Obtaining the Gradient for Logistic Regression

1 Definitions for Logistic Regression

These are the four equations we'll use to completely define *Logistic Regression* and the cost function this algorithm should minimize. You should watch the theory videos for an explanation of what each of these equations mean.

The notation here is slightly different than in the videos, so make sure you understand these equations and where they come from before you continue.

$$J(\mathbf{w}, b) = \frac{1}{m}\sum_{i=1}^{m} L^{(i)}$$

$$L^{(i)} = -y^{(i)} log(\hat{y}^{(i)}) - (1 - y^{(i)}) log(1 - \hat{y}^{(i)})$$

$$\hat{y}^{(i)} = \frac{1}{1 + e^{-z^{(i)}}}$$

$$z^{(i)} = \mathbf{x}^{(i)} \cdot \mathbf{w} + b$$

2 Decomposition of the w derivative

Taking the partial derivative of this whole set of equations (all four of which are needed to fully define the cost function J) w.r.t. a specific parameter w_j might seem like a daunting task. However, there are several steps we can take to make taking this derivative a lot more doable.

First, we'll move the derivative operator inside the summation (sum and constant factor rules) and then see that these functions applications can be decomposed into several smaller partial derivatives using the **chain rule**.

$$\begin{aligned}
\frac{\partial J}{\partial w_j} &= \frac{\partial}{\partial w_j}\frac{1}{m}\sum_{i=1}^{m} L^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial w_j} L^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m} \frac{\partial L^{(i)}}{\partial \hat{y}^{(i)}} \frac{\partial \hat{y}^{(i)}}{\partial w_j} \\
&= \frac{1}{m}\sum_{i=1}^{m} \frac{\partial L^{(i)}}{\partial \hat{y}^{(i)}} \frac{\partial \hat{y}^{(i)}}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial w_j}
\end{aligned}$$

3 Cost function term

The first partial derivative from the decomposition is to take the derivative of one loss term $L^{(i)}$, i.e. the part of the cost caused by the i^{th} training example, w.r.t. $\hat{y}^{(i)}$, which is the model prediction for the i^{th} training example. Even if this prediction is the result of several function applications, we can just treat that result as a new partial derivative variable.

The steps of this derivative have been omitted here, as they will be part of the assignment for this week.

$$\begin{aligned}\frac{\partial L^{(i)}}{\partial \hat{y}^{(i)}} &= \frac{\partial}{\partial \hat{y}^{(i)}} - y^{(i)} log(\hat{y}^{(i)}) - (1 - y^{(i)}) log(1 - \hat{y}^{(i)}) \\ &= \ldots \\ &= \frac{-y^{(i)}}{\hat{y}^{(i)}} + \frac{1 - y^{(i)}}{1 - \hat{y}^{(i)}}\end{aligned}$$

4 Sigmoid activation function

This is definitely the most complicated step of the whole derivative. Not only is taking the derivative of the *Sigmoid* function a little more involved, but we also need to rewrite this derivative to be purely expressed in terms of $\hat{y}^{(i)}$.

The fact that the derivative of the Sigmoid function can be described by the function output, instead of just the function input, is one of the reasons the Sigmoid function is so useful to use as an activation function. This rewrite step will make the whole expression a lot nicer at the end.

You're not required to completely understand all of the steps here, as some of them are a bit tricky. This part is included here for completeness and only as optional material.

4.1 Taking the derivative

$$\begin{aligned}\frac{\partial \hat{y}^{(i)}}{\partial z^{(i)}} &= \frac{\partial}{\partial z^{(i)}} \frac{1}{1 + e^{-z^{(i)}}} \\ &= \frac{0\ (1 + e^{-z^{(i)}}) - 1\ (\frac{\partial}{\partial z^{(i)}} 1 + e^{-z^{(i)}})}{(1 + e^{-z^{(i)}})^2} \\ &= \frac{-(0 + e^{-z^{(i)}}(-1))}{(1 + e^{-z^{(i)}})^2} \\ &= \frac{e^{-z^{(i)}}}{(1 + e^{-z^{(i)}})^2}\end{aligned}$$

4.2 Rewriting into terms of $\hat{y}^{(i)}$

$$\begin{aligned}
\frac{\partial \hat{y}^{(i)}}{\partial z^{(i)}} &= \frac{e^{-z^{(i)}}}{(1+e^{-z^{(i)}})^2} \\
&= \frac{1}{1+e^{-z^{(i)}}} \; \frac{e^{-z^{(i)}}}{1+e^{-z^{(i)}}} \\
&= \hat{y}^{(i)} \; \frac{e^{-z^{(i)}}}{1+e^{-z^{(i)}}} \\
&= \hat{y}^{(i)} \; \frac{e^{-z^{(i)}} + (1-1)}{1+e^{-z^{(i)}}} \\
&= \hat{y}^{(i)} \; \frac{1+e^{-z^{(i)}} - 1}{1+e^{-z^{(i)}}} \\
&= \hat{y}^{(i)} \; \left(\frac{1+e^{-z^{(i)}}}{1+e^{-z^{(i)}}} - \frac{1}{1+e^{-z^{(i)}}} \right) \\
&= \hat{y}^{(i)} \; \left(1 - \frac{1}{1+e^{-z^{(i)}}} \right) \\
&= \hat{y}^{(i)} \; (1-\hat{y}^{(i)})
\end{aligned}$$

5 Linear input function

This is the same partial derivative we've already taken for linear regression, in a slightly different notation.

It is just the partial derivative of the dot product between the input vector $\mathbf{x}^{(i)}$ and the parameter vector $\mathbf{w}$ plus the bias b w.r.t. a specific parameter w_j.

$$\begin{aligned}
\frac{\partial z^{(i)}}{\partial w_j} &= \frac{\partial}{\partial w_j} \mathbf{x}^{(i)} \cdot \mathbf{w} + b \\
&= \frac{\partial}{\partial w_j} \sum_k x_k^{(i)} w_k + \frac{\partial}{\partial w_j} b \\
&= \sum_k \frac{\partial}{\partial w_j} x_k^{(i)} w_k \\
&= x_j^{(i)}
\end{aligned}$$

6 Recombining all the terms

Here we take the results of the three decomposed partial derivatives, as derived in sections 3, 4 and 5, and combine them again to a single partial derivative for the whole cost, as decomposed in section 2.

A lot of these terms end up cancelling each other out, simplifying the final expression for the partial derivative a lot. This is mostly due to rewrite step we did in section 4.2, and we end up with a partial derivative term that is surprisingly easy to compute.

$$\begin{aligned}
\frac{\partial J}{\partial w_j} &= \frac{1}{m}\sum_{i=1}^{m} \frac{\partial L^{(i)}}{\partial \hat{y}^{(i)}} \frac{\partial \hat{y}^{(i)}}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial w_j} \\
&= \frac{1}{m}\sum_{i=1}^{m} \left(\frac{-y^{(i)}}{\hat{y}^{(i)}} + \frac{1-y^{(i)}}{1-\hat{y}^{(i)}} \right) \ \hat{y}^{(i)}(1-\hat{y}^{(i)}) \ \ x_j^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m} \left(-y^{(i)}(1-\hat{y}^{(i)}) + (1-y^{(i)})\hat{y}^{(i)} \right) \ x_j^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m} \left(-y^{(i)} + y^{(i)}\hat{y}^{(i)} + \hat{y}^{(i)} - y^{(i)}\hat{y}^{(i)} \right) \ x_j^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right) x_j^{(i)}
\end{aligned}$$

7 Decomposition of the b derivative

The derivative for the bias b is very similar, and can also be decomposed as

$$\frac{\partial J}{\partial b} = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial L^{(i)}}{\partial \hat{y}^{(i)}} \frac{\partial \hat{y}^{(i)}}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial b}$$

All these terms are identical to the w_j derivative, except for $\frac{\partial z^{(i)}}{\partial b}$, which is just

$$\begin{aligned}
\frac{\partial z^{(i)}}{\partial b} &= \frac{\partial}{\partial b} \mathbf{x}^{(i)} \cdot \mathbf{w} + b \\
&= \frac{\partial}{\partial b} \sum_{k} x_k^{(i)} w_k + \frac{\partial}{\partial b} b \\
&= 1
\end{aligned}$$

Then, recombining the terms in the same way, the derivative for b just becomes

$$\frac{\partial J}{\partial b} = \frac{1}{m}\sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right)$$