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.

$$J_\theta(\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{N} E_\theta^i$$

$$E_\theta^i = -y^i log(h_\theta^i) - (1 - y^i) log(1 - h_\theta^i)$$

$$h_\theta^i = \frac{1}{1 + e^{-g_\theta^i}}$$

$$g_\theta^i = \theta^T \mathbf{x}^i$$

2 Decomposition of the 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. some θ_j might seem like a daunting task. However, there are several steps we can take to make taking this derivative a lot more doable.

The first of these steps, is to realise that all these functions applications can be derived using the **chain rule**, and so we can decompose the whole cost into several smaller partial derivatives.

$$\begin{aligned}
\frac{\partial J_\theta}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j} \sum_{i=1}^{N} E_\theta^i \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial \theta_j} E_\theta^i \\
&= \sum_{i=1}^{N} \frac{\partial E_\theta^i}{\partial h_\theta^i} \frac{\partial h_\theta^i}{\partial \theta_j} \\
&= \sum_{i=1}^{N} \frac{\partial E_\theta^i}{\partial h_\theta^i} \frac{\partial h_\theta^i}{\partial g_\theta^i} \frac{\partial g_\theta^i}{\partial \theta_j}
\end{aligned}$$

3 Cost function term

The first partial derivative from the decomposition is to take the derivative of one specific error term E_θ^i, so the part of the cost caused a specific training example $\mathbf{x}^i, y^i$ w.r.t. h_θ^i. Here h_θ^i is the hypothesis output for the i^{th} training example $\mathbf{x}^i, y^i$, as parameterized by θ. Even if this value is the result of (several) function application, we can just treat it as a partial derivative variable here.

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

$$\begin{aligned}\frac{\partial E_\theta^i}{\partial h_\theta^i} &= \frac{\partial}{\partial h_\theta^i} - y^i log(h_\theta^i) - (1-y^i) log(1-h_\theta^i) \\ &= \dots \\ &= \frac{-y^i}{h_\theta^i} + \frac{1-y^i}{1-h_\theta^i}\end{aligned}$$

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 h_θ^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

$$\begin{aligned}\frac{\partial h_\theta^i}{\partial g_\theta^i} &= \frac{\partial}{\partial g_\theta^i} \frac{1}{1+e^{-g_\theta^i}} \\ &= \frac{0\ (1+e^{-g_\theta^i}) - 1\ (\frac{\partial}{\partial g_\theta^i} 1 + e^{-g_\theta^i})}{(1+e^{-g_\theta^i})^2} \\ &= \frac{-(0+e^{-g_\theta^i})(-1)}{(1+e^{-g_\theta^i})^2} \\ &= \frac{e^{-g_\theta^i}}{(1+e^{-g_\theta^i})^2}\end{aligned}$$

4.2 Rewriting into terms of h_θ^i

$$\begin{aligned}
\frac{\partial h_\theta^i}{\partial g_\theta^i} &= \frac{e^{-g_\theta^i}}{(1+e^{-g_\theta^i})^2} \\
&= \frac{1}{1+e^{-g_\theta^i}} \; \frac{e^{-g_\theta^i}}{1+e^{-g_\theta^i}} \\
&= h_\theta^i \; \frac{e^{-g_\theta^i}}{1+e^{-g_\theta^i}} \\
&= h_\theta^i \; \frac{e^{-g_\theta^i} + (1-1)}{1+e^{-g_\theta^i}} \\
&= h_\theta^i \; \frac{1+e^{-g_\theta^i} - 1}{1+e^{-g_\theta^i}} \\
&= h_\theta^i \; \left(\frac{1+e^{-g_\theta^i}}{1+e^{-g_\theta^i}} - \frac{1}{1+e^{-g_\theta^i}} \right) \\
&= h_\theta^i \; \left(1 - \frac{1}{1+e^{-g_\theta^i}} \right) \\
&= h_\theta^i \; (1 - h_\theta^i)
\end{aligned}$$

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 θ w.r.t. some specific parameter θ_j.

$$\begin{aligned}
\frac{\partial g_\theta^i}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j} \theta^T \mathbf{x}^i \\
&= \frac{\partial}{\partial \theta_j} \sum_k \theta_k x_k^i \\
&= \sum_k \frac{\partial}{\partial \theta_j} \theta_k x_k^i \\
&= x_j^i
\end{aligned}$$

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.

$$
\begin{aligned}
\frac{\partial J_\theta}{\partial \theta_j} &= \sum_{i=1}^{N} \frac{\partial E_\theta^i}{\partial h_\theta^i} \frac{\partial h_\theta^i}{\partial g_\theta^i} \frac{\partial g_\theta^i}{\partial \theta_j} \\
&= \sum_{i=1}^{N} \left(\frac{-y^i}{h_\theta^i} + \frac{1-y^i}{1-h_\theta^i} \right) h_\theta^i (1-h_\theta^i) \; x_j^i \\
&= \sum_{i=1}^{N} \left(-y^i (1-h_\theta^i) + (1-y^i) h_\theta^i \right) \; x_j^i \\
&= \sum_{i=1}^{N} \left(-y^i + y^i h_\theta^i + h_\theta^i - y^i h_\theta^i \right) \; x_j^i \\
&= \sum_{i=1}^{N} \left(h_\theta^i - y^i \right) x_j^i
\end{aligned}
$$