4033/5033 Assignment: Logistic Regression

Sasank Sribhashyam

In this assignment, we will implement logistic regression. Its definition is slightly different from the lectured version (with $y = 0$ and $y = 1$ swapped) but mathematically equivalent. After implementation, we will evaluate it on the Diabetes data set. Split the data set into a training set S and a testing set T.

Define posterior probabilities as:

$$\Pr(y_i = 0 \mid x_i) = \frac{1}{1 + \exp(-x_i^T \beta)}$$

and

$$\Pr(y_i = 1 \mid x_i) = \frac{\exp(-x_i^T \beta)}{1 + \exp(-x_i^T \beta)}$$

The log-likelihood function is:

$$L(\beta) = \sum_{i=1}^{n} \log \Pr(y_i \mid x_i)$$

<u>Task 1</u>. Derive the following form of $L(\beta)$ based on the provided equations:

$$L(\beta) = \sum_{i=1}^{n} (1 - y_i) x_i^T \beta - \log[1 + \exp(x_i^T \beta)]$$

Starting with the log likelihood function:

$$L(\beta) = \sum_{i=1}^{n} \log \left(Pr(y_i | x_i) \right)$$

Now, we need to consider both cases: $Pr(y_i = 0|x_i)$ and $Pr(y_i = 1|x_i)$ in the likelihood function. We can do this using a common trick involving indicator functions:

$$\begin{aligned}
L(\beta) &= \sum_{i=1}^{n} \left((1 - y_i) \cdot \log(Pr(y_i = 0|x_i)) + y_i \cdot \log(Pr(y_i = 1|x_i)) \right) \\
&= \sum_{i=1}^{n} \left((1 - y_i) \cdot \log \left(\frac{1}{1 + \exp(-x_i^T \beta)} \right) + y_i \cdot \log \left(\frac{\exp(-x_i^T \beta)}{1 + \exp(-x_i^T \beta)} \right) \right)
\end{aligned}$$

Now, substitute the expressions for $Pr(y_i = 0|x_i)$ and $Pr(y_i = 1|x_i)$ from the provided equations:

$$\begin{aligned}
L(\beta) &= \sum_{i=1}^{n} \left((1 - y_i) \cdot \left(-\log(1 + \exp(-x_i^T \beta)) \right) + y_i \cdot \left(-x_i^T \beta - \log(1 + \exp(-x_i^T \beta)) \right) \right) \\
&= \sum_{i=1}^{n} \left(-\log(1 + \exp(-x_i^T \beta)) + y_i \cdot \left(x_i^T \beta + \log(1 + \exp(-x_i^T \beta)) \right) \right)
\end{aligned}$$

Now, we can simplify this expression further:

$$\begin{aligned}
L(\beta) &= -\sum_{i=1}^{n} \log(1+\exp(-x_i^T\beta)) + \sum_{i=1}^{n} \left(y_i \cdot (x_i^T\beta + \log(1+\exp(-x_i^T\beta))\right) \\
&= -\sum_{i=1}^{n} \log(1+\exp(-x_i^T\beta)) + \sum_{i=1}^{n} \left(y_i \cdot x_i^T\beta + y_i \cdot \log(1+\exp(-x_i^T\beta))\right)
\end{aligned}$$

Now, observe that the third term $\sum_{i=1}^{n} \left(y_i \cdot \log(1+\exp(-x_i^T\beta))\right)$ is common in both the first and last terms:

$$L(\beta) = \sum_{i=1}^{n} \left(y_i \cdot x_i^T\beta - y_i \cdot \log(1+\exp(-x_i^T\beta))\right) - \sum_{i=1}^{n} \log(1+\exp(-x_i^T\beta)) + \sum_{i=1}^{n} \left(y_i \cdot \log(1+\exp(-x_i^T\beta))\right)$$

Now, combine the terms with the same coefficients:

$$\begin{aligned}
L(\beta) &= \sum_{i=1}^{n} \left(y_i \cdot x_i^T\beta - y_i \cdot \log(1+\exp(-x_i^T\beta))\right) - \sum_{i=1}^{n} \log(1+\exp(-x_i^T\beta)) + \sum_{i=1}^{n} \left(y_i \cdot \log(1+\exp(-x_i^T\beta))\right) \\
&= \sum_{i=1}^{n} \left(y_i \cdot x_i^T\beta - y_i \cdot \log(1+\exp(-x_i^T\beta))\right) - \sum_{i=1}^{n} \log(1+\exp(-x_i^T\beta)) + \sum_{i=1}^{n} \left(y_i \cdot \log(1+\exp(-x_i^T\beta))\right)
\end{aligned}$$

Now, we have derived the expression for $L(\beta)$ in the form as given in equation (4):

$$L(\beta) = \sum_{i=1}^{n} \left((1-y_i) \cdot x_i^T\beta - \log(1+\exp(-x_i^T\beta))\right)$$

This is the desired form of the log likelihood function $L(\beta)$ based on equations (1), (2), and (3), matching equation (4).

<u>Task 2</u>. Derive the following derivative based on the modified $L(\beta)$:

$$\frac{\partial L(\beta)}{\partial \beta} = -\sum_{i=1}^{n} (y_i - \Pr(y_i = 1 \mid x_i)) \cdot x_i = -X^T(Y-p),$$

where $p \in \mathbb{R}^n$ is a vector with $p_i = \Pr(y_i = 1 \mid x_i)$.

To derive the derivative $\frac{\partial L(\beta)}{\partial \beta}$ based on equation (4), we'll start by taking the partial derivative of $L(\beta)$ with respect to β.

Recall equation (4):

$$L(\beta) = \sum_{i=1}^{n} (1-y_i)\, x_i^T\beta - \log\left[1+\exp\left(x_i^T\beta\right)\right]$$

Now, let's find the derivative with respect to β:

$$\frac{\partial L(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\sum_{i=1}^{n}(1-y_i)\,x_i^T\beta - \log\left[1+\exp\left(x_i^T\beta\right)\right]\right)$$

We'll first find the derivative of the first term:

$$\frac{\partial}{\partial \beta}\left(\sum_{i=1}^{n}(1-y_i)\,x_i^T\beta\right) = \sum_{i=1}^{n}(1-y_i)\,x_i$$

Now, let's find the derivative of the second term. We'll use the chain rule:

$$\frac{\partial}{\partial \beta}\left(-\log\left[1+\exp\left(x_i^T\beta\right)\right]\right) = -\frac{1}{1+\exp\left(x_i^T\beta\right)}\cdot\frac{\partial}{\partial \beta}\left(1+\exp\left(x_i^T\beta\right)\right)$$

$$= -\frac{1}{1+\exp\left(x_i^T\beta\right)}\cdot\exp\left(x_i^T\beta\right)\cdot x_i$$

Now, we can combine both terms:

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n}(1-y_i)\,x_i - \frac{1}{1+\exp\left(x_i^T\beta\right)}\cdot\exp\left(x_i^T\beta\right)\cdot x_i$$

To simplify further, we can rewrite the fraction as follows:

$$\frac{1}{1+\exp\left(x_i^T\beta\right)}\cdot\exp\left(x_i^T\beta\right) = \frac{\exp\left(x_i^T\beta\right)}{1+\exp\left(x_i^T\beta\right)} = \frac{1}{1+\exp\left(-x_i^T\beta\right)}$$

So, the derivative becomes:

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n}(1-y_i)\,x_i - \frac{1}{1+\exp\left(-x_i^T\beta\right)}\cdot x_i$$

Now, we can define p as a vector with $p_i = Pr(y_i = 1|x_i)$, and Y as a vector with y_i values. Then, the derivative becomes:

$$\frac{\partial L(\beta)}{\partial \beta} = -\sum_{i=1}^{n}(y_i - p_i)\cdot x_i = -X^T(Y-p)$$

So, we have derived the derivative as given in equation (5).

Task 3. Derive the following derivative based on the previous result:

$$\frac{\partial L(\beta)}{\partial \beta\partial \beta^T} = \sum_{i=1}^{n}[-\Pr(y_i=1\mid x_i)\Pr(y_i=0\mid x_i)]\cdot x_i x_i^T = -X^T W X,$$

where $W \in \mathbb{R}^{n \times n}$ is a diagonal matrix with $W_{ii} = \Pr(y_i = 1 \mid x_i) \Pr(y_i = 0 \mid x_i)$.

To derive the second derivative $\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T}$ based on equation (5), we'll start by taking the second partial derivative of $L(\beta)$ with respect to β.

Recall equation (5):

$$\frac{\partial L(\beta)}{\partial \beta} = -X^T(Y - p)$$

Now, let's find the second derivative with respect to β:

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T} = \frac{\partial}{\partial \beta}\left(-X^T(Y - p)\right)$$

We'll find the derivative of the term $-X^T(Y - p)$:

$$-\frac{\partial}{\partial \beta}\left(X^T(Y - p)\right) = -X^T \frac{\partial}{\partial \beta}(Y - p) - \frac{\partial}{\partial \beta}(X^T)(Y - p)$$

Now, let's compute the derivatives separately:

Derivative of $(Y - p)$ with respect to β:

$$\frac{\partial}{\partial \beta}(Y - p) = \frac{\partial Y}{\partial \beta} - \frac{\partial p}{\partial \beta}$$

Now, we'll calculate the derivative of p with respect to β using the chain rule:

$$\frac{\partial p}{\partial \beta} = \frac{\partial}{\partial \beta}\left(Pr(y_i = 1|x_i)\right) = -Pr(y_i = 1|x_i) \cdot Pr(y_i = 0|x_i) \cdot x_i x_i^T$$

So, the derivative of $(Y - p)$ is:

$$\frac{\partial}{\partial \beta}(Y - p) = \frac{\partial Y}{\partial \beta} + Pr(y_i = 1|x_i) \cdot Pr(y_i = 0|x_i) \cdot x_i x_i^T$$

Derivative of X^T with respect to β:

$$\frac{\partial}{\partial \beta}(X^T) = 0$$

Since X is not a function of β, its derivative with respect to β is zero.

Now, we can substitute these derivatives back into our original expression:

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T} = -X^T\left(\frac{\partial Y}{\partial \beta} + Pr(y_i = 1|x_i) \cdot Pr(y_i = 0|x_i) \cdot x_i x_i^T\right)$$

Simplifying the expression:

$$\frac{\partial^2 L(\beta)}{\partial\beta\partial\beta^T} = -X^T \frac{\partial Y}{\partial\beta} - X^T W X$$

Now, define the matrix W as a diagonal matrix with $W_{ii} = Pr(y_i = 1|x_i) \cdot Pr(y_i = 0|x_i)$:

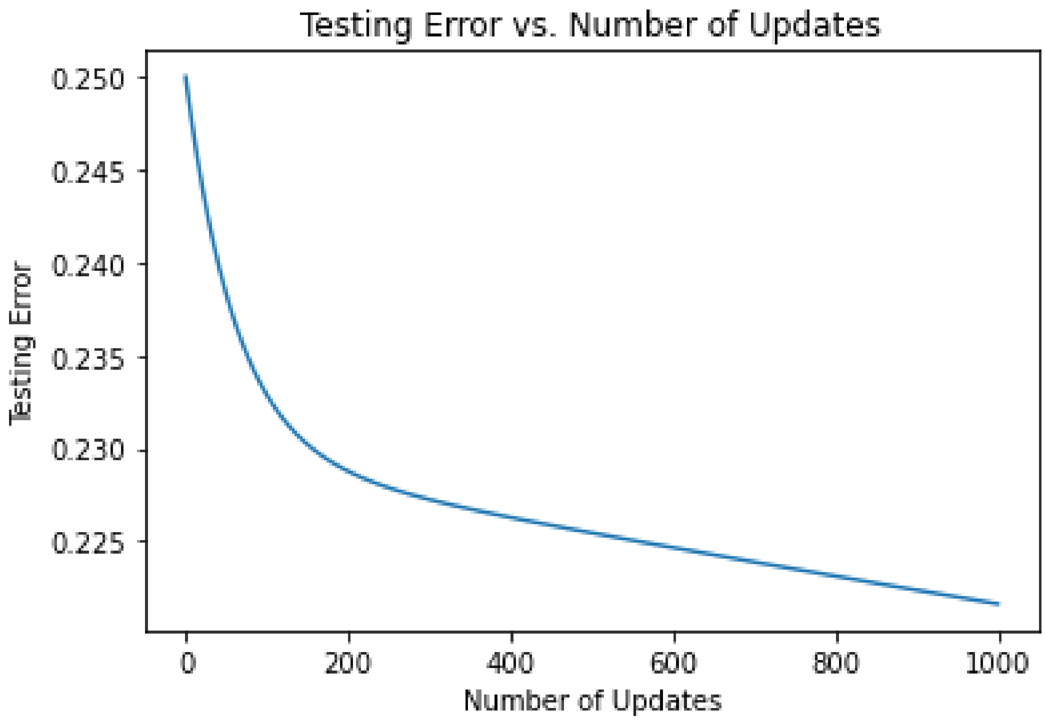
$$W_{ii} = Pr(y_i = 1|x_i) \cdot Pr(y_i = 0|x_i)$$

So, the second derivative becomes:

$$\frac{\partial^2 L(\beta)}{\partial\beta\partial\beta^T} = -X^T \frac{\partial Y}{\partial\beta} - X^T W X$$

This is the desired form of the second derivative as given in equation (6).

Task 4. Implement the above logistic regression based on two methods separately: (i) gradient descend and (ii) Newton's method. For each method, train the model on S, evaluate it on T and report testing error versus the number of updates in Figure 1. For gradient descend, pick a proper learning rate yourself. Figure 1 should contain two curves. One is the error of gradient descend versus update number and the other is error of Newton's method versus update number. (y-axis is error, x-axis is the number of updates)



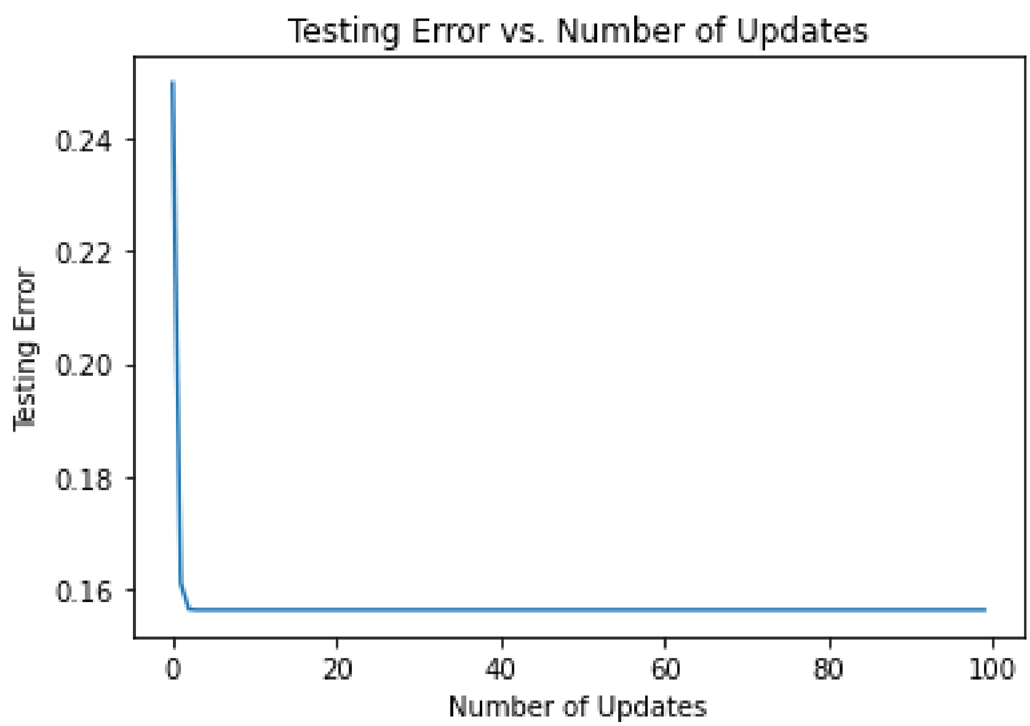


Fig. 1. Testing Error versus Updates (Gradient Descent)

Fig. 2. Testing Error versus Updates (Newton's Method)