ML Homework 1 Solution

Han Wang (hw2435@nyu.edu)

September 16, 2019

Problem 1

Multi-dimensional regression with D as the degree of the polynomial:

$$f(x;\theta) = \theta_0 + \theta_1 + \theta_2 x^2 + \ldots + \theta_d x^d = \theta * \mathbf{X} \tag{1-1}$$

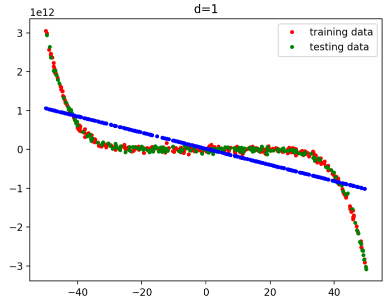
For θ minimizes the empirical risk:

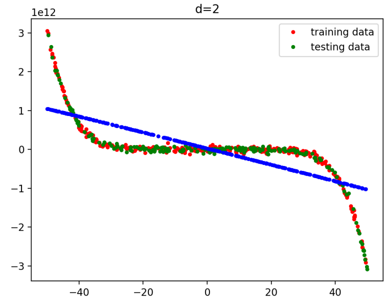
$$R_{emp}(\theta) = \frac{1}{N}\sum_{i=1}^{D}\frac{1}{2}(y_i - f(x;\theta))^2 \tag{1-2}$$

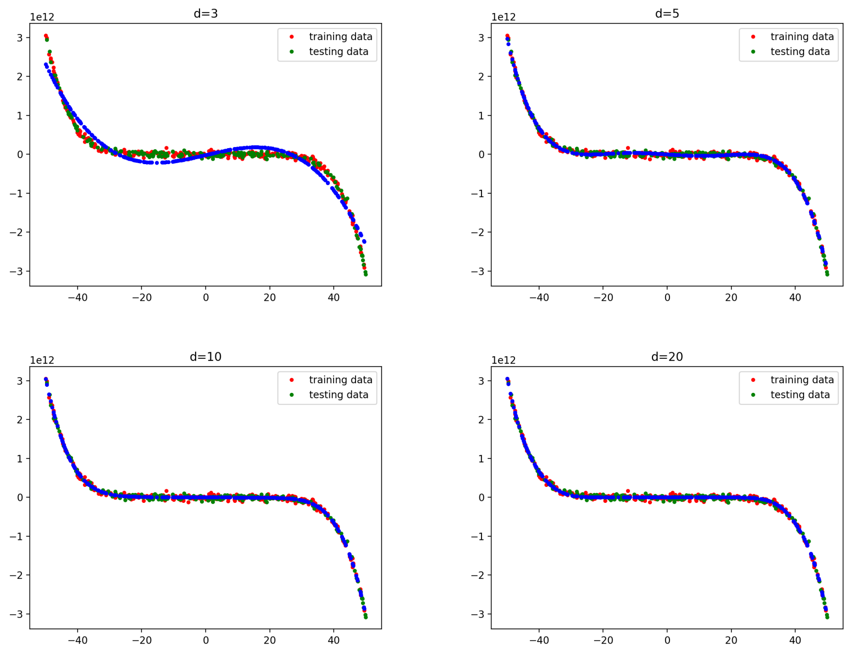
When $\nabla_\theta R(\theta) = 0$, find θ^*

$$\theta^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \tag{1-3}$$

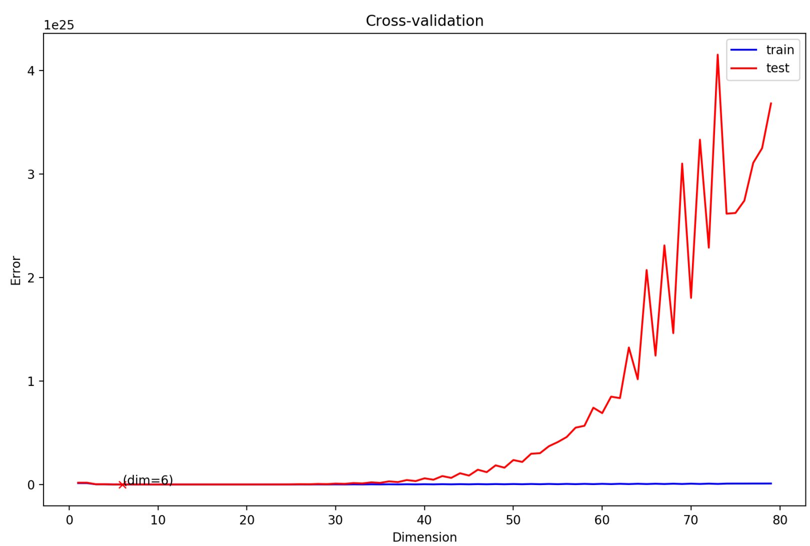
By randomly split the data into two halves, when d=(1,2,3,5,10,20), the plot is as below:







After run a cross-validation on dimension in the range of (1,80)



According to the cross-validation chart, as degree of polynomial starts to grow, the traning error basically remains low, while testing error grow with degree tremendously.

When d=6, testing error reaches its lowest, which finds us the best θ^*.

Problem 2

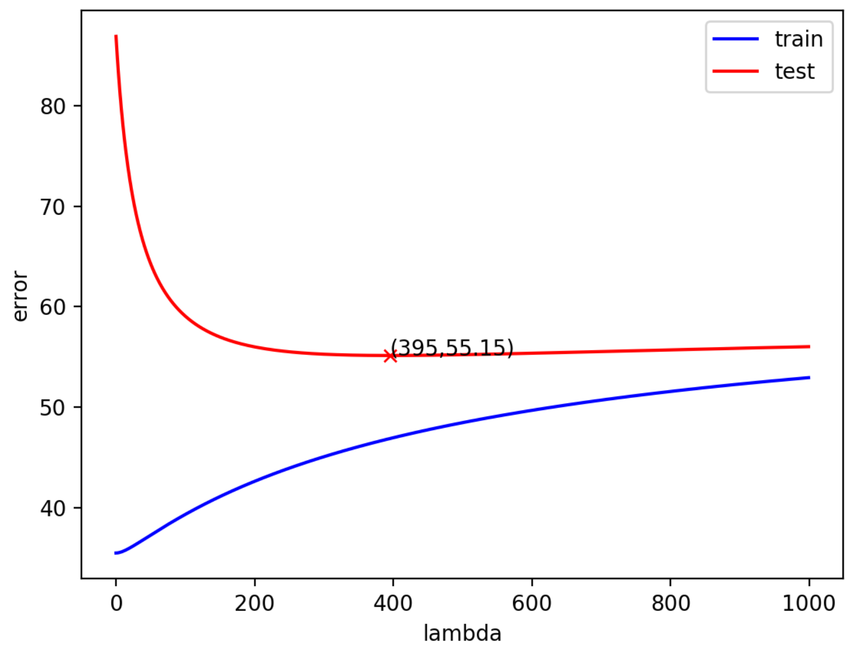
Polynomial regression with l2 regularization, which alternates the empirical risk into:

$$
\begin{aligned}
R_{reg}(\theta) &= R_{emp}(\theta) + Penalty(\theta) \\
&= \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i, \theta)) + \frac{\lambda}{2N}
\end{aligned}
\tag{2-1}
$$

When gradient=0, empirical risk reaches its lowest with θ^* as:

$$
\theta^* = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T y \tag{2-2}
$$

Apply two-fold cross-validation to find the best λ.The plot is as below:



According to the graph, when $\lambda = 395$, testing error is the lowest.

Problem 3

Given $g(z) = \frac{1}{1+e^{-z}}$, proof for the property $g(-z) = 1 - g(z)$ is as below:

$$\begin{aligned} g(-z) &= \frac{1}{1+e^{z}} \\ &= \frac{1+e^{z}-e^{z}}{1+e^{z}} \\ &= 1 - \frac{e^{z}}{1+e^{z}} \\ &= 1 - \frac{1}{\frac{1}{e^{z}}+1} \\ &= 1 - \frac{1}{e^{-z}+1} \\ &= 1 - g(z) \end{aligned} \tag{3-1}$$

Proof for the inverse of logistic function is $g^{-1}(y) = ln\frac{y}{1-y}$ is given below:

$$\begin{aligned} ln\frac{y}{1-y} &= ln\frac{\frac{1}{1+e^{-z}}}{1-\frac{1}{1+e^{-z}}} \\ &= ln\frac{\frac{1}{1+e^{-z}}}{\frac{1+e^{-z}}{1+e^{-z}}-\frac{1}{1+e^{-z}}} \\ &= ln\frac{1}{1+e^{-z}-1} \\ &= ln\frac{1}{e^{-z}} \\ &= ln(e^{z}) \\ &= z \end{aligned} \tag{3-2}$$

Thus, $g^{-1}(g(z)) = z$.

Problem 4

To minimize the empirical risk with logistic loss of logistic regression:

$$R_{emp}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - 1)log(1 - f(x_i;\theta)) - y_i log(f(x_i;\theta)). \tag{4-1}$$

The gradient of R_{emp} is:

$$\begin{aligned}
\nabla_\theta R(\theta) &= -\frac{1}{N}\sum_{i=1}^{N}(y_i\frac{1}{f(x_i;\theta)}\frac{\partial}{\partial\theta}f(x_i;\theta) - (1-y_i)\frac{1}{1-f(x_i;\theta)}\frac{\partial}{\partial\theta}f(x_i;\theta)) \\
&= -\frac{1}{N}\sum_{i=1}^{N}(y_i\frac{1}{f(x_i;\theta)} - (1-y_i)\frac{1}{1-f(x_i;\theta)})\frac{\partial}{\partial\theta}f(x_i;\theta)) \\
&= -\frac{1}{N}\sum_{i=1}^{N}(y_i\frac{1}{g(\theta^T\mathbf{x})} - (1-y_i)\frac{1}{1-g(\theta^T\mathbf{x})})\frac{\partial}{\partial\theta}g(\theta^T\mathbf{x})) \\
&= -\frac{1}{N}\sum_{i=1}^{N}(y_i\frac{1}{g(\theta^T\mathbf{x})} - (1-y_i)\frac{1}{1-g(\theta^T\mathbf{x})})g(\theta^T\mathbf{x})(1-g(\theta^T\mathbf{x}))\frac{\partial}{\partial\theta}\theta^T\mathbf{x}) \\
&= -\frac{1}{N}\sum_{i=1}^{N}(y_i\frac{1}{g(\theta^T\mathbf{x})} - (1-y_i)\frac{1}{1-g(\theta^T\mathbf{x})})g(\theta^T\mathbf{x})(1-g(\theta^T\mathbf{x}))\mathbf{x}) \\
&= -\frac{1}{N}\sum_{i=1}^{N}(y_i(1-g(\theta^T\mathbf{x}) - (1-y_i)g(\theta^T\mathbf{x}))\mathbf{x}) \\
&= -\frac{1}{N}\sum_{i=1}^{N}(y - g(\theta^T\mathbf{x}))\mathbf{x}
\end{aligned} \tag{4-2}$$

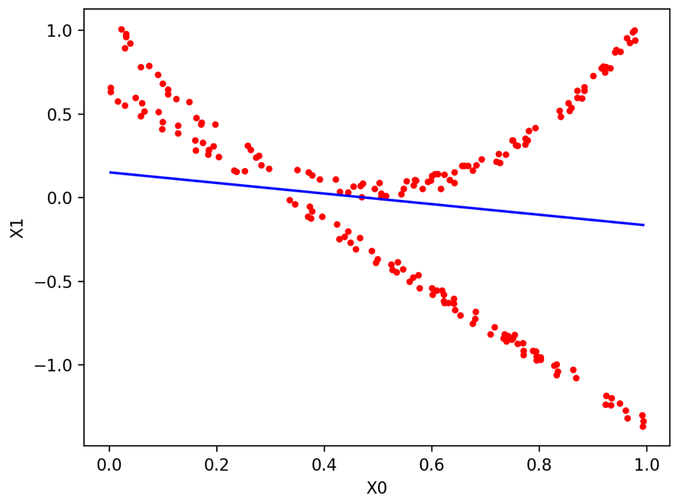
To minimize the gradient, apply batch gradient descent to solve θ^*. While θ^0 is initiated as a small random parameter matrix. For each iteration,

$$\theta^{t+1} = \theta^t - \eta\nabla_\theta R(\theta^t) \tag{4-3}$$

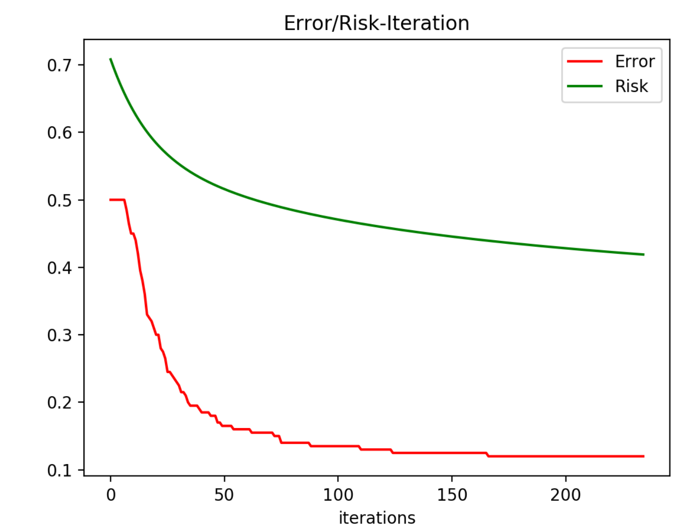
Where η denotes the learning rate of the gradient descent algorithm.

Iterations stops when the decrement $\theta^t - \theta^{t-1}$ is small than the tolerance ϵ. In this case, Both learning rate η and tolerance value ϵ are hyperparameters.

When set $\eta = 0.1, \epsilon = 0.005$, the logistic regression model takes 235 iterations until convergence, its final accuracy is 80%. The graph showing the decision boundary is :

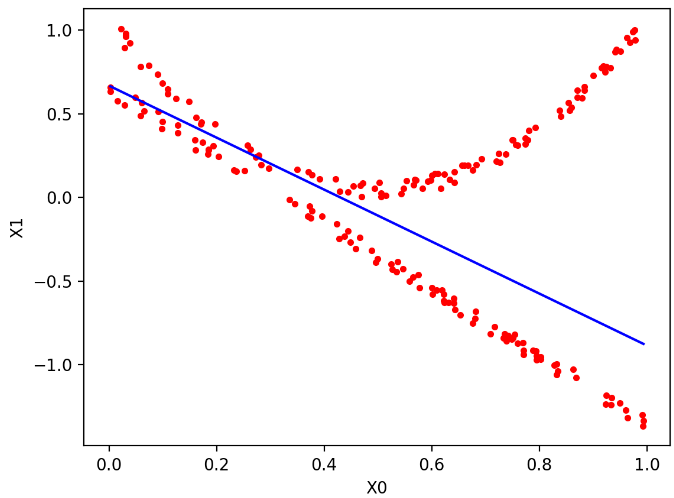


The graph plotting the binary error and the empirical risk along iterations is as below:

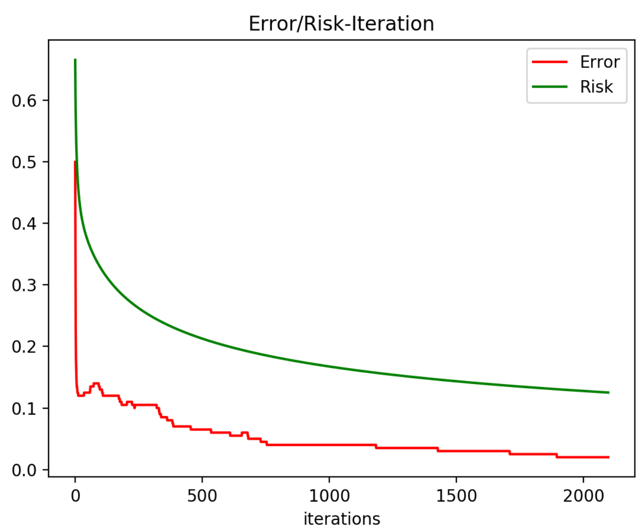


Where the performance is quite pool, for the current learning rate is small to tolerance so that iteration stops easily.

When set $\eta = 1, \epsilon = 0.005$, it takes 2098 iterations until convergence at the accuracy of 98%. The graph showing the decision boundary is as below:

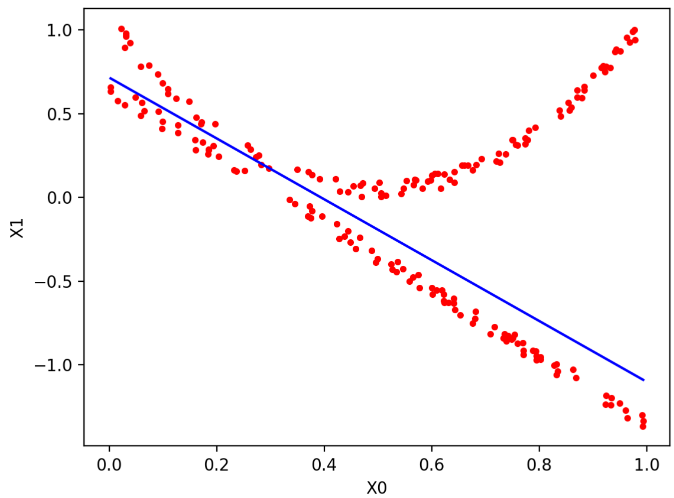


The graph plotting the binary error and the empirical risk along iterations is as below:

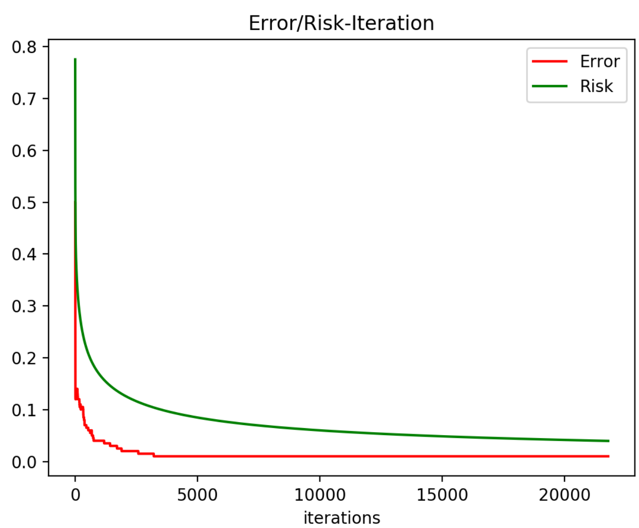


Where the total accuracy is good, though the decision boundary may not seem perfect.

When set $\eta = 1, \epsilon = 0.001$, it takes 21768 iterations until convergence at the accuracy of 100%. The graph showing the decision boundary is as below:



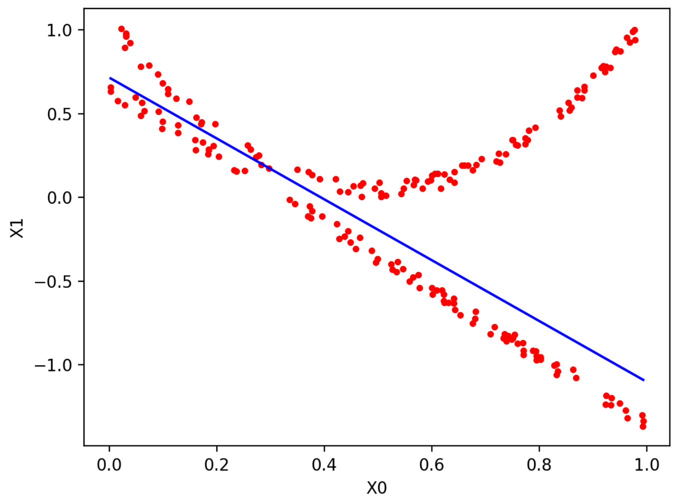
The graph plotting the binary error and the empirical risk along iterations is as below:



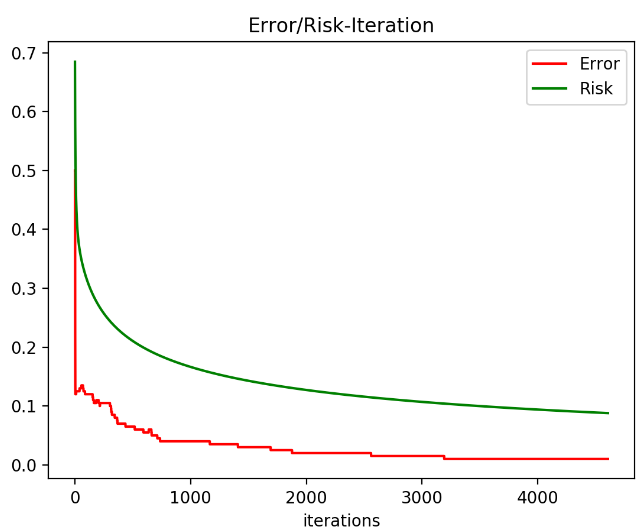
Where the total accuracy is excellent. But the error reaches zero way before convergence, the tolerance might be set a little bit too low.

According to the graph, for this learning rate, it reaches zero error after roughly around 4000 iterations.

When set $\eta = 1, \epsilon = 0.003$, it takes 4608 iterations until convergence at the accuracy of 100%. The graph showing the decision boundary is as below:



The graph plotting the binary error and the empirical risk along iterations is as below:



Where the total performance is excellent, iteration stops just after binay error gets to the value zero.