$$\begin{aligned}\sigma'(x) &= \frac{\partial}{\partial x}\frac{1}{1+e^{-x}} \\ &= \frac{e^{-x}}{(1+e^{-x})^2} \\ &= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\ &= \sigma(x)(1-\sigma(x))\end{aligned}$$

And now, let's recall that if we have m points labelled $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, the error formula is:

$$E = -\frac{1}{m}\sum_{i=1}^{m} \left(y_i \ln(\hat{y}_i) + (1-y_i)\ln(1-\hat{y}_i)\right)$$

where the prediction is given by $\hat{y}_i = \sigma(Wx^{(i)} + b)$.

Our goal is to calculate the gradient of E, at a point $x = (x_1, \ldots, x_n)$, given by the partial derivatives

$$\nabla E = \left(\frac{\partial}{\partial w_1}E, \cdots, \frac{\partial}{\partial w_n}E, \frac{\partial}{\partial b}E\right)$$

To simplify our calculations, we'll actually think of the error that each point produces, and calculate the derivative of this error. The total error, then, is the average of the errors at all the points. The error produced by each point is, simply,

$$E = -y\ln(\hat{y}) - (1-y)\ln(1-\hat{y})$$

In order to calculate the derivative of this error with respect to the weights, we'll first calculate $\frac{\partial}{\partial w_j}\hat{y}$. Recall that $\hat{y} = \sigma(Wx+b)$, so:

$$\begin{aligned}\frac{\partial}{\partial w_j}\hat{y} &= \frac{\partial}{\partial w_j}\sigma(Wx+b) \\ &= \sigma(Wx+b)(1-\sigma(Wx+b)) \cdot \frac{\partial}{\partial w_j}(Wx+b) \\ &= \hat{y}(1-\hat{y}) \cdot \frac{\partial}{\partial w_j}(Wx+b) \\ &= \hat{y}(1-\hat{y}) \cdot \frac{\partial}{\partial w_j}(w_1x_1 + \cdots + w_jx_j + \cdots + w_nx_n + b) \\ &= \hat{y}(1-\hat{y}) \cdot x_j\end{aligned}$$

The last equality is because the only term in the sum which is not a constant with