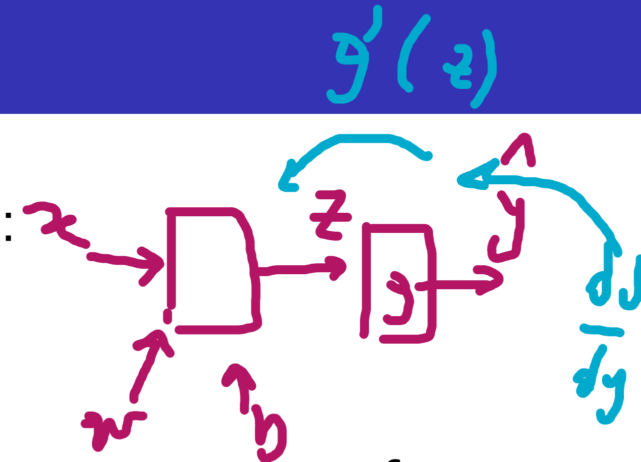
Redes neuronales (I) $\mathbb{R} \to \mathbb{R}$



Un perceptrón/neurona es un estimador de la forma:

$$\hat{y} = g(\overbrace{w \cdot x + b}^{z})$$

donde en su forma más simple $x, y, w, b \in \mathbb{R}$ y $g : \mathbb{R} \to \mathbb{R}$ es una función no lineal como puede ser la sigmoidea $\sigma(z) = \frac{1}{1+e^{-z}}$.

Si se define la función $J(W, b)$ de error respecto de los parámetros W y b se puede comprobar que, definiendo $z = w \cdot x + b$ y suponiendo conocido $\frac{dJ}{d\hat{y}} = dY \in \mathbb{R}$:

$$\frac{\partial \hat{y}}{\partial z} = g'(z) \quad \in \mathbb{R}$$

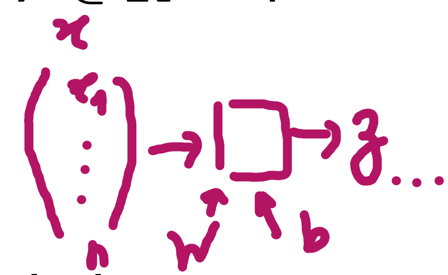
$$\frac{\partial J}{\partial W} = \frac{dJ}{d\hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial W} = \overbrace{dY}^{\mathbb{R}} \cdot \overbrace{g'(z)}^{\mathbb{R}} \cdot \overset{\in \mathbb{R}}{x} \quad \in \mathbb{R}$$

$$\frac{\partial J}{\partial b} = \frac{dJ}{d\hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial b} = dY \cdot g'(z) \cdot 1 \quad \in \mathbb{R}$$

Redes neuronales (II) $\mathbb{R}^n \to \mathbb{R}$

Si ahora consideramos múltiples entradas, es decir $x \in \mathbb{R}^n$, $W \in \mathbb{R}^{1\times n}$:

$$\hat{y} = g(\overbrace{W}^{1\times n} \cdot \overbrace{x}^{n\times 1} + b)$$



Entonces ahora para cada elemento de $\underbrace{W}_{1\times n} = (w_1, \ldots, w_n)$ vale lo anterior, y por tanto se puede comprobar que

$$\frac{\partial J}{\partial w_i} = dY \cdot g'(z) \cdot x_i$$

$$\frac{\partial J}{\partial W} = \nabla_J(W) = (dY \cdot g'(z) \cdot x_1, \ldots, dY \cdot g'(z) \cdot x_n) = \underbrace{dY}_{1} \cdot \underbrace{g'(z)}_{1} \cdot \underbrace{x^T}_{1\times n} \overset{1\times n}{\in \mathbb{R}}$$

$$= dY \cdot g'(z) \cdot \underbrace{(x_1, \ldots, x_n)}_{x^T} =$$

$$\frac{\partial J}{\partial b} = dY \cdot g'(z) \cdot 1 \quad \in \mathbb{R}$$

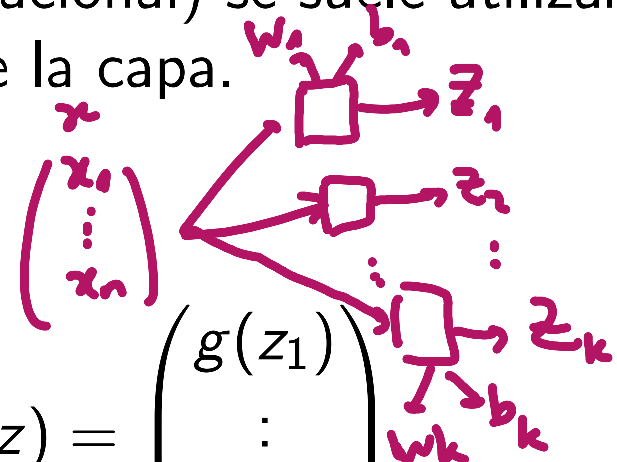
$$! \; z = w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_n \cdot x_n + b \rightarrow \frac{\partial z}{\partial w_i} = x_i, \; \frac{\partial z}{\partial x_i} = w_i$$

Redes neuronales (III) $\mathbb{R}^n \to \mathbb{R}^k$

Una capa en una red neuronal se define como un vector de k neuronas en paralelo. Una propiedad atractiva de este formato es que se puede considerar a la salida de una capa $y \in \mathbb{R}^k$ como simplemente el x de la capa siguiente. Por convención (y eficiencia computacional) se suele utilizar la misma no-linealidad g para todas las neuronas de la capa.

Nuevamente tenemos:

$$\hat{y} = g(W \cdot x + b)$$



donde $x \in \mathbb{R}^n$, $W \in \mathbb{R}^{k \times n}$, $b \in \mathbb{R}^k$ y se conviene $g(z) = \begin{pmatrix} g(z_1) \\ \vdots \\ g(z_k) \end{pmatrix}$

¿Y ahora cómo se calculan las derivadas para W y b?

$$\begin{cases} z_1 = W_1 \cdot x + b_1 \\ \vdots \\ z_k = W_k \cdot x + b_k \end{cases} \quad W_i \in \mathbb{R}^{1 \times n}, \; b_i \in \mathbb{R}$$

$$\equiv \begin{pmatrix} z_1 \\ \vdots \\ z_k \end{pmatrix} = \begin{pmatrix} -W_1- \\ \vdots \\ -W_k- \end{pmatrix} \cdot x + \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}$$

$$z = W \cdot x + b$$

En el caso de b es simple:

$\frac{\partial z_i}{\partial b_i} = 1 \rightarrow \forall i = 1, \ldots, k$

$$\frac{\partial J}{\partial b} = \overbrace{\frac{\partial J}{\partial \hat{y}}}^{k\times 1} \overbrace{\frac{\partial \hat{y}}{\partial z}}^{k\times 1} \overbrace{\frac{\partial z}{\partial b}}^{k\times 1} = \begin{pmatrix} dY_1 \\ \vdots \\ dY_k \end{pmatrix} \odot \begin{pmatrix} g'(z_1) \\ \vdots \\ g'(z_k) \end{pmatrix} \odot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = dY \odot g'(z) \in \mathbb{R}^k$$

Ahora para cada elemento de W tenemos:

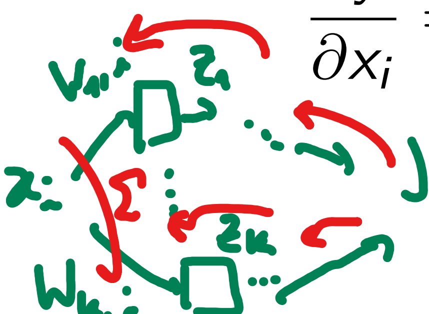
$$\begin{pmatrix} dY_1 \cdot g'(z_1) \\ \vdots \\ dY_k \cdot g'(z_k) \end{pmatrix} \cdot x^T$$

$$\frac{\partial J}{\partial W} = \begin{pmatrix} \frac{\partial J}{\partial W_{1,1}} & \cdots & \frac{\partial J}{\partial W_{1,n}} \\ \frac{\partial J}{\partial W_{2,1}} & \cdots & \frac{\partial J}{\partial W_{2,n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial J}{\partial W_{k,1}} & \cdots & \frac{\partial J}{\partial W_{k,n}} \end{pmatrix} = \begin{pmatrix} \nabla_J(W_{1,:}) \\ \vdots \\ \nabla_J(W_{k,:}) \end{pmatrix} = \begin{pmatrix} dY_1 \cdot g'(z_1) \cdot x^T \\ \vdots \\ dY_k \cdot g'(z_k) \cdot x^T \end{pmatrix} =$$

$\left(\frac{\partial J}{\partial W_{1,1}} \cdots \frac{\partial J}{\partial W_{1,n}}\right) = \nabla_J(W_1)$

$$= \begin{pmatrix} dY_1 \\ \vdots \\ dY_k \end{pmatrix} \odot \begin{pmatrix} g'(z_1) \\ \vdots \\ g'(z_k) \end{pmatrix} \cdot x^T = \overbrace{dY}^{k\times 1} \odot \overbrace{g'(z)}^{k\times 1} \cdot \overbrace{x^T}^{1\times n} \in \mathbb{R}^{k\times n}$$

¿Cómo se encadena esto? Nosotros estamos dando por conocida la derivada del error respecto de la salida de la capa, $dY = \frac{dJ}{d\hat{y}}$, *pero en realidad no tenemos idea si estamos en una capa intermedia o no.*



$$\frac{\partial \hat{y}}{\partial x_i} = \sum_{j=1}^{k} g'(z_j) \cdot W_{j,i} = \langle \begin{pmatrix} dY_1 \cdot g'(z_1) \\ \vdots \\ dY_k \cdot g'(z_k) \end{pmatrix}, \begin{pmatrix} W_{1,i} \\ \vdots \\ W_{k,i} \end{pmatrix} \rangle =$$

$$= \langle dY \odot g'(z), W_{:,i} \rangle = W_{i,:}^T \cdot dY \odot g'(z)$$

$$\begin{pmatrix} W_{1,1} & \cdots & W_{n,1} \\ \vdots & \ddots & \vdots \\ W_{k,1} & \cdots & W_{k,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

En forma vectorizada:

$$W^T \cdot dY \odot g'(z)$$

$$dX = \frac{\partial J}{\partial x} = \begin{pmatrix} \frac{\partial J}{\partial x_1} \\ \vdots \\ \frac{\partial J}{\partial x_n} \end{pmatrix} = \begin{pmatrix} W_{1,:}^T \cdot dY \odot g'(z) \\ \vdots \\ W_{n,:}^T \cdot dY \odot g'(z) \end{pmatrix} = \underbrace{W^T}_{n \times k} \cdot \underbrace{dY}_{k \times 1} \odot \underbrace{g'(z)}_{k \times 1} \in \mathbb{R}^n$$

Y ese dX no es otra cosa que el dY de la capa anterior!