning Rate(η), we get the following observations:

- For $\eta \in \{0.1, 0.5, 0.9\}$, the error function converges to the minima.
- For $\eta \in \{1.3, 1.7\}$, the error function overshoots to the other side of the minima and toggles around the minima and finally converges.
- For $\eta \in \{2.1, 2.5\}$, the error fuction diverges from the minima.
- The number of iterations required decreases as we increase the learning rate(η), but after a certain value when the error function overshoots it takes more iterations to converge and finally the error function diverges on large learning rates.

2 Locally Weighted Linear Regression:

(a) Linear Regression (unweighted):

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 0.327680 \\ 0.175316 \end{bmatrix}$$

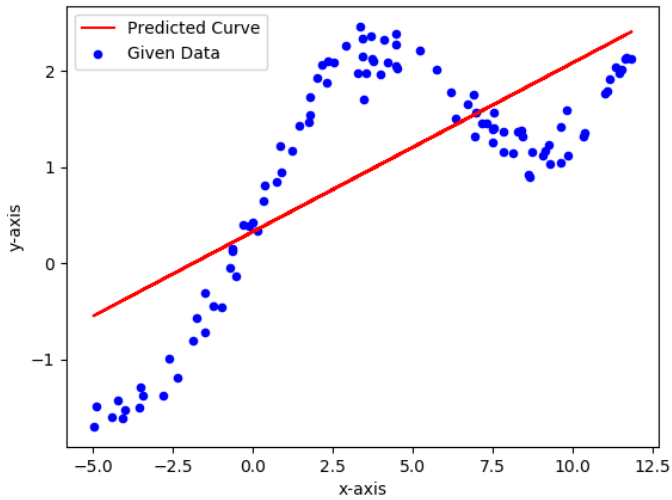

Figure 4: Linear Regressin Plot (Underfitting)

Linear Regression(unweighted) is not a good fit for the data as shown in the Figure 1(underfitting).

(b) Locally Weighted Linear Regression:

Weights : $$w^{(i)} = \exp\left(-\frac{(x - x^{(i)})^2}{2\tau^2}\right) \qquad \text{(where } \tau = \text{Bandwidth Parameter)}$$

Error Function: $$J(\theta) = \frac{1}{2m}(X\theta - Y)^T W (X\theta - Y) \qquad \text{(where } W = diag(w^{(i)})\text{)}$$

Minima: $$\nabla_\theta J(\theta) = 0 \quad \Rightarrow \quad \theta = (X^T W X)^{-1} X^T W Y$$

Plots:

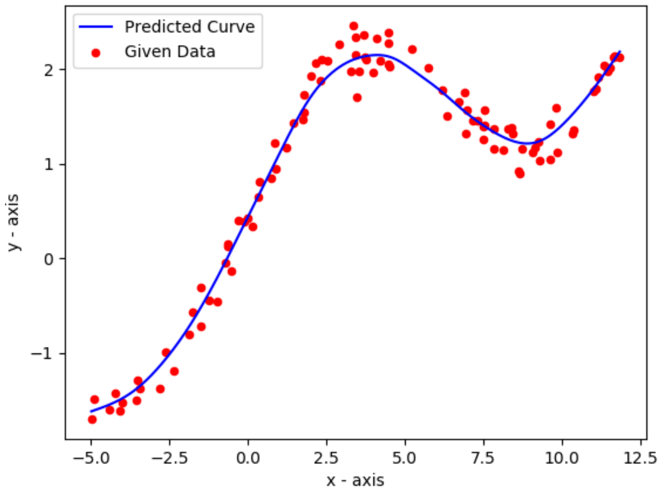

Figure 5: Locally Weighted Linear Regressin Plot ($\tau = 0.8$)

(c) Plots on Varying Bandwidth Parameter(τ)

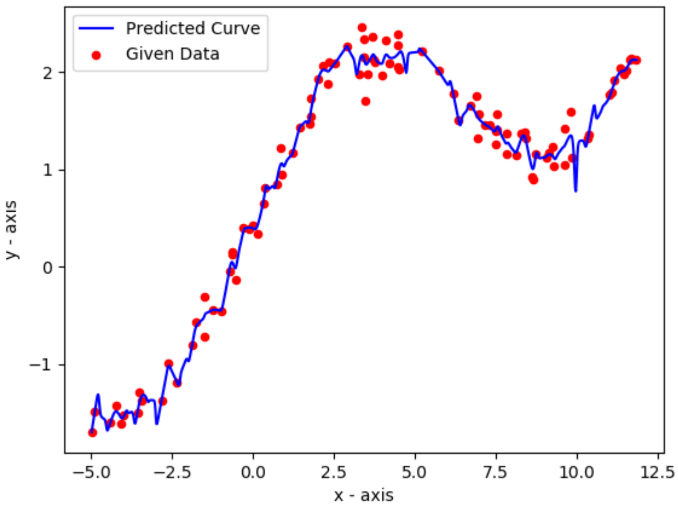

Figure 6: $\tau = 0.1$

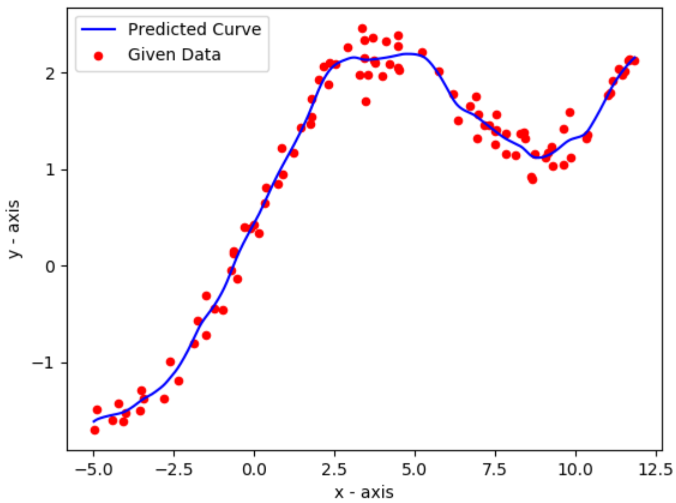

Figure 7: $\tau = 0.3$

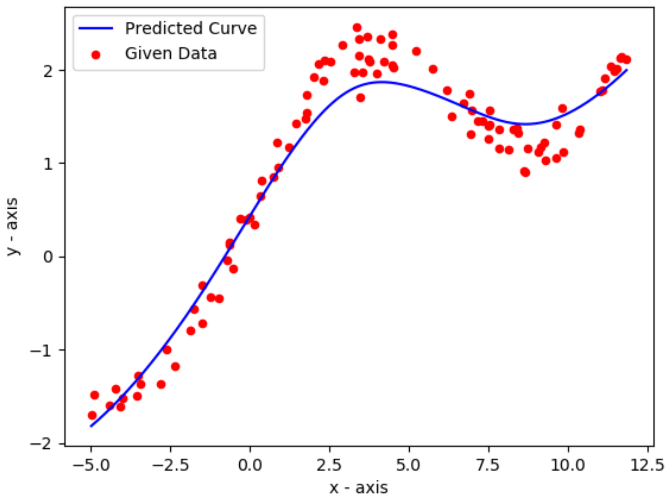

Figure 8: $\tau = 2$

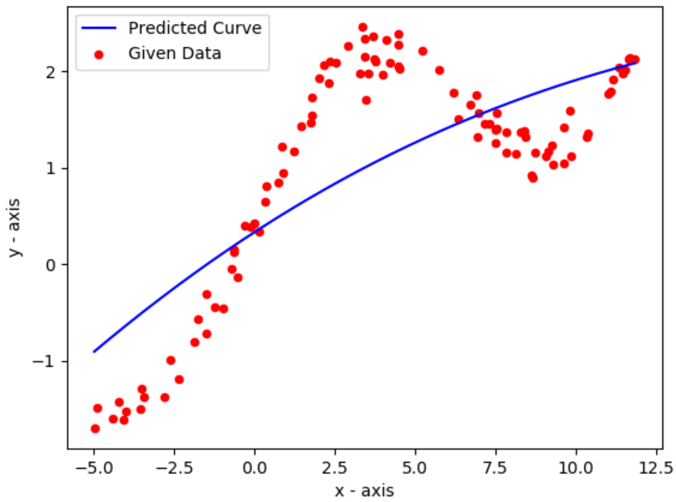

Figure 9: $\tau = 10$

Analysis:

- When Bandwidth Parameter(τ) is:
  - **Too Small:** It results in **overfitting**.
  - **Too Large:** It results in **underfitting**.
- If τ is too small the model looks at really close data-points to the query data point x which results in overfitting and if τ is too large it predicts the query data point based on giving all the data same weights(overgeneralization) which results in underfitting.
- $\tau = 0.8$ works the best in our case.

3 Logistic Regression:

Log Likelihood: $$LL(\theta) = \sum_{i=1}^{m} y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))$$

$$\nabla_\theta LL(\theta) = X^T(Y - g(X\theta)) \qquad \text{(where } g(x) = \frac{1}{1 + \exp(-x)}\text{)}$$

Hessian Matrix: $$H = \nabla_\theta^2 LL(\theta) = -X^T D X$$

$$\text{(where } D = diag(\, g(x^{(i)T}\theta)(1 - g(x^{(i)T}\theta))\,)\text{)}$$

Newton's Method:

$$\theta^{(t+1)} = \theta^{(t)} - H^{-1} \nabla_\theta LL(\theta)\big|_{\theta_t}$$

Stopping Criteria:

$$|\theta_j^{(t+1)} - \theta_j^{(t)}| < \epsilon \qquad \forall j \in 1, 2 \ldots n \qquad \text{(for a sufficiently small } \epsilon \text{ (took } \epsilon = 10^{-8}\text{))}$$

Resulting Parameters:

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 0.223295 \\ 1.962616 \\ -1.964861 \end{bmatrix}$$

Decision Boundary is the straight line boundary separating the region where $h_\theta(x) \geq 0.5$ (class $y = 1$) from where $h_\theta(x) \leq 0.5$ (class $y = 0$).

Plot:

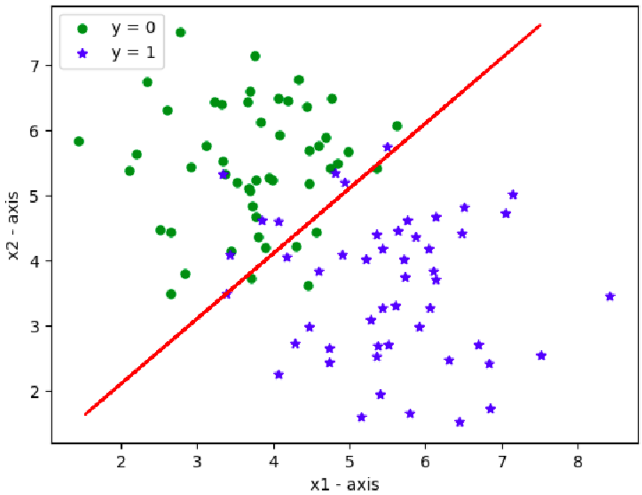

Figure 10: Logistic Regression : Given data and Linear Separator

4 Gaussian Discrmimant Analysis:

(a):

$$\phi = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 1\} \qquad \mu_0 = \frac{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 0\}} \qquad \mu_1 = \frac{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 1\}}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

which can be written as: $\Sigma = \frac{1}{m}W^TW$ (where $W = X - Y\mu_1^T - (1-Y)\mu_0^T$)

Resulting Parameters:

$$\phi = 0.5 \qquad \mu_0 = \begin{bmatrix} 98.38 \\ 429.66 \end{bmatrix} \qquad \mu_1 = \begin{bmatrix} 137.46 \\ 366.62 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 287.482 & -26.748 \\ -26.748 & 1123.25 \end{bmatrix}$$

(b) & (c):

(i) Linear Boundary Equation:

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

(ii) Plot:

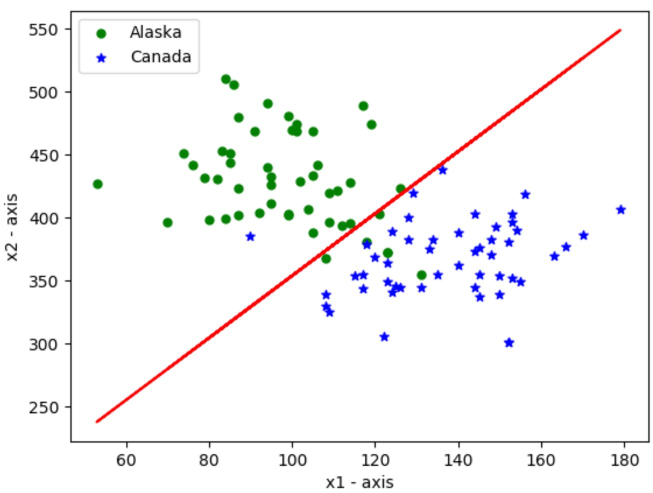

Figure 11: GDA : Given data and Linear Separator

(d):

$$\phi = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 1\} \qquad \mu_0 = \frac{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 0\}} \qquad \mu_1 = \frac{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 1\}}$$

$$\Sigma_0 = (X - \mu_0^{ext})^T diag(y^{(i)})(X - \mu_0^{ext}) \quad \text{(where } \mu_0^{ext} \text{ is } m \times n \text{ matrix with each row} = \mu_0^T)$$

Resulting Parameters:

$$\phi = 0.5 \qquad \mu_0 = \begin{bmatrix} 98.38 \\ 429.66 \end{bmatrix} \qquad \mu_1 = \begin{bmatrix} 137.46 \\ 366.62 \end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix} 255.3956 & -184.3308 \\ -184.3308 & 1371.1044 \end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix} 319.5684 & 130.8348 \\ 130.8348 & 875.3956 \end{bmatrix}$$

(e):

(i) Quadratic Boundary Equation:

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

(ii) Plot:

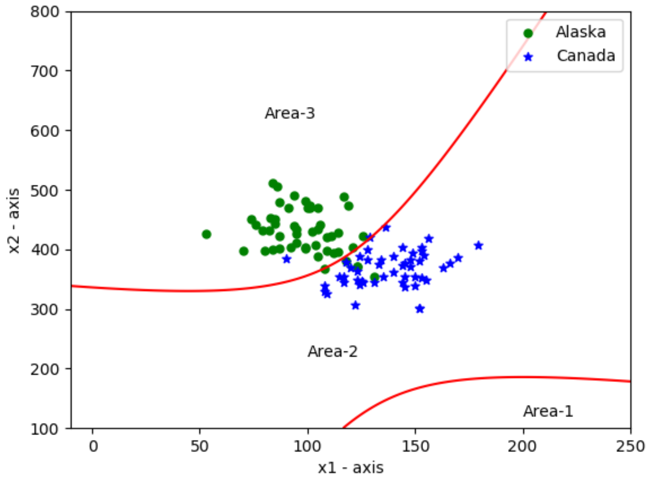

Figure 12: GDA : Given data and Quadratic Separator

(f) Analysis:

- In our case both linear and hyperbolic boundaries separates the data to a fair extend.
- But hyperbola gives 3 classes(Area-1, Area-2, Area-3 as shown in Figure 12) in the graph:

$$\text{Area-1 and Area-3} \quad \Rightarrow \quad y = 0$$
$$\text{Area-2} \quad \Rightarrow \quad y = 1$$

  This classifier classifies Area-1 in $y = 0$ class, which does not seem to be true in this case. This dataset does not contain any point in Area-1.
  Similarly ellipse and hyperbola are conics which won't be able to classify linearly separated data.

- As we make stronger assumption in case of linear separator($\Sigma_0 = \Sigma_1$) which could not be true some times, so quadratic separator would perform better than the linear separtor in these type of the cases.