Machine Learning Homework 1

Mengchuan Fu (Mike)
A14008047

February 19, 2016

1 Problem 1: Polynomial Regression

Hypothesis:

$$f(x,\Theta) = \Theta_0 = \Theta_0 + \Theta_1 x + \Theta_2 x^2 + \ldots + \Theta_d x^d$$

Parameters:

$$d; \Theta_0, \Theta_1 \ldots \Theta_d$$

Cost Function:

$$J(\Theta_0, \Theta_1, \ldots, \Theta_d) = \frac{1}{2m} \sum (f_\Theta(x_i) - y_i)^2$$

Goal:

$$\min_\Theta J(\Theta)$$

Method:
Least Square (OLS)

Derivation:
(1).Count the partial derivatives for the cost function for each "d"

$$-2\sum[y - (a_0 + a_1 x + \ldots + a_k x^k)] = 0$$
$$-2\sum[y - (a_0 + a_1 x + \ldots + a_k x^k)]x = 0$$

$$\vdots$$

$$-2\sum[y-(a_0+a_1x+\ldots+a_kx^k)]x^d=0$$

(2).Simplify

$$\begin{bmatrix}1 & x_1 & \ldots & x_1^d \\ 1 & x_2 & \ldots & x_2^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^d\end{bmatrix}\begin{bmatrix}a_0 \\ a_1 \\ \vdots \\ a_k\end{bmatrix}=\begin{bmatrix}y_1 \\ y_2 \\ \vdots \\ y_n\end{bmatrix} \tag{1}$$

(3).Formula

$$A=[(X^TX]^{-1}x^Ty$$

(4).Conclusion

The value of d: 1

Reason: the optimization model of the regression get the lowest MSE when d=1

MSE on training set

d	1	2	3	4	5
Mse	3.96e+24	4.44e+24	1.63e+25	9.31e+24	9.31e+24
d	6	7	8	9	10
Mse	9.32e+24	9.34e+24	1.08e+25	1.08e+25	1.08e+25
d	11	12	13	14	15
Mse	1.08e+25	1.16e+25	1.16e+25	1.46e+25	1.46e+25
d	16	17	18	19	20
Mse	1.46e+25	1.46e+25	1.46e+25	1.46e+25	1.96e+25

MSE on testing set

d	1	2	3	4	5
Mse	3.20e+22	1.54e+23	6.50e+24	2.47e+24	2.47e+24
d	6	7	8	9	10
Mse	4.72e+24	4.72e+24	4.73e+24	4.75e+24	4.76e+24
d	11	12	13	14	15
Mse	4.76e+24	4.76e+24	4.76e+24	5.25e+24	5.25e+24
d	16	17	18	19	20
Mse	5.25e+24	5.25e+24	5.25e+24	9.58e+24	9.58e+24

Final regression function:

$$y = -14987033392 - 20185819214 * x$$

Result visualization:

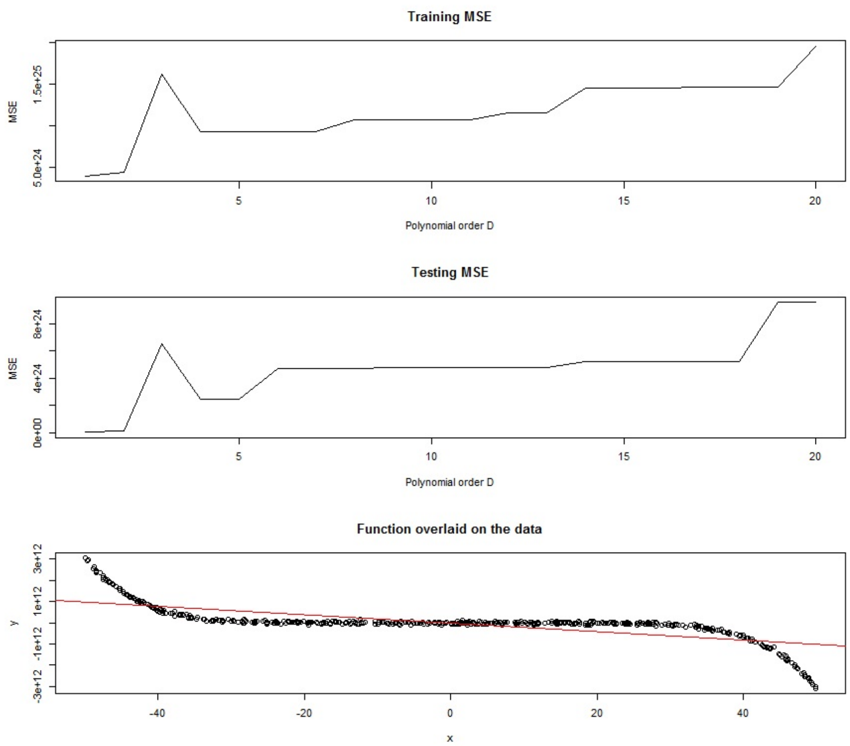


Figure 1: The plot of the training, testing error for different choices of d and the function overlaid on the data for the best choice of d on the testing set

2 Problem 2: Multivariable Ridge Regression

Cost Function:

$$\beta^{ridge} = argmin(\sum(y_i - \beta_0 - \Sigma x_{ij}\beta_j)^2 + \lambda\Sigma\beta_j^2)$$

Goal:

$$\min_{\lambda} \beta^{ridge}$$

Background:
When the x of the predictors contains severe multi-collinarity,
the $R = (x^T x)$ in the least-square formular $\hat{\beta} = (x^T x)^{-1} x^T y$ would be irreversible $(x^T x = 0)$, which will result the failure of least-square method

Objective:
Use an l_2 loss function to penalize the complexity of the model which reduce the possibility that R become singulation
Sacrificing the Unbiasness to exchange for low Variance

Formula:

$$LeastSquare : A = [(X^T X]^{-1} x^T y$$

$$\Downarrow$$

$$RidgeRegression : A = [(X^T X + \lambda I]^{-1} x^T y$$

Conclusion:
The λ that minimize the testing error is: 421
The corresponding error is: 24835.8
Discovery:
The MSE of the training set is increasing along with the increase of λ, but the MSE of the testing set is dropping down when λ increase

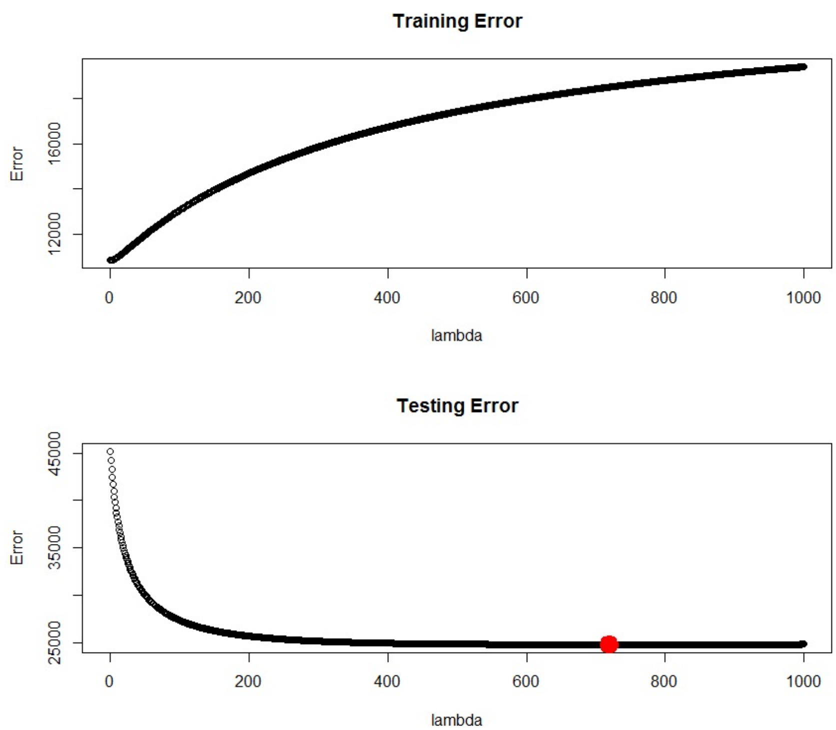


Figure 2: The plot of the training and testing error for different values of and mark the which minimizes the testing error on the data set

3 Problem 3: Sigmoid Functions

3.1

Function:

$$g(z) = \frac{1}{1+e^{-z}}$$

Objective 1:

$$g(z) = 1 - g(-z)$$

Derivation:

$$g(-z) = 1 - g(z)$$
$$\Downarrow$$
$$g(z) + g(-z) = \frac{1}{1+e^{-z}} + \frac{1}{1+e^{z}}$$
$$g(z) + g(-z) = \frac{1+e^{z}+1+e^{-z}}{(1+e^{-z})(1+e^{z})}$$
$$g(z) + g(-z) = \frac{1+e^{z}+1+e^{-z}}{1+e^{z}+1+e^{-z}}$$
$$g(z) + g(-z) = 1$$
$$\Downarrow$$
$$g(-z) = 1 - g(z)$$
$$\Downarrow$$
$$g(z) = 1 - g(-z)$$

Objective 2:

$$g^{-1}(y) = ln(\frac{y}{1-y})$$

Derivation:

$$g(z) = \frac{1}{1+e^{-z}}$$
$$\Downarrow$$

$$1 + e^{-z} = \frac{1}{g(z)}$$

$$e^{-z} = \frac{1 - g(z)}{g(z)}$$

$$-z = ln \frac{1 - g(z)}{g(z)}$$

$$z = -ln \frac{1 - g(z)}{g(z)}$$

$$z = ln \frac{g(z)}{1 - g(z)}$$

$$\Downarrow$$

$$g^{-1}(y) = ln(\frac{y}{1 - y})$$

3.2

Function:

$$x = \frac{sinh(x)}{cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Objective:

$$\frac{1 + tanh(x)}{1 - tanh(x)} = e^{2x}$$

Derivation:

$$\frac{1 + tanh(x)}{1 - tanh(x)} = \frac{1 + \frac{e^x - e^{-x}}{e^x + e^{-x}}}{1 - \frac{e^x - e^{-x}}{e^x + e^{-x}}}$$

$$= \frac{e^x + e^{-x} + (e^x - e^{-x})}{e^x + e^{-x} - (e^x - e^{-x})}$$

$$= \frac{2e^x}{2e^{-x}} = \frac{e^x}{e^{-x}} = e^{2x}$$

$$\Downarrow$$

$$\frac{1 + tanh(x)}{1 - tanh(x)} = e^{2x}$$

4 Problem 4: Logistic Regression - Gradient Descent

Hypothesis:

$$f(x;\Theta) = \frac{1}{(1+e^{-\theta^T x})}$$

Parameters:

$$x;\Theta$$

Cost Function:

$$J(\Theta) = \frac{1}{N}\sum Cost(f_\theta(x^i), y^i)$$

$$= \frac{1}{N}[\sum(y_i - 1)log(1 - f(x;\theta)) - y_i log(f(x;\theta)]$$

Goal:

$$\min_\Theta J(\Theta)$$

Method:
Gradient Descent

Algorithm:
Repeat{

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\Theta)$$

}

$$\Downarrow$$

Repeat{

$$\theta_j := \theta_j - \alpha\sum(f_\theta(x^i) - y^i)x_j^i$$

}

Conclusion:
step size $\varepsilon = 0.5$

When the tolerance $\eta = 0.01$

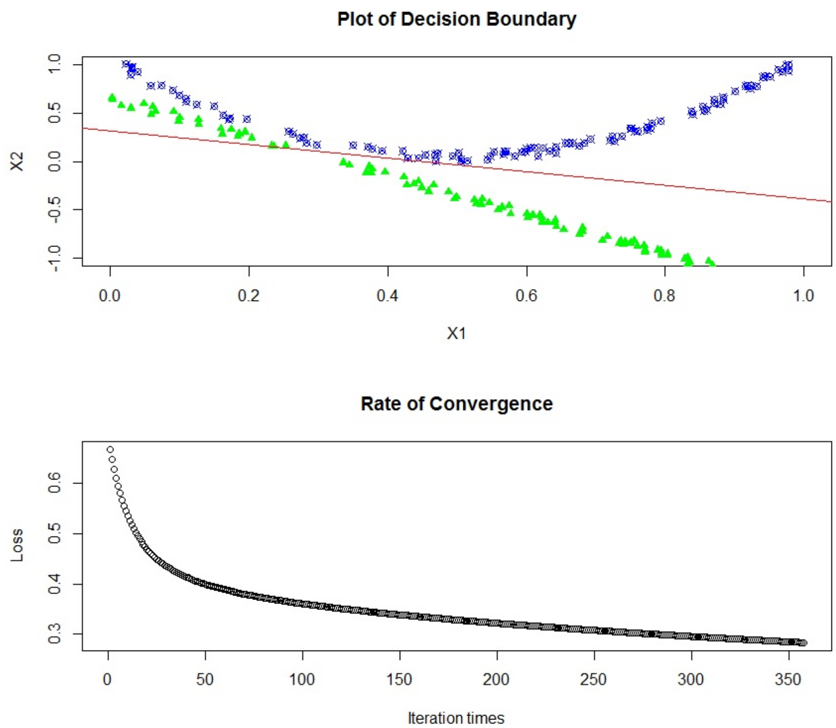


Figure 3: The plot for Gradient Descent when $\eta = 0.01$

When the tolerance $\eta = 0.001$

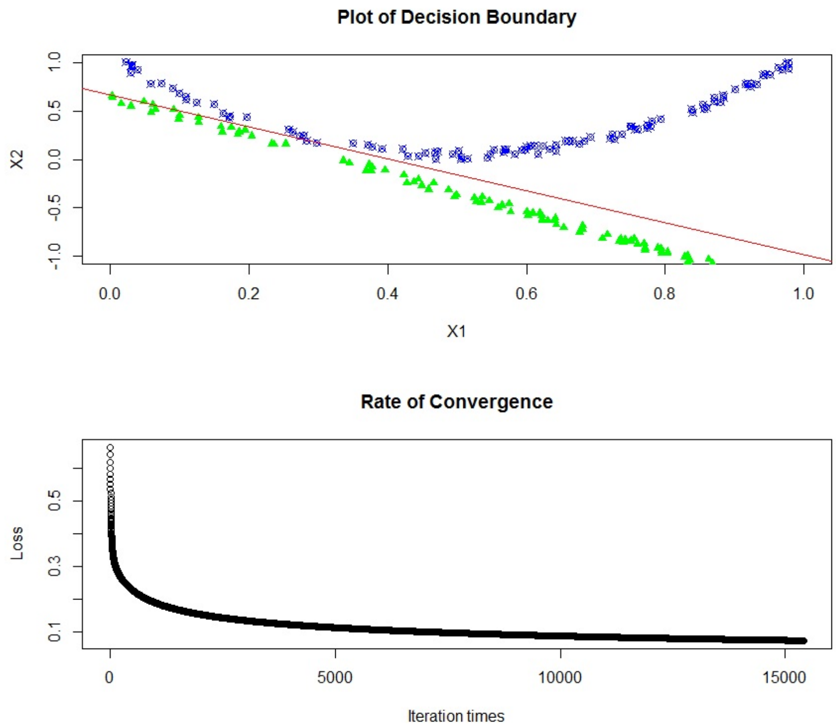


Figure 4: The plot for Gradient Descent

$\Theta = [33.52727, 19.26946, -13.72924]$
So, the regression line is: $x_2 = -1.739917x_1 + 0.7124871$

5 Problem 5: Logistic Regression - Newtons Method

Hypothesis:

$$f(x;\Theta) = \frac{1}{(1+e^{-\theta^T x})}$$

Parameters:

$$x;\Theta$$

Cost Function:

$$J(\Theta) = \frac{1}{N}\sum Cost(f_\theta(x^i), y^i)$$

$$= \frac{1}{N}[\sum(y_i - 1)log(1 - f(x;\theta)) - y_i log(f(x;\theta)]$$

Goal:

$$\min_\Theta J(\Theta)$$

Method:
Newtons Method

Algorithm:
Repeat{

$$\theta := \theta - \frac{f(\theta)}{f'(\theta)}$$

}

$$\Downarrow$$

Repeat{

$$\theta := \theta - H^{-1}\nabla_\theta J(\theta)$$

$$(H_{ij} = \frac{\partial^2 l(\theta)}{\partial\theta_i \partial\theta_j})$$

}

$$\nabla J(\theta) = \Sigma(f_{\theta)-y)x}$$

$$
\begin{aligned}
H_{J(\theta)} &= (\nabla J_\theta)' \\
&= \Sigma f'(\theta) \\
&= \Sigma(-\frac{e^{-\theta^T x}(-x)}{(1+e^{-\theta^T x})^2}x) \\
&= \Sigma x^T(\frac{1}{1+e^{-\theta^T x}})(\frac{e^{-\theta^T x}}{1+e^{-\theta^T x}})x \\
&= x^T * diag(h) * diag(1-h) * x
\end{aligned}
$$

Conclusion:
When the tolerance $\eta = 0.01$

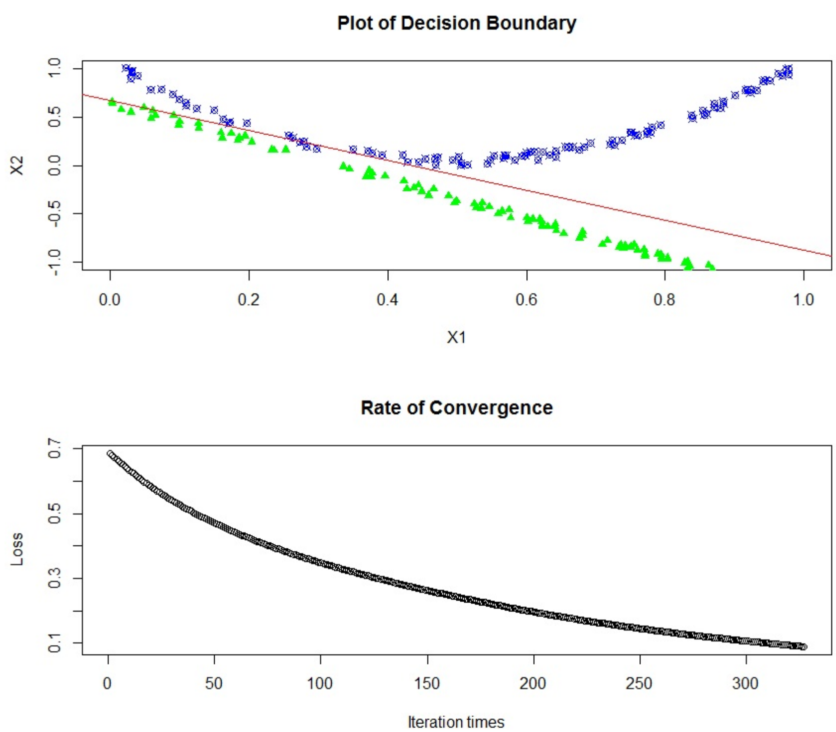
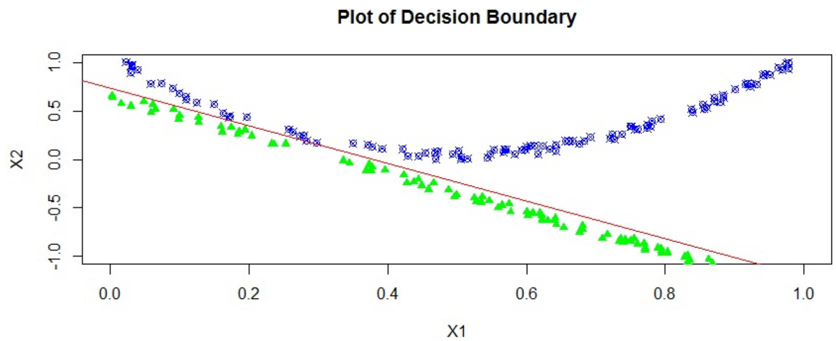


Figure 5: The plot for Newton's Method

When the tolerance $\eta = 0.001$



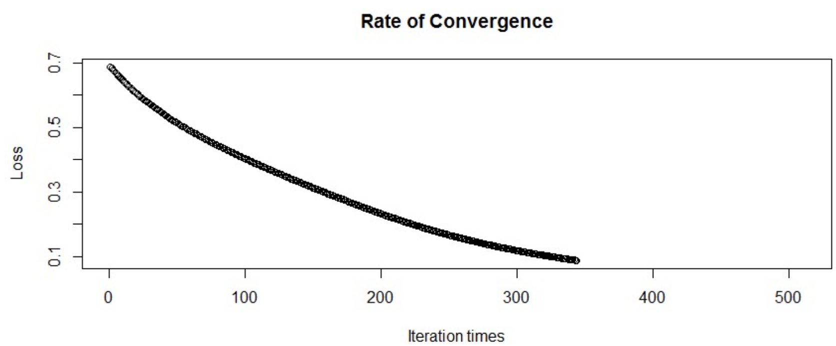


Figure 6: The plot for Newton's Method

$\Theta = [68.75510, 36.31431, -26.67597]$
So, the regression line is: $x_2 = -1.888912x_1 + 0.7353629$