$$= \hat{y}(1-\hat{y}) \cdot x_j$$

The last equality is because the only term in the sum which is not a constant with respect to w_j is precisely $w_j x_j$, which clearly has derivative x_j.

Now, we can go ahead and calculate the derivative of the error E at a point x, with respect to the weight w_j.

$$\begin{aligned}
\frac{\partial}{\partial w_j} E &= \frac{\partial}{\partial w_j}[-y\log(\hat{y}) - (1-y)\log(1-\hat{y})] \\
&= -y\frac{\partial}{\partial w_j}\log(\hat{y}) - (1-y)\frac{\partial}{\partial w_j}\log(1-\hat{y}) \\
&= -y \cdot \frac{1}{\hat{y}} \cdot \frac{\partial}{\partial w_j}\hat{y} - (1-y) \cdot \frac{1}{1-\hat{y}} \cdot \frac{\partial}{\partial w_j}(1-\hat{y}) \\
&= -y \cdot \frac{1}{\hat{y}} \cdot \hat{y}(1-\hat{y})x_j - (1-y) \cdot \frac{1}{1-\hat{y}} \cdot (-1)\hat{y}(1-\hat{y})x_j \\
&= -y(1-\hat{y}) \cdot x_j + (1-y)\hat{y} \cdot x_j \\
&= -(y-\hat{y})x_j
\end{aligned}$$

A similar calculation will show us that

$$\frac{\partial}{\partial b}E = -(y-\hat{y})$$

This actually tells us something very important. For a point with coordinates $(x_1, \ldots, x_n)$, label y, and prediction $\hat{y}$, the gradient of the error function at that point is $(-(y-\hat{y})x_1, \cdots, -(y-\hat{y})x_n, -(y-\hat{y}))$. In summary, the gradient is

$\nabla E = -(y-\hat{y})(x_1, \ldots, x_n, 1)$.

If you think about it, this is fascinating. The gradient is actually a scalar times the coordinates of the point! And what is the scalar? Nothing less than a multiple of the difference between the label and the prediction. What significance does this have?