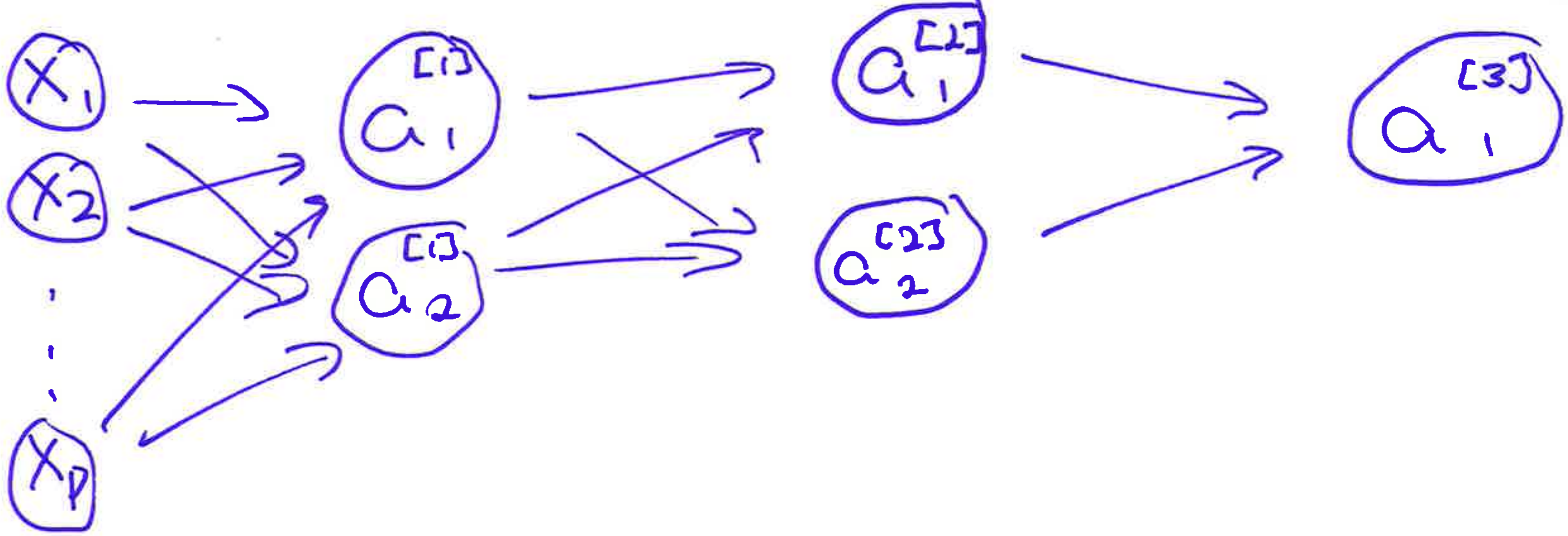
forward: starts with inputs

$$z^{[1]} = b^{[1]} + w^{[1]T} x \qquad a^{[1]} = g^{[1]}(a^{[1]})$$

$$\longrightarrow \quad z^{[2]} = b^{[2]} + w^{[2]T} a^{[1]} \qquad a^{[2]} = g^{[2]}(z^{[2]})$$

$$\longrightarrow \quad z^{[3]} = b^{[3]} + w^{[3]T} a^{[2]} \qquad a^{[3]} = g^{[3]}(z^{[3]})$$



Backward: starts with last layer

$$\frac{\partial J}{\partial a^{[3]}}$$

$$\frac{\partial J}{\partial z^{[3]}} = \frac{\partial J}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}}$$

$$\frac{\partial J}{\partial a^{[2]}} = \frac{\partial J}{\partial z^{[3]}} \cdot \frac{\partial z^{[3]}}{\partial a^{[2]}}$$

$$\frac{\partial J}{\partial z^{[2]}} = \frac{\partial J}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}}$$

$$\frac{\partial J}{\partial a^{[1]}} = \frac{\partial J}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial a^{[1]}}$$

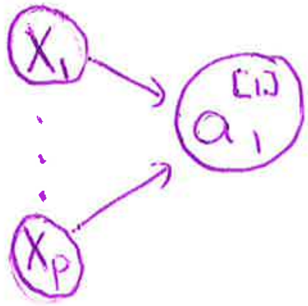
$$\frac{\partial J}{\partial z^{[1]}} = \frac{\partial J}{\partial a^{[1]}} \cdot \frac{\partial a^{[1]}}{\partial z^{[1]}}$$

$$\boxed{\begin{array}{lll} \frac{\partial J}{\partial b^{[1]}} = \frac{\partial J}{\partial z^{[1]}} \cdot \frac{\partial z^{[1]}}{\partial b^{[1]}} & \frac{\partial J}{\partial b^{[2]}} = \frac{\partial J}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial b^{[2]}} & \frac{\partial J}{\partial b^{[3]}} = \frac{\partial J}{\partial z^{[3]}} \cdot \frac{\partial z^{[3]}}{\partial b^{[3]}} \\ \frac{\partial J}{\partial w^{[1]}} = \frac{\partial J}{\partial z^{[1]}} \cdot \frac{\partial z^{[1]}}{\partial w^{[1]}} & \frac{\partial J}{\partial w^{[2]}} = \frac{\partial J}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial w^{[2]}} & \frac{\partial J}{\partial w^{[3]}} = \frac{\partial J}{\partial z^{[3]}} \cdot \frac{\partial z^{[3]}}{\partial w^{[3]}} \end{array}}$$

These are the numbers we need to do gradient descent

Previously:

Backpropagation for logistic regression:



$$z_1^{(i)[1]} = b^{[1]} + (w^{[1]})^T x^{(i)}$$

$$a_1^{(i)[1]} = g(z_1^{(i)[1]})$$

$\uparrow$ sigmoid

$$J(b,w) = -\sum_{i=1}^{m} \{ y^{(i)} \cdot \log(a_1^{(i)[1]}) + (1-y^{(i)}) \cdot \log(1-a_1^{(i)[1]}) \}$$

We found:

$$dJdz1 = a1 - y = [a_1^{(1)[1]} - y^{(1)} \quad \cdots \quad a_1^{(m)[1]} - y^{(m)}]$$

$$dJdb1 = np.mean(dJdz1,\ axis = 1,\ keepdims = True)$$

$$= \frac{1}{m} \sum_{i=1}^{m} (a_1^{(i)[1]} - y^{(i)})$$ as an array of shape (1,1) (same as b1)

$$dJdw1 = (1/m) * np.dot(X,\ dJdz1.T)$$

$\uparrow$ note: same as a0 ($a^{[0]}$)

$$= \frac{1}{m} \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(m)} \\ \vdots & \ddots & \vdots \\ x_p^{(1)} & \cdots & x_p^{(m)} \end{bmatrix} \begin{bmatrix} \frac{\partial J^{(1)}}{\partial z_1^{(1)[1]}} \\ \vdots \\ \frac{\partial J^{(m)}}{\partial z_1^{(m)[1]}} \end{bmatrix}$$

$\leftarrow$ resulting shape same as w1: (p,1) ($w1^T$ shape (1,p))

$$\frac{\partial z_1^{(1)[1]}}{\partial w_{1p}^{[1]}}$$

$$= \frac{1}{m} \begin{bmatrix} \frac{\partial J^{(1)}}{\partial z_1^{(1)[1]}} \cdot \frac{\partial z_1^{(1)[1]}}{\partial w_{11}^{[1]}} + \cdots + \frac{\partial J^{(m)}}{\partial z_1^{(m)[1]}} \cdot \frac{\partial z_1^{(m)[1]}}{\partial w_{11}^{[1]}} \\ \vdots \\ \frac{\partial J^{(1)}}{\partial z_1^{(1)[1]}} \cdot \frac{\partial z_1^{(1)[1]}}{\partial w_{1p}^{[1]}} + \cdots + \frac{\partial J^{(m)}}{\partial z_1^{(m)[1]}} \cdot \frac{\partial z_1^{(m)[1]}}{\partial w_{1p}^{[1]}} \end{bmatrix}$$

Back propagation

Main idea: One layer at a time, starting with the last layer, calculate: (suppose working on $l=2$)

For logistic regression we did not write these two calculations down, we directly computed $\frac{\partial J}{\partial z}$

- $\frac{\partial J}{\partial a2}$: shape (n_2, m)
    - for each unit in this layer and each observation i, how does cost function change based on value of activation output for that unit and observation?
    - np.dot(W3, dJdz3)

      ↑ dJdz for <u>next</u> layer if not currently working on output layer
    - derivation of above formula is complicated use of chain rule.
- $\frac{\partial J}{\partial z2}$: shape (n_2, m)
    - $\underbrace{\text{dJda2} * \text{dadz2}}$

      element-wise product of numbers like
      $$\frac{\partial J}{\partial a_j^{(i)[2]}} \cdot \frac{\partial a_j^{(i)[2]}}{\partial z_j^{(i)[2]}}$$
- $\frac{\partial J}{\partial b2}$: shape $(n_2, 1)$
    - np.mean(dJdz2, axis=1, keepdims=True)
      $$= \frac{1}{m} \sum_{i=1}^{m} \frac{\partial J}{\partial z^{(i)[2]}} \quad \text{since } \frac{\partial J}{\partial b_j^{[2]}} = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial J}{\partial z_j^{(i)[2]}} \cdot \cancel{\frac{\partial z_j^{(i)[2]}}{\partial b_j}} \to 1$$
- $\frac{\partial J}{\partial w2} = (1/m) * \text{np.dot(a1, dJdz2.T)}$ shape (n_1, n_2)

  ↑ a1 shape (n_1, m) ; after transpose, shape (m, n_2)

  same as for logistic reg, but based on a1 instead of a0.

Multivariate Chain Rule from Calculus:

(simpler case)

Suppose $f: \mathbb{R}^k \to \mathbb{R}$ is differentiable wrt each of its arguments.

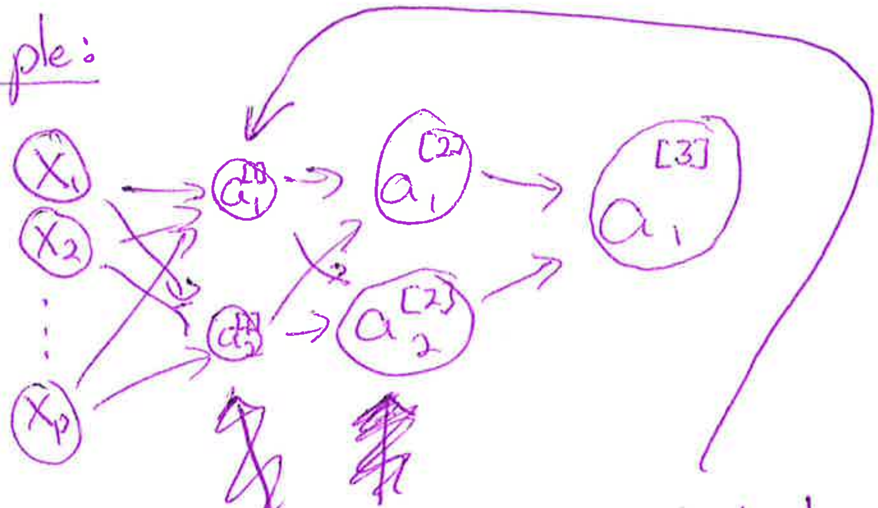
Then
$$\frac{\partial}{\partial x} f(g_1(x), g_2(x), \ldots, g_k(x))$$
$$= \left\{\frac{d}{dx} g_1(x)\right\} \cdot \frac{\partial}{\partial g_1} f(g_1, \ldots, g_k) + \cdots + \left\{\frac{d}{dx} g_k(x)\right\} \frac{\partial}{\partial g_k} f(g_1, \ldots, g_k)$$

Example:
$$f(g_1, g_2) = 2g_1 + g_2$$
$$g_1(x) = x^2$$
$$g_2(x) = x$$
$$f(g_1(x), g_2(x)) = 2g_1(x) + g_2(x)$$
$$= 2x^2 + x$$
$$\frac{d}{dx} f(g_1(x), g_2(x)) = 4x + 1$$

OR
$$\frac{d}{dx} f(g_1(x), g_2(x)) = \left\{\frac{d}{dx} g_1(x)\right\} \cdot \frac{\partial}{\partial g_1} f(g_1, g_2) + \left\{\frac{d}{dx} g_2(x)\right\} \cdot \frac{\partial}{\partial g_2} f(g_1, g_2)$$
$$= \{2x\} \cdot 2 + \{1\} \cdot 1$$
$$= 4x + 1$$

In Backpropagation, multivariate chain rule is used to find $\dfrac{\partial J}{\partial a_j^{(i)[l-1]}}$ ← rate of change of cost function with respect to the activation value for unit j in layer l-1 for observation number i.

Example:



Suppose we want to know how much a change in this unit's output for observation i affects the cost function.

Need to find $\dfrac{\partial J}{\partial a_1^{(i)[1]}}$

How can we find this?

Note: $a_1^{(i)[1]}$ feeds into calculation of z's for next layer:

$$z_1^{(i)[2]} = b_1^{[2]} + w_{11}^{[2]} a_1^{(i)[1]} + w_{12}^{[2]} a_2^{(i)[1]}$$

$$z_2^{(i)[2]} = b_2^{[2]} + w_{21}^{[2]} a_1^{(i)[1]} + w_{22}^{[2]} a_2^{(i)[1]}$$

so $\dfrac{\partial z_1^{(i)[2]}}{\partial a_1^{(i)[1]}} = w_{11}^{[2]}$ and $\dfrac{\partial z_2^{(i)[2]}}{\partial a_1^{(i)[1]}} = w_{21}^{[2]}$

Think of the cost function as a function of Z's from layer l: (here, layer $l=2$)

$$J = J\left(z_1^{(i)[2]}, z_2^{(i)[2]}, \ldots, z_{n_2}^{(i)[2]}\right)$$

Each of these z's depends on the activation output from the previous layer (in particular, on $a_2^{(i)[1]}$)

$$\text{So } \frac{\partial J}{\partial a_1^{(i)[1]}} = \frac{\partial z_1^{(i)[2]}}{\partial a_1^{(i)[1]}} \cdot \frac{\partial J}{\partial z_1^{(i)[2]}} + \frac{\partial z_2^{(i)[2]}}{\partial a_1^{(i)[1]}} \cdot \frac{\partial J}{\partial z_2^{(i)[2]}} + \cdots + \frac{\partial z_{n_2}^{(i)[2]}}{\partial a_1^{(i)[1]}} \cdot \frac{\partial J}{\partial z_{n_2}^{(i)[2]}}$$

weight given to $a_1^{(i)[1]}$ when calculating $z_1^{(i)[2]}$

$$= w_{11}^{[2]} \cdot \frac{\partial J}{\partial z_1^{(i)[2]}} + w_{21}^{[2]} \cdot \frac{\partial J}{\partial z_2^{(i)[2]}} + \cdots + w_{n_2 1}^{[2]} \cdot \frac{\partial J}{\partial z_{n_2}^{(i)[2]}}$$

Put this together for all n_1 units in layer 1:

$$\begin{bmatrix} \frac{\partial J}{\partial a_1^{(i)[1]}} \\ \vdots \\ \frac{\partial J}{\partial a_{n_1}^{(i)[1]}} \end{bmatrix} = \begin{bmatrix} w_{11}^{[2]} \cdot \frac{\partial J}{\partial z_1^{(i)[2]}} + \cdots + w_{n_2 1}^{[2]} \cdot \frac{\partial J}{\partial z_{n_2}^{(i)[2]}} \\ \vdots \\ w_{1 n_1}^{[2]} \frac{\partial J}{\partial z_1^{(i)[2]}} + \cdots + w_{n_2 n_1}^{[2]} \frac{\partial J}{\partial z_{n_2}^{(i)[2]}} \end{bmatrix} = \underbrace{\begin{bmatrix} | & & | \\ w_1^{[2]} & \cdots & w_{n_2}^{[2]} \\ | & & | \end{bmatrix}}_{W^{[2]}} \begin{bmatrix} \frac{\partial J}{\partial z_1^{(i)[2]}} \\ \vdots \\ \frac{\partial J}{\partial z_{n_2}^{(i)[2]}} \end{bmatrix}$$

now, stack observations next to each other

dJda1 = W $\cdot$ dJdz1