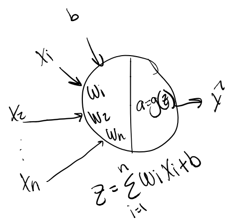


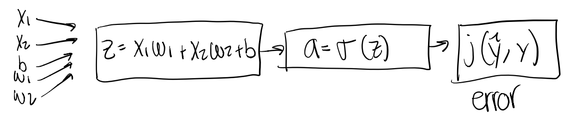
costo

$$j(\hat{y}, y) = -(y \log a + (1-y)\log(1-a))$$

error

Parámetros:

w_1, w_2, b



$$\boxed{a = \frac{1}{1+e^{-z}}}$$ ← función de activación sigmoidal

$\boxed{a = z}$ → net

$$w_1 = w_1 - \alpha \frac{\partial j}{\partial w_1}$$

$$w_2 = w_2 - \alpha \frac{\partial j}{\partial w_2}$$

$$\frac{dj}{dw_1} = \frac{dj}{da} \cdot \frac{da}{dz} \cdot \frac{dz}{dw_1}$$

$$b = b - \alpha \frac{\partial j}{\partial b}$$

$$da = \frac{dj}{da} = -\frac{y}{a} + \frac{1-y}{1-a}$$

$$dz = \frac{dj}{dz} = \left(-\frac{y}{a} + \frac{1-y}{1-a}\right) a(1-a) = -y(1-a) + (1-y)a = -y + \cancel{ya} + a - \cancel{ya} = \boxed{-y + a}$$

$$\frac{da}{dz} = (1+e^{-z})^{-2} e^{-z} = \frac{e^{-z}}{(1+e^{-z})^2} = a(1-a)$$

$$\frac{dz}{dw_1} = x_1$$

$$dw_1 = \frac{dj}{dw_1} = x_1 \frac{dj}{dz}$$

$$dw_2 = \frac{dj}{dw_2} = x_2 \frac{dj}{dz}$$

$$\frac{da}{dz} = a(1-a)$$

$$db = \frac{dj}{db} = \frac{dj}{dz}$$

$$J = \frac{1}{2}(a-y)^2$$

$$\frac{dJ}{dz} = dz$$

$a = g(z) = z$

$da = a - y$

$$\frac{da}{dz} = 1$$

$$dz = \frac{dJ}{dz} = \underbrace{\frac{dJ}{da}}_{da} \cdot \underbrace{\frac{da}{dz}}_{a(1-a)} \quad a-y$$

$(a-y)$ $\quad 1 \quad$ $a-y$

Para m ejemplos:

$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m} j(a^{(i)}, y^{(i)})$$

$$a^{(i)} = \hat{y}^{(i)} = \sigma(z^{(i)}) = \sigma(w^T x^{(i)} + b)$$

$$\frac{\partial J}{\partial w_i} = \frac{1}{m}\sum_{i=1}^{m} \underbrace{\frac{\partial}{dw_1} j(a^{(i)}, y^{(i)})}_{dw_1^{(i)}}$$

Pseudocódigo.

$J=0$, $dw_1=0$, $dw_2=0$, $db=0$

for $i=1$ to m {

Forward propagation
$$\begin{cases} z^{(i)} = w^T x^{(i)} + b \\ a^{(i)} = \sigma(z^{(i)}) \\ J += -[y^{(i)}\log(a^{(i)}) + (1-y^{(i)})\log(1-a^{(i)})] \end{cases}$$

Backward propagation
$$\begin{cases} dz^{(i)} = a^{(i)} - y^{(i)} \\ dw_1 += x_1^{(i)} dz^{(i)} \\ dw_2 += x_2^{(i)} dz^{(i)} \\ db += dz^{(i)} \end{cases}$$

}

$J /= m$, $dw_1 /= m$, $dw_2 /= m$, $db /= m$

$w_1 = w_1 - \alpha\, dw_1$

$w_2 = w_2 - \alpha\, dw_2$

$b = b - \alpha\, db$

Vectorizado:

$$X = \begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(m)} \\ | & | & & | \end{bmatrix} \; n \times m$$

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \; n \times 1$$

$$Z = [z^{(1)}\ z^{(2)} \cdots z^{(m)}] = \overset{1\times m}{W^T X} + \overset{1\times m}{[b\ b\ b\ b\ b]}$$

$= np.dot(W.T, X) + b$

$A = [a^{(1)}\ a^{(2)} \ldots a^{(m)}] = \sigma(Z)$

$dZ = A - Y$

$dw = \emptyset$ (vector)

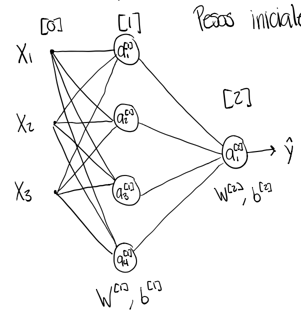
$dw = X\, dz^T / m$

$db = np.sum(dz)/m$

$w = w - \alpha dw$

$b = b - \alpha\, db$

NN Multicapa



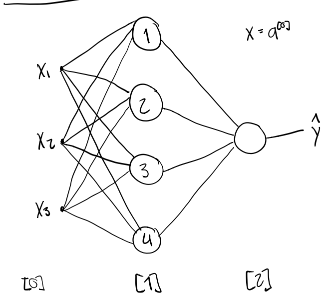
Pesos iniciales aleatoreos entre 0-1

Parámetros: $W^{[1]}, b^{[1]}, W^{[2]}, b^{[2]}$

$$x,\ W^{[1]},\ b^{[1]} \rightarrow \boxed{Z^{[1]} = W^{[1]T}x + b^{[1]}} \rightarrow \boxed{A^{[1]} = g(z)} \rightarrow \boxed{Z^{[2]} = W^{[2]T}A^{[1]} + b^{[2]}} \rightarrow \boxed{A^{[2]} = g(Z^{[2]})} \rightarrow \boxed{J(A^{[2]}, y)}$$

$dW^{[1]}$, $db^{[1]}$, $dZ^{[1]}$, $dA^{[1]}$, $dW^{[2]}$, $W^{[2]}$, $b^{[2]}$, $db^{[2]}$, $dZ^{[2]}$, $dA^{[2]}$

$$W^{[1]} = W^{[1]} - \alpha\, dW^{[1]}$$
$$b^{[1]} = b^{[1]} - \alpha\, db^{[1]}$$
$$W^{[2]} = W^{[2]} - \alpha\, dW^{[2]}$$
$$b^{[2]} = b^{[2]} - \alpha\, db^{[2]}$$



$x = a^{[0]}$

$$\boxed{z^{[1]} = W^{[1]}x + b^{[1]}} \rightarrow \boxed{a^{[1]} = g^{[1]}(z^{[1]})} \rightarrow \boxed{z^{[2]} = W^{[2]}a^{[1]} + b^{[2]}} \rightarrow \boxed{a^{[2]} = g^{[2]}(z^{[2]})} \rightarrow \boxed{J(a^{[2]}, y)}$$

$\hat{y}$

$$\frac{dJ}{dz^{[1]}} \equiv dz^{[1]} = \underset{da^{[1]}}{\frac{dJ}{da^{[1]}}} \cdot \underset{g^{[1]\prime}(z^{[1]})}{\frac{da^{[1]}}{dz^{[1]}}}$$

$$\frac{dJ}{da^{[1]}} \equiv da = \underset{dz^{[2]}}{\frac{dJ}{dz^{[2]}}} \cdot \underset{W^{[2]}}{\frac{dz^{[2]}}{da^{[1]}}}$$

$$\frac{dJ}{dz^{[2]}} \equiv dz^{[2]} = \underbrace{\frac{dJ}{da^{[2]}}}_{da^{[2]}} \cdot \underset{g^{[2]\prime}(z^{[2]})}{\frac{da^{[2]}}{dz^{[2]}}} \begin{cases} a(1-a) \\ 1 \end{cases}$$

$$\frac{dJ}{da^{[2]}} \equiv da^{[2]} = \begin{cases} \text{Sig} & -\frac{y}{a} + \frac{(1-y)}{(1-a)} \\ \text{Lin} & (a-y) \end{cases}$$

$$\frac{dJ}{dW^{[2]}} \equiv dW^{[2]} = \underset{dz^{[2]}}{\frac{dJ}{dz^{[2]}}} \cdot \underset{a^{[1]}}{\frac{dz^{[2]}}{dW^{[2]}}}$$

$$\frac{dJ}{db^{[2]}} \equiv db^{[2]} = \underset{dz^{[2]}}{\frac{dJ}{dz^{[2]}}} \cdot \underset{1}{\frac{dz^{[2]}}{db^{[2]}}}$$

$$\frac{dJ}{dW^{[1]}} \equiv dW^{[1]} = \underset{dz^{[1]}}{\frac{dJ}{dz^{[1]}}} \cdot \underset{a^{[0]} = x}{\frac{dz^{[1]}}{dW^{[1]}}}$$

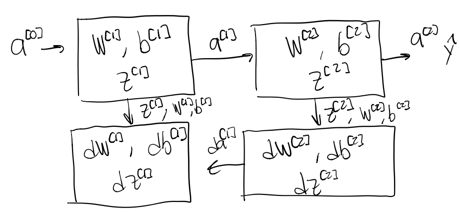
$$\frac{dJ}{db^{[1]}} \equiv db^{[1]} = \underbrace{\frac{dJ}{dz^{[1]}}}_{dz^{[1]}} \cdot \underbrace{\frac{dz^{[1]}}{db^{[1]}}}_{1}$$

features $n \times m$ ejemplos

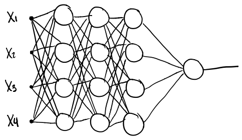
$$X = \begin{bmatrix} X_1^{(1)} & | & | & & | \\ X_2^{(1)} & X^{(2)} & X^{(3)} & \cdots & X^{(m)} \\ X_3^{(1)} & | & | & & | \end{bmatrix}$$

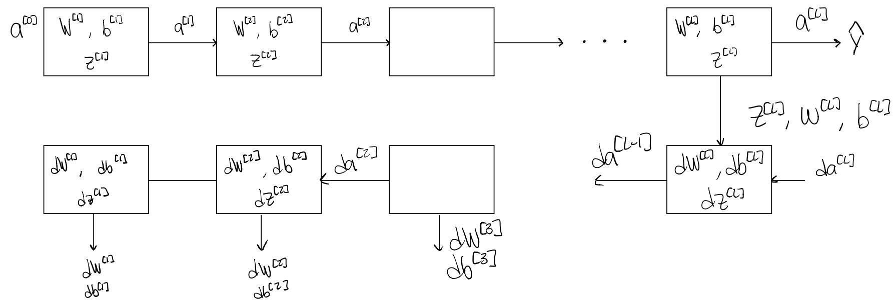
$$W = \begin{bmatrix} — w_1 — \\ — w_2 — \\ — w_3 — \\ — w_4 — \end{bmatrix} \quad 4 \times n$$

$$W^{[1]} = W^{[1]} - \alpha\, dW^{[1]}$$
$$b^{[1]} = b^{[1]} - \alpha\, db^{[1]}$$
$$W^{[2]} = W^{[2]} - \alpha\, dW^{[2]}$$
$$b^{[2]} = b^{[2]} - \alpha\, db^{[2]}$$



Red neuronal de L capas:





l = capa actual
$L \equiv$ número de capas (sin la 0).

FW propagation para la capa l

Entrada: $a^{[l-1]}$
Salida: $a^{[l]}, z^{[l]}, W^{[l]}, b^{[l]}$
$z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]}$
$a^{[l]} = g^{[l]}(z^{[l]})$

Vectorizado
$Z^{[l]} = W^{[l]} A^{[l-1]} + b^{[l]}$
$A^{[l]} = g^{[l]}(Z^{[l]})$

BW propagation:

Entrada: $da^{[l]}$
Salida: $da^{[l-1]}, dW^{[l]}, db^{[l]}$

iguales, usa una u otra.
$dz^{[l]} = da^{[l]} * g^{[l]\prime}(z^{[l]})$
$dw^{[l]} = dz^{[l]} \cdot a^{[l-1]}$
$db^{[l]} = dz^{[l]}$
$da^{[l-1]} = W^{[l]T} \cdot dz^{[l]}$
$dz^{[l]} = W^{[l+1]T} dz^{[l+1]} * g^{[l]\prime}(z^{[l]})$

Vectorizado:
$dz^{[l]} = dA^{[l]} * g^{[l]\prime}(z^{[l]})$
$dW^{[l]} = \frac{1}{m} dz^{[l]} \cdot A^{[l-1]T}$
$db^{[l]} = \frac{1}{m}$ np.sum($dz^{[l]}$, axis=1, keepdims=true)
$dA^{[l-1]} = W^{[l]T} \cdot dz^{[l]}$

L=total de capas
$k \equiv$ total de neuronas en la capa de salida
$S_l \equiv$ No. de neuronas (sin b) en la capa l

$$J = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log\left(h_\theta(x^{(i)})_k + (1-y_k^{(i)})\log(1-h_\theta(x^{(i)})_k\right)\right] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{S_l}\sum_{j=1}^{S_{l+1}} \left(\theta_{ij}^{(l)}\right)^2$$

W_{ij}

Resumen Clase Pasada:

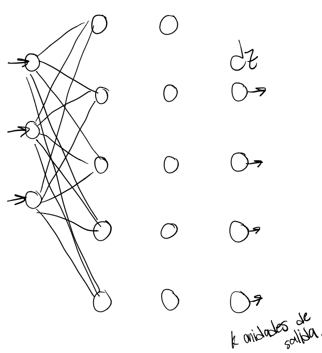
$Z^{[1]} = W^{[1]}X + b^{[1]}$
$A^{[1]} = g^{[1]}(Z^{[1]})$
$Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$
$A^{[2]} = g^{[2]}(Z^{[2]})$
⋮
$A^{[L]} = g^{[L]}(Z^{[L]})$

$L \equiv$ # capas

$dz^{[L]} = A^{[L]} - Y$
$dW^{[L]} = \frac{1}{m} dz^{[L]} A^{[L]T}$
$db^{[L]} = \frac{1}{m}$ np.sum($dz^{[L]}$, axis=1, keepdims=true)

$dz^{[L-1]} = dW^{[L]T} dz^{[L]} * g^{\prime[L-1]}(z^{[L-1]})$
⋮
$dz^{[1]} = dW^{[2]T} dz^{[2]} * g^{\prime[1]}(z^{[1]})$
$dW^{[1]} = \frac{1}{m} dz^{[1]} A^{[1]T}$
$db^{[1]} = \frac{1}{m}$ np.sum($dz^{[1]}$, axis=1, keepdims=true)

dz→ error en todas las neuronas.
El error en una de las capas depende del error de la capa siguiente. Todos dependen del error en la capa de salida.



Ejemplo: (x_i, y_i)

$$x_i = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ x_2^{(i)} \end{bmatrix} \quad y_i \in \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0\\1\end{bmatrix} \right\}$$

clase 1 2 3 4

$$J = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)} \log a_k^{(i)} + (1-y_k^{(i)})\log(1-a_k^{(i)})\right.$$

$$dz = \frac{dJ}{dz} = A^{[3]} - y$$

(4×1)

general $\rightarrow (k\times1)$

Esto es solo para la última capa, solo cambia el cálculo de a en la última capa.

Softmax layer.

hardmax	Softmax
$\begin{bmatrix}1\\0\\0\\0\end{bmatrix}$	$\begin{bmatrix}0.72\\0.16\\0.05\\0.\end{bmatrix}$

$$z^{[L]} = W^{[L]}a^{[L-1]} + b^{[L]}$$

Función de evaluación:

$$t = e^{z[L]}$$

$$a^{[L]} = \frac{e^{z[L]}}{\sum_{j=1}^{c} t_j}, \quad a_i^{[L]} = \frac{t_i}{\sum_{j=1}^{c} t_i}$$

$C \equiv$ # de clases

eg. $$z^{[L]} = \begin{bmatrix}5\\2\\-1\\3\end{bmatrix} \quad t = \begin{bmatrix}e^5\\e^2\\e^{-1}\\e^3\end{bmatrix} = \begin{bmatrix}148.4\\7.4\\0.4\\20.1\end{bmatrix} \quad \sum_{j=1}^{c} t_i = 176.3$$

$$a^{[L]} = \frac{t}{176.3} = \begin{bmatrix}0.842\\0.042\\0.002\\0.114\end{bmatrix}$$

$$J = \frac{1}{m}\sum j(\hat{y}, y) \qquad \frac{dJ}{dz^{[L]}} = dz^{[L]} = \hat{y} - y$$

$$j(\hat{y}, y) = -\sum_{j=1}^{c} y_j \log \hat{y}_j$$

$$y = \begin{bmatrix}0\\1\\0\\0\end{bmatrix} \quad a^{[L]} = \hat{y} = \begin{bmatrix}0.3\\0.2\\0.1\\0.4\end{bmatrix}$$

$$\underbrace{J(\hat{y}, y)}_{\text{pequeño}} = -y_2 \log \underbrace{\hat{y}_2}_{\text{grande}}$$

$$= -\log \hat{y}_2$$

$$J = -\sum_j y_j \log a_j$$

$$\frac{dJ}{dz_i} = \frac{d}{dz_i}\left(-\sum_j y_j \log a_j\right)$$

$$= -\sum_j \frac{d}{dz_i} y_j \log a_j$$

$$= -\sum_j y_j \frac{d}{dz_i}\log a_j$$

$$dz = \frac{dJ}{dz_i} = a_i - y_i = \hat{y} - y$$

$$a_j = \frac{e^{z_j}}{\sum_k e^{z_k}}$$

Caso 1: $i = j$

$$\frac{d}{dz_i}\log a_i =$$

Caso 2: $i \neq j$

$$\frac{d}{dz_i}\log a_j = \frac{1}{a_j}\frac{d}{dz_i}a_j = \frac{d}{dz_i}\frac{e^{z_j}}{\sum_k e^{z_k}}$$

$$= \frac{d}{dz_i} e^{z_j}\left(\sum e^{z_k}\right)^{-1} = -e^{z_j}\left(\sum e^{z_k}\right)^{-2}\frac{d}{dz_i}\sum_k e^{z_k}$$

$$= \frac{-e^{z_j}e^{z_i}}{\left(\sum e^{z_k}\right)^2} = \frac{-e^{z_j}}{\sum e^{z_k}} \cdot \frac{e^{z_i}}{\sum e^{z_k}} = -a_j a_i$$

$$-\frac{1}{\cancel{a_j}}\cancel{a_j}a_i = -a_i$$