Implementation Assignment 2

Rong Yu and Finn Womack

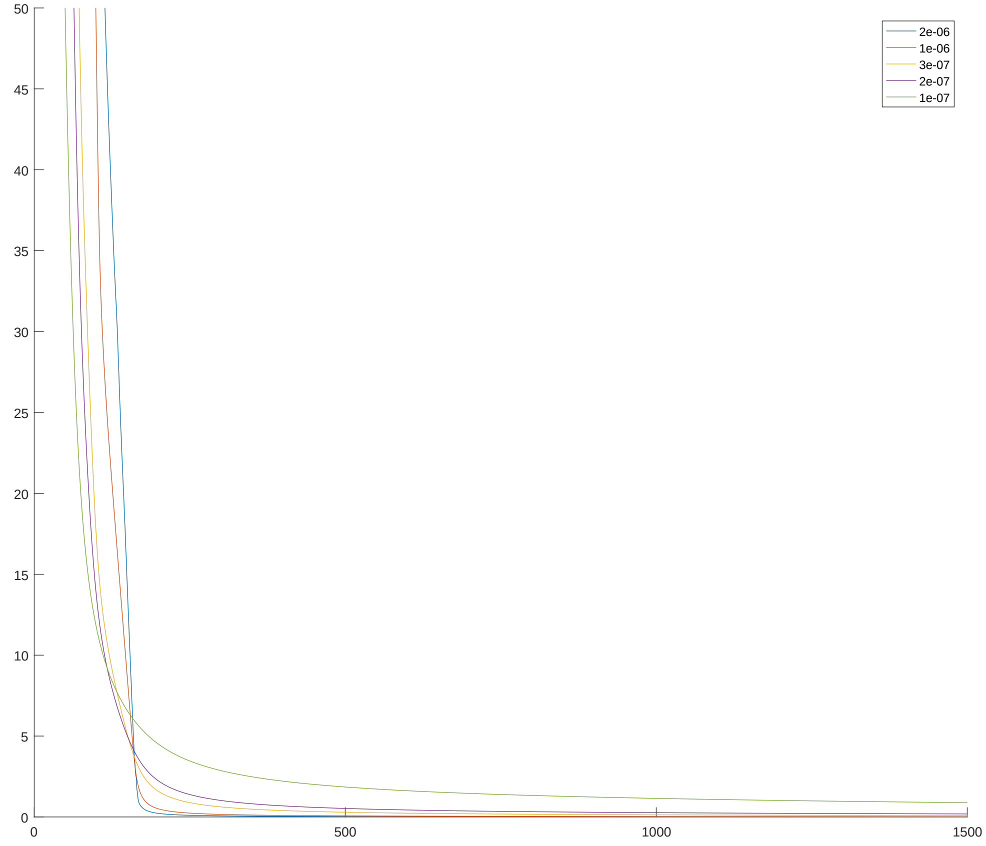
April 21, 2017

Part 1: Loading Data and Implementing Gradient Batch

After loading the testing and training data into feature and response matrices we implemented the batch gradient decent algorithm with the following learning rates:

$$Rates = \begin{bmatrix} 2 \cdot 10^{-6} \\ 1 \cdot 10^{-6} \\ 3 \cdot 10^{-7} \\ 2 \cdot 10^{-7} \\ 1 \cdot 10^{-7} \end{bmatrix} \tag{1}$$

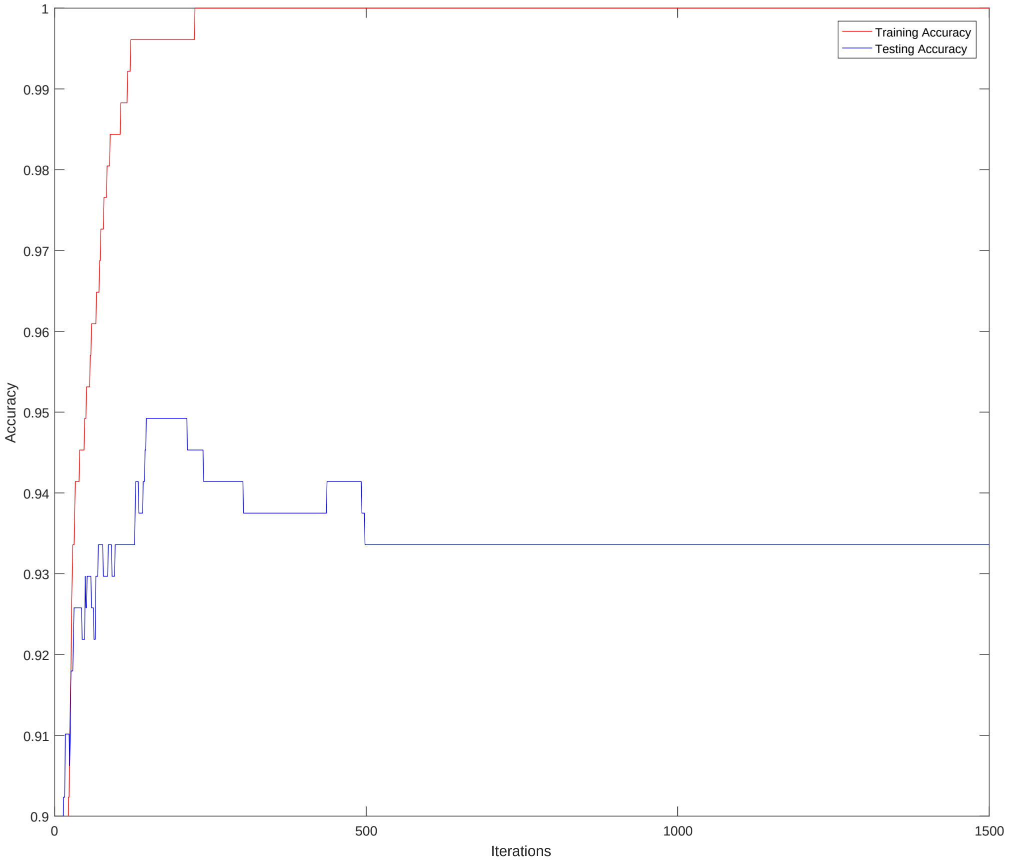
Then after running the algorithm on the above rates we plotted the iterations onto the loss function to get a sense of the convergence rate for the different learning rates:



The learning rate of $2 \cdot 10^{-6}$ gives the fastest convergence. When we picked larger learning rates we started to get oscillations.

Part 2: Testing and Training Accuracies

We then selected the learning rate of $1 \cdot 10^{-7}$ to test the effects of iterations on the testing and training accuracies:



As we can see the training accuracy converges to 1 and the testing accuracy peaks at around 175 iterations and then decreases.

Part 3: Deriving the Regularization Term

Consider the new objective function:

$$L(w) = \sum_{i=1}^{n} l(w^T x_i, y_i) + \frac{\lambda}{2}\|W\|_2^2$$

This is the the same as the original objective function with an added term and since the gradient of a sum of functions in the sum of the gradients all we need to do is find the gradient of $\frac{\lambda}{2}\|W\|_2^2$:

$$
\begin{aligned}
\nabla\frac{\lambda}{2}\|W\|_2^2 &= \frac{\lambda}{2}\nabla\|W\|_2^2 && (2)\\
&= \frac{\lambda}{2}\nabla\sum_{i=1}^{m} w_i^2 && (3)\\
&= \frac{\lambda}{2}\begin{bmatrix} \frac{\partial}{\partial w_1}\sum_{i=1}^{m} w_i^2 \\ \frac{\partial}{\partial w_2}\sum_{i=1}^{m} w_i^2 \\ \vdots \\ \frac{\partial}{\partial w_m}\sum_{i=1}^{m} w_i^2 \end{bmatrix} && (4)\\
&= \frac{\lambda}{2}\begin{bmatrix} 2w_1 \\ 2w_2 \\ \vdots \\ 2w_m \end{bmatrix} && (5)\\
&= \lambda W && (6)
\end{aligned}
$$

Thus the new batch gradient code would be as follows:

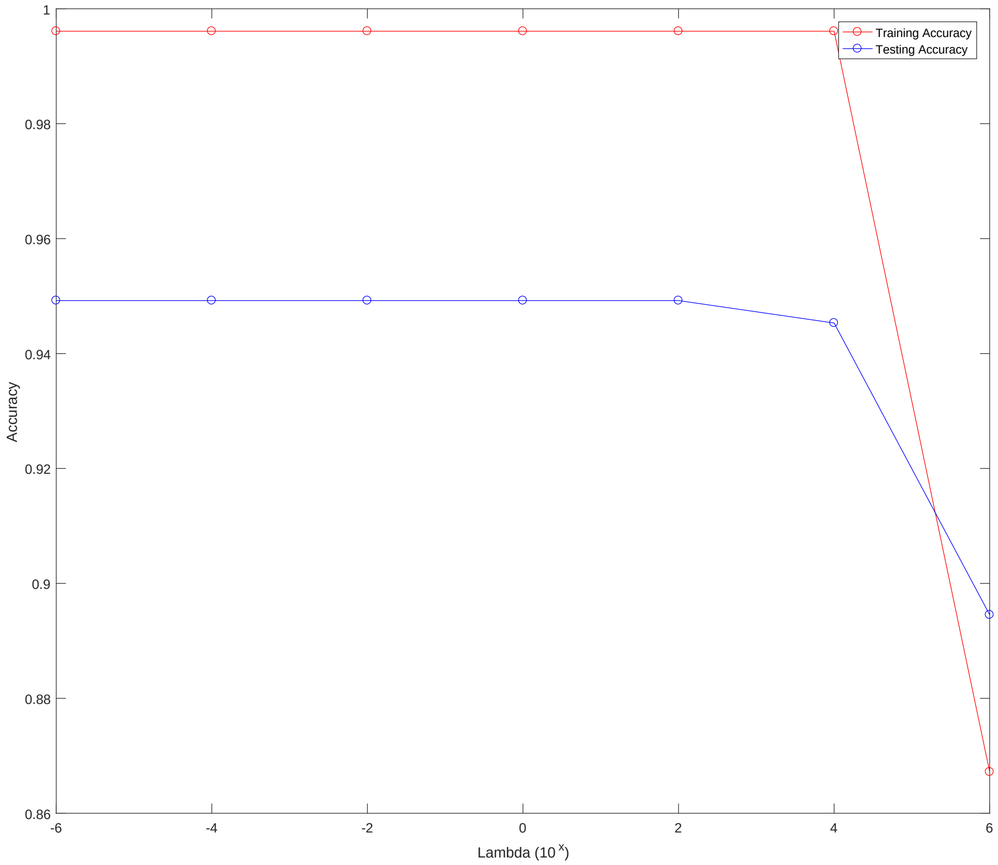
```
Given: training examples $(x_i, y_i), i = 1, ..., N$
$W \leftarrow [0, 0, ..., 0]$
Repeat until convergence:
    $d \leftarrow [0, 0, ..., 0]$
    For i = 1 to N do
        $\hat{y}_i \leftarrow \frac{1}{1+e^{-w \cdot x_i}}$
        $error = y_i - \hat{y}_i$
        $d = d + error \cdot (x_i + 2w)$
    $w \leftarrow w + \eta d$
```

Part 4: Implementing Regularization

To examine the effect of regularization on the algorithm we plotted the testing and training accuracies against the following choices of λ:

$$
\lambda \begin{bmatrix} 10^{-6} \\ 10^{-4} \\ 10^{-2} \\ 1 \\ 10^{2} \\ 10^{4} \\ 10^{6} \end{bmatrix} \qquad (7)
$$

The resulting plot is as follows:


1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
Accuracy
-6
-4
-2
0
2
4
6
Lambda (10^x)
Training Accuracy
Testing Accuracy