$$\frac{\partial C}{\partial w_j} = x_j (a - y) \quad \text{——— (1)}$$

$$\frac{\partial C}{\partial b} = (a - y) \quad \text{——— (2)}$$

These are the equation we desire to have.

$$\frac{\partial C}{\partial b} = \frac{\partial C}{\partial a} \cdot \sigma'(z)$$

→ The real equation

$$\frac{\partial C}{\partial b} = \frac{\partial C}{\partial a} \times a(1-a) \quad \text{——— (3)}$$

comparing (2) and (3)

$$(a-y) = \frac{\partial C}{\partial a} \cdot a(1-a)$$

$$\frac{\partial C}{\partial a} = \frac{a-y}{a(1-a)}$$

Integrating w.r.t. a,

$$C = -[y \cdot \ln a + (1-y) \cdot \ln(1-a)] + C$$

constant

Summing and averaging over all the training examples, we get

$$C = \frac{-1}{n} \sum_{x} \left[y \cdot \ln a + (1-y) \cdot \ln (1-a) \right] + \text{constant}$$

And this is how people figured out the cross entropy cost function.