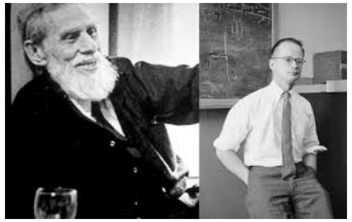
Universidad Nacional de General Sarmiento

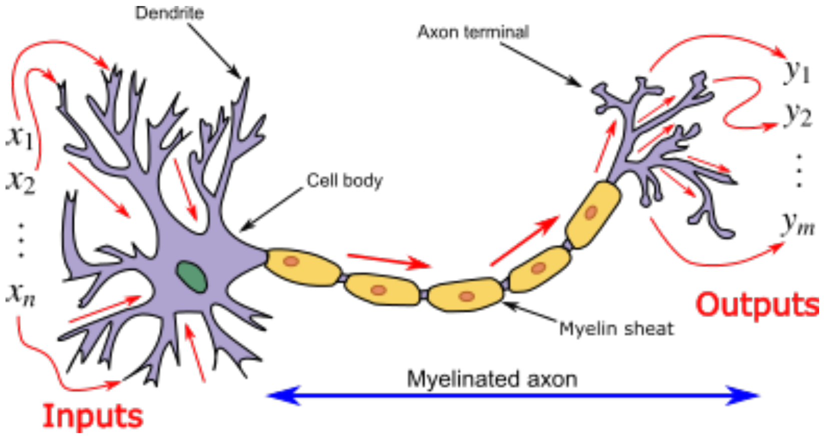
Licenciatura en Sistemas

Taller de Tesina

1943 – Warren McCulloch y Walter Pitts.
Modelo matemático de la neurona

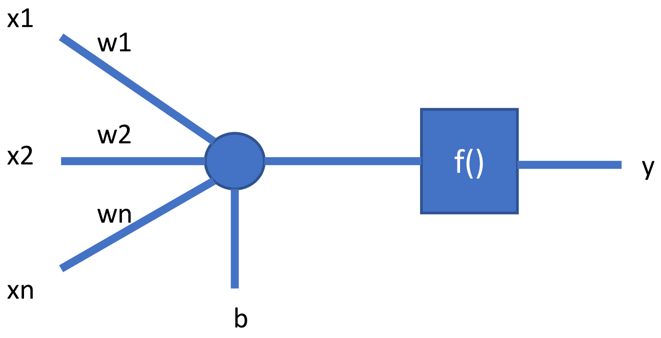


"Because of the all-or-none character of nervous activity, neural events and the relations among them can be treated by means of propositional logic"

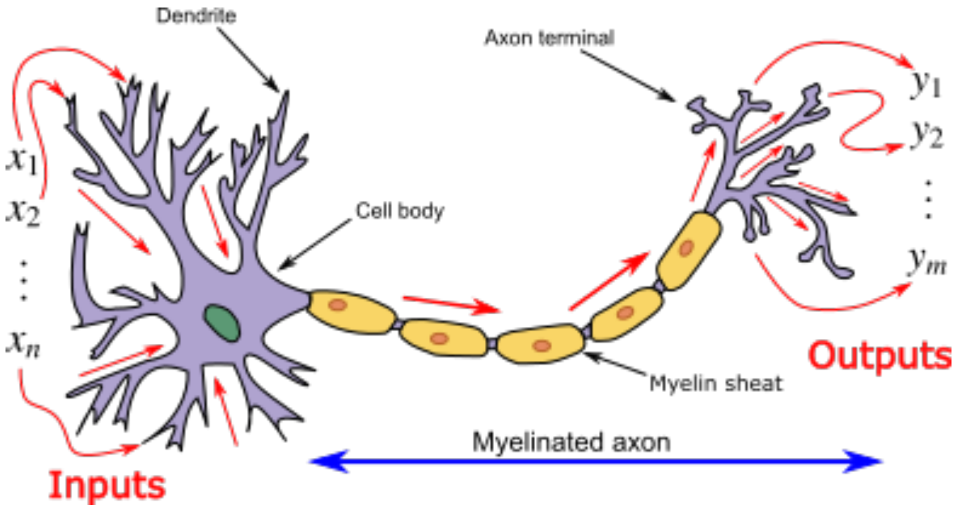


$$y = f\left(\sum_j w_j\, x_j + b\right)$$

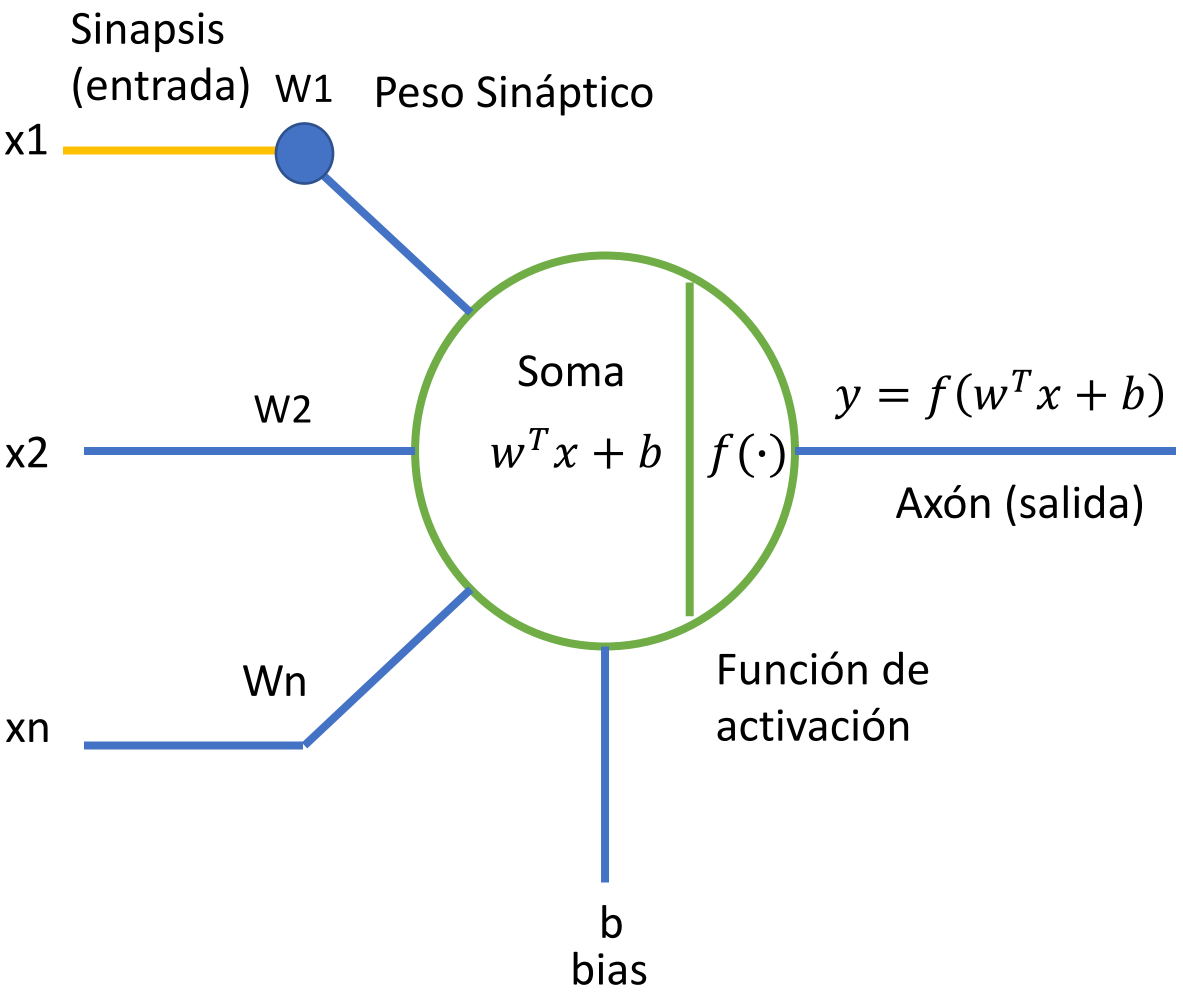
Donde f() es una función de activación:



$$\mathrm{f}(v) = \begin{cases} 1 \; si \; v > \theta \\ 0 \; si \; v \leq \theta \end{cases}$$



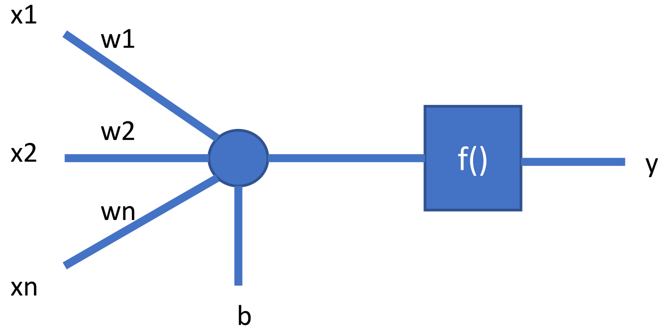
- Es un modelo matemático
- Las sinapsis biológicas son sistemas dinámicos mucho más complejos
- Lo mismo sucede en el soma neuronal.
- La función de activación emula al potencial de acción neuronal, no obstante hay transmisión de información entre neuronas biológicas sin potenciales de acción.



$$y = f(w^T x + b)$$

¿Cómo definir W?

¿Cómo definir b?



1949 – Donald Hebb : Aprendizaje Asociativo

No Supervisado

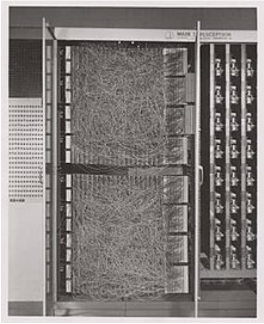
$$w_i = y x_i \qquad w_i(n+1) = w_i(n) + \Delta w_i$$

$$\Delta w_i = \eta y x_i$$

1962 – Frank Rosemblatt : Perceptrón.

Supervisado

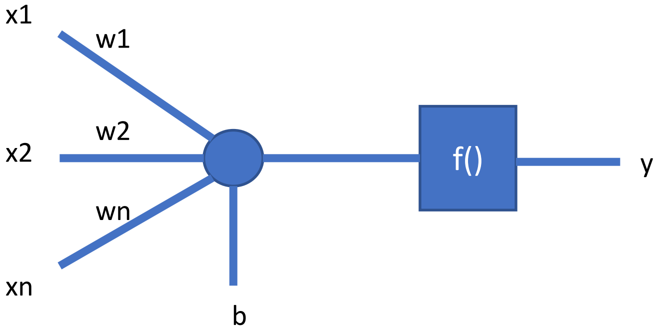
$$w_j(n+1) = w_j(n) + \rho(l_k - y_i)x_{j,k}$$



$$y = f(w^T x + b)$$

¿Cómo definir W?

¿Cómo definir b?



$$x = [X0, \quad X1, \quad X2, \ldots]^T \qquad w = [W0, \quad W1, \quad W2, \ldots]^T \qquad b = b$$

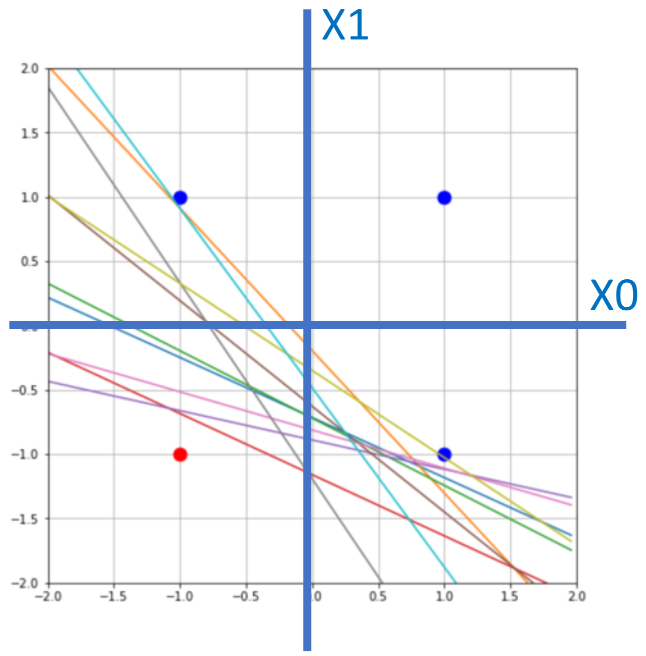
$$(w^T x + b) = W0\,X0 + W1\,X1 + W2\,X2 + \ \ldots + b$$

$$(w^T x + b) = [W0, W1, W2, \ldots, b]\,[X0, X1, X2, \ldots, 1]^T$$

$$w_j = \ [W0, W1, W2, \ldots, b]^T \qquad x_j = [X0, X1, X2, \ldots, 1]^T$$

$$y = f\left(w_j{}^T x_j\right) = f(\,[W0, W1, b][X0, X1, 1]^T) \qquad f(x) = \begin{cases} 1\ si\ x \geq 0 \\ -1\ si\ x < 0 \end{cases}$$

$$w_{j+1} = w_j + \rho(l_j - y_j)x_j \qquad \rho\ velocidad\ de\ convergencia$$



X0	X1	L
-1	-1	-1
-1	1	1
1	-1	1
1	1	1

Aprendizaje Supervisado en el Perceptrón.

Estado inicial aleatorio:

$$\boxed{w_0 = [0.06, 0.48, 0.4]^T = [W0, W1, b]^T}$$

Muestra:

$$x_0 = [1,1,1]^T = [X0, X1, 1]^T$$

$$l_0 = 1$$

Estimación:

$$y_0 = f({w_0}^T x_0) = f(0.94) = 1$$

Error:

$$e_0 = (l_0 - y_0) = 0$$

Aprendizaje (adaptación):

$$w_1 = w_0 + \rho\, e_0\, x_0 = w_0$$

$$w_1 = [0.06, 0.48, 0.4]^T$$

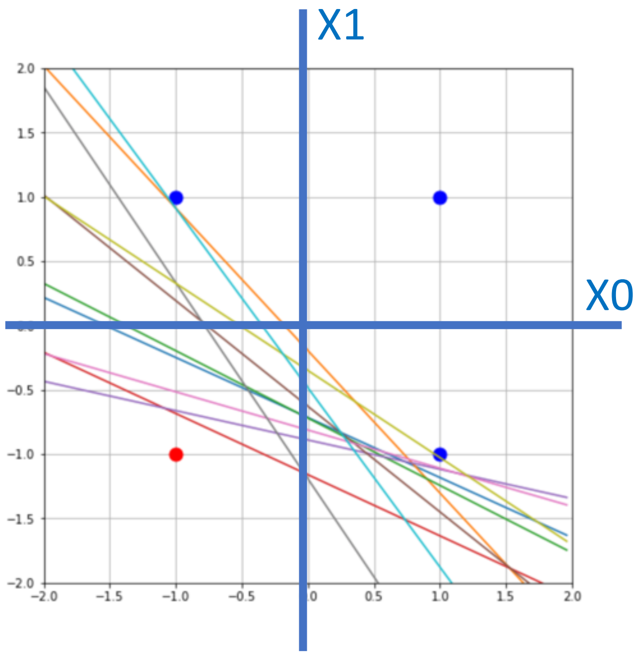
$$y = f(w^T x + b) = f([W0, W1, b][X0, X1, 1]^T)$$

$$f(x) = \begin{cases} 1 \; si \; x \geq 0 \\ -1 \; si \; x < 0 \end{cases}$$

$$w_j(n+1) = w_j(n) + \rho(l_k - y_i)x_j$$

$$\rho = 0.1$$

Aprendizaje Supervisado en el Perceptrón.



X0	X1	L
-1	-1	-1
-1	1	1
1	-1	1
1	1	1

Estado inicial aleatorio:

$$\boxed{w_0 = [0.06{,}0.48{,}0.4]^T = [W0, W1, b]^T}$$

Muestra:

$$x_0 = [1{,}1{,}1]^T = [X0, X1{,}1]^T$$

$$l_0 = 1$$

Estimación:

$$y_0 = f({w_0}^T x_0) = f(0.94) = 1$$

Error:

$$e_0 = (l_0 - y_0) = 0$$

Aprendizaje (adaptación):

$$w_1 = w_0 + \rho\, e_0\, x_0 = w_0$$

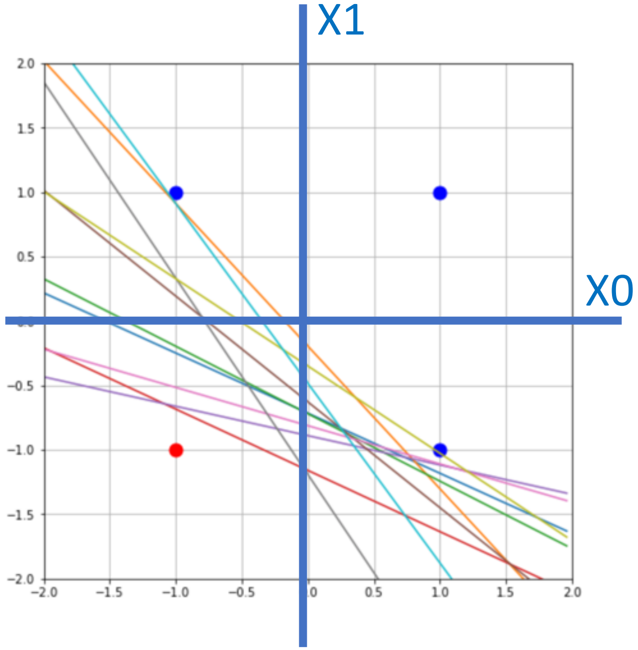
$$w_1 = [0.06, 0.48\,, 0.4]^T$$

$$y = f(w^T x + b)$$

$$f(x) = \begin{cases} 1\ si\ x \geq 0 \\ -1\ si\ x < 0 \end{cases}$$

$$w_j(n+1) = w_j(n) + \rho(l_k - y_i)x_j$$

$$\rho = 0.1$$



X0	X1	L
-1	-1	-1
-1	1	1
1	-1	1
1	1	1

$\boxed{w_1 = [0.06, 0.48, 0.4]^T}$

Muestra:

$x_1 = [1, -1, 1]^T = [X0, X1, 1]^T$

$l_1 = 1$

Estimación:

$y_1 = f({w_1}^T x_1) = f(-0.02) = -1$

Error:

$e_1 = (l_1 - y_1) = 2$

Aprendizaje (adaptación):

$w_2 = w_1 + \rho\ e_0\ x_0 = [0.06, 0.48, 0.4]^T +$
$0.1\ (2)\ [1, -1, 1]^T$

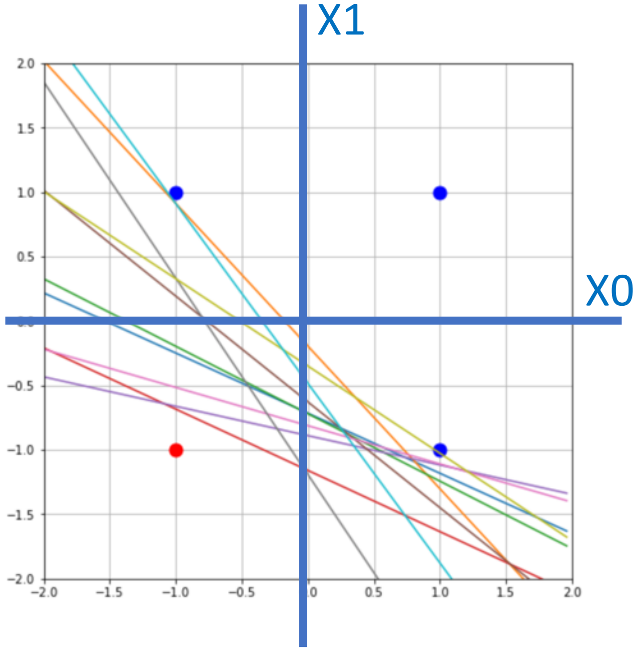
$w_2 = [0.26, 0.28\ , 0.6]^T$

$y = f(w^T x + b)$

$f(x) = \begin{cases} 1\ si\ x > 0 \\ -1\ si\ x \leq 0 \end{cases}$

$w_j(n+1) = w_j(n) + \rho(l_k - y_i)x_j$

$\rho = 0.1$



X0	X1	L
-1	-1	-1
-1	1	1
1	-1	1
1	1	1

$\boxed{w_2 = [0.26, 0.28, 0.6]^T}$

Muestra:

$x_2 = [-1,-1,1]^T = [X0, X1, 1]^T$

$l_2 = -1$

Estimación:

$y_2 = f(w_2{}^T x_2) = f(0.02) = 1$

Error:

$e_2 = (l_2 - y_2) = -2$

Aprendizaje (adaptación):

$w_3 = w_2 + \rho\, e_2\, x_2 =$

$[0.26, 0.28, 0.6]^T + 0.1\ (-2)\ [-1,-1,1]^T$

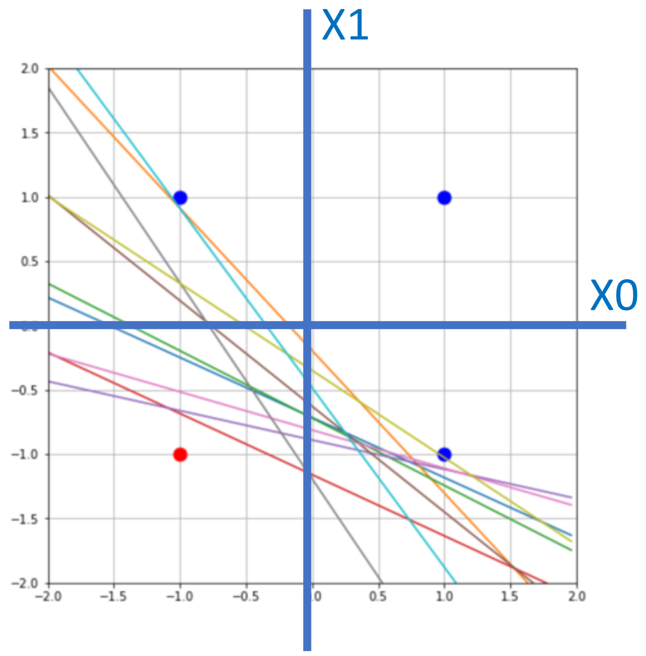
$w_3 = [0.46, 0.48, 0.4]^T$

$y = f(w^T x + b)$

$f(x) = \begin{cases} 1\ si\ x > 0 \\ -1\ si\ x \leq 0 \end{cases}$

$w_j(n+1) = w_j(n) + \rho(l_k - y_i)x_j$

$\rho = 0.1$



X0	X1	L
-1	-1	-1
-1	1	1
1	-1	1
1	1	1

$\boxed{w_2 = [0.26, 0.28\,, 0.6]^T}$

Muestra:

$x_2 = [-1, -1, 1]^T = [X0, X1, 1]^T$

$l_2 = -1$

Estimación:

$y_2 = f({w_2}^T x_2) = f(0.02) = 1$

Error:

$e_2 = (l_2 - y_2) = -2$

Aprendizaje (adaptación):

$w_3 = w_2 + \rho\, e_2\, x_2 =$

$[0.26, 0.28\,, 0.6]^T + 0.1\,(-2)\,[-1, -1, 1]^T$

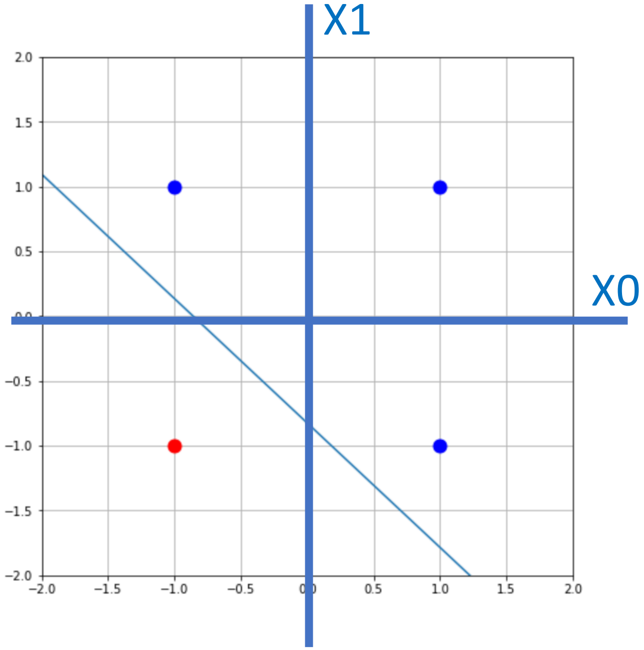
$$w_3 = [0.46, 0.48\,, 0.4]^T$$

$y = f(w^T x + b)$

$$f(x) = \begin{cases} 1 \; si \; x > 0 \\ -1 \; si \; x \leq 0 \end{cases}$$

$w_j(n+1) = w_j(n) + \rho(l_k - y_i)x_j$

$\rho = 0.1$



X0	X1	L
-1	-1	-1
-1	1	1
1	-1	1
1	1	1

$$w_3 = [0.46, 0.48\, , 0.4]^T$$

Hiperplano de clasificación:

$$w^T x + b = {w_3}^T\ [X0, X1, 1] = 0$$

$$0.46\ X0 + 0.48\ X1 + 0.4 = 0$$

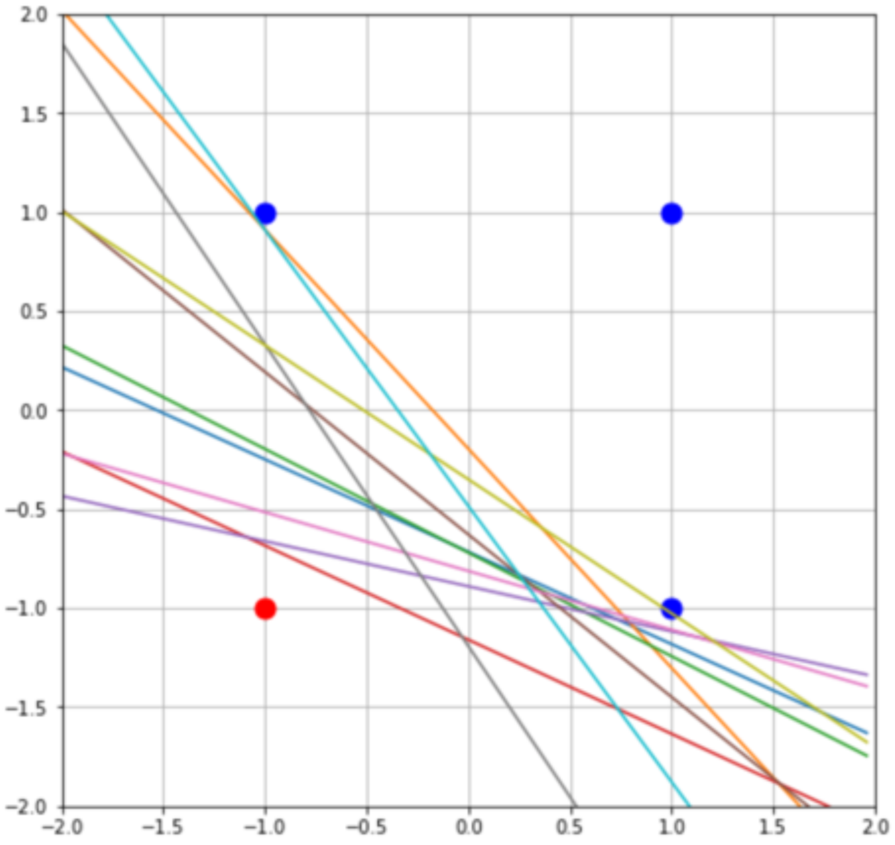
$$X1 = -\frac{0.46}{0.48}\ X0 - \frac{0.4}{0.48}$$

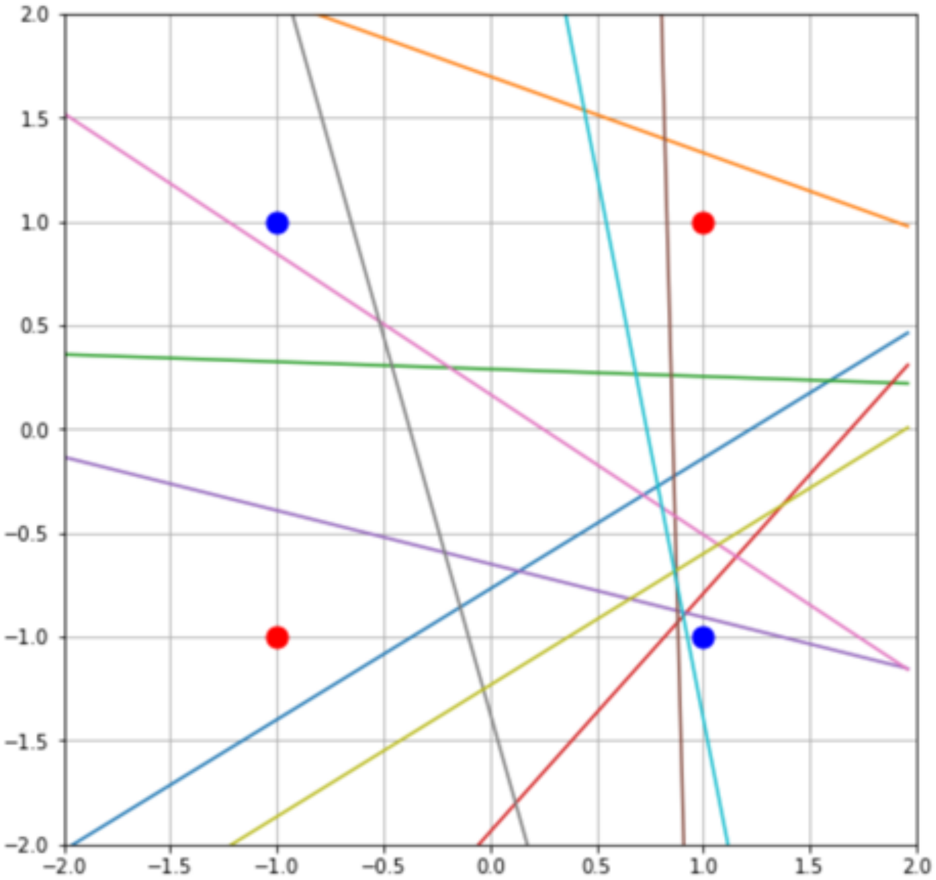
$$y = f(w^T x + b)$$

$$f(x) = \begin{cases} 1\ si\ x > 0 \\ -1\ si\ x \leq 0 \end{cases}$$

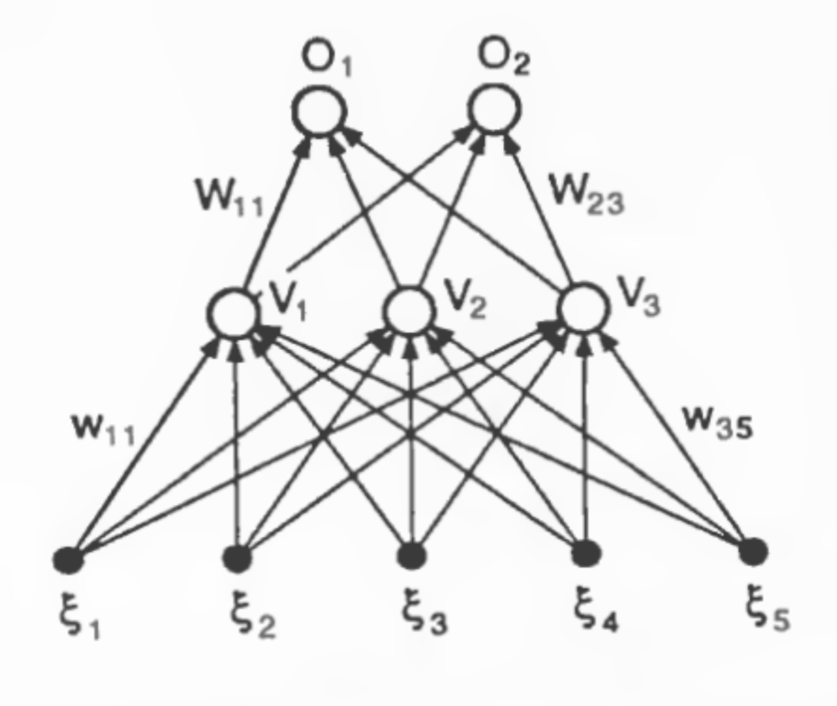
$$w_j(n+1) = w_j(n) + \rho(l_k - y_i)x_j$$

$$\rho = 0.1$$


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0

Redes multicapas – Perceptrón Multicapa



Datos de entrada y etiquetas de entrenamiento

$$\xi^{\mu} = [\xi_1, \xi_2, \cdots, \xi_k, \cdots]^{\mu} \leftrightarrow \lambda^{\mu} = [\lambda_1, \lambda_2]^{\mu}$$

Salida Nivel oculto

$$V_j = f\left(\sum_k w_{jk}\xi_k\right)$$ $f(.)$ es una función diferenciable

Salida de la Red

$$O_i = f\left(\sum_j W_{ij} f\left(\sum_k w_{jk}\xi_k\right)\right)$$

Error cuadrático

$$E(W_{ij}, w_{jk}) = \frac{1}{2}\sum_{\mu i}\left\|\lambda_i^{\mu} - O_i^{\mu}\right\|^2$$

¿W?

Obtener los parámetros W_{ij}, w_{jk} tal que el error,

$$E(W_{ij}, w_{jk}) = \frac{1}{2} \sum_{\mu i} \left\| \lambda_i^\mu - O_i^\mu \right\|^2$$

sea mínimo.

Obtener los parámetros W_{ij}, w_{jk} tal que la función de **loss**,

$$L(W_{ij}, w_{jk}) = E(W_{ij}, w_{jk}) + \sum \|W\|^2 = \frac{1}{2} \sum_{\mu i} \left\| \lambda_i^\mu - O_i^\mu \right\|^2 + + \sum \|W\|^2$$

sea mínima.

¿Cómo?

¿W?

Obtener los parámetros W_{ij}, w_{jk} tal que el error,

$$E(W_{ij}, w_{jk}) = \frac{1}{2}\sum_{\mu i} \left\|\lambda_i^\mu - O_i^\mu\right\|^2$$

sea mínimo.

Obtener los parámetros W_{ij}, w_{jk} tal que la función de **loss**,

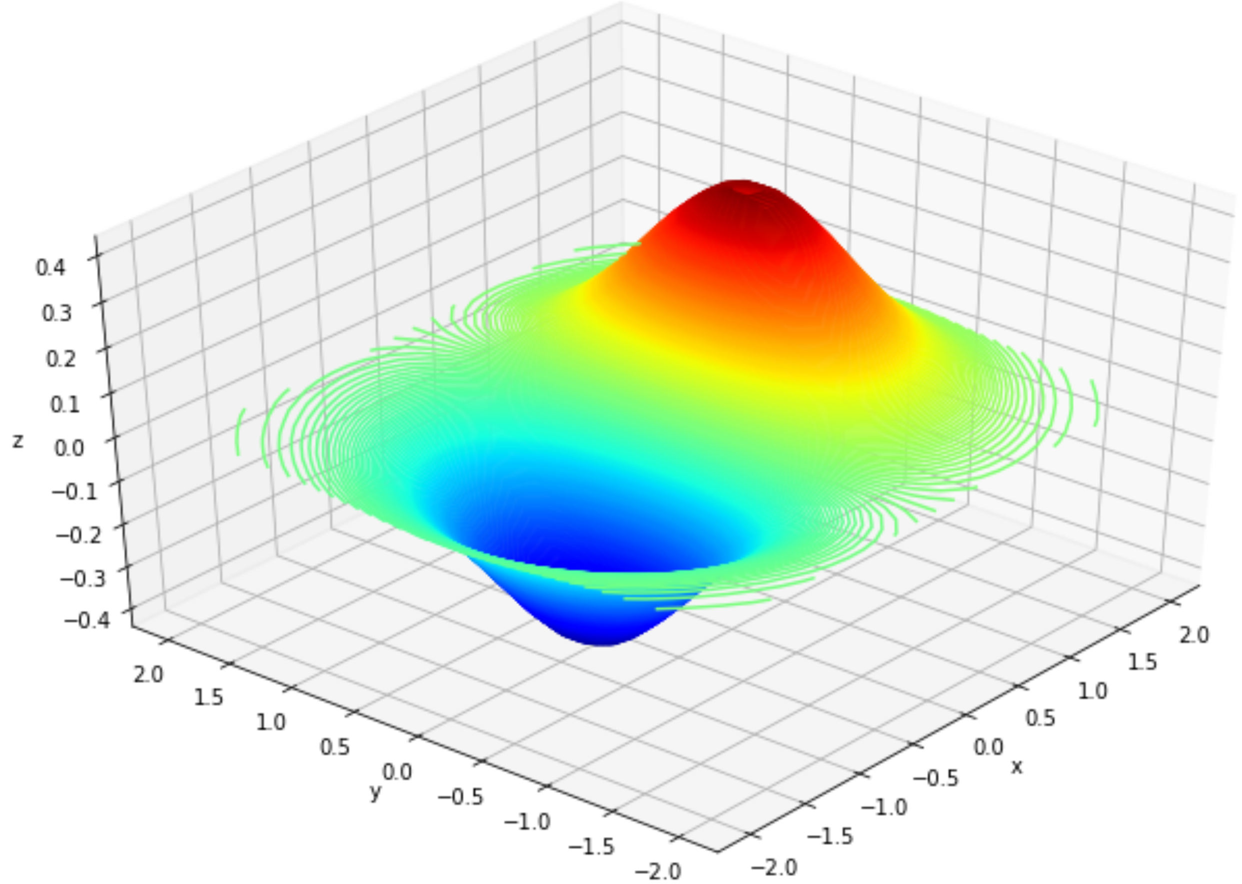
$$L(W_{ij}, w_{jk}) = E(W_{ij}, w_{jk}) + \sum \|W\|^2 = \frac{1}{2}\sum_{\mu i} \left\|\lambda_i^\mu - O_i^\mu\right\|^2 + + \sum \|W\|^2$$

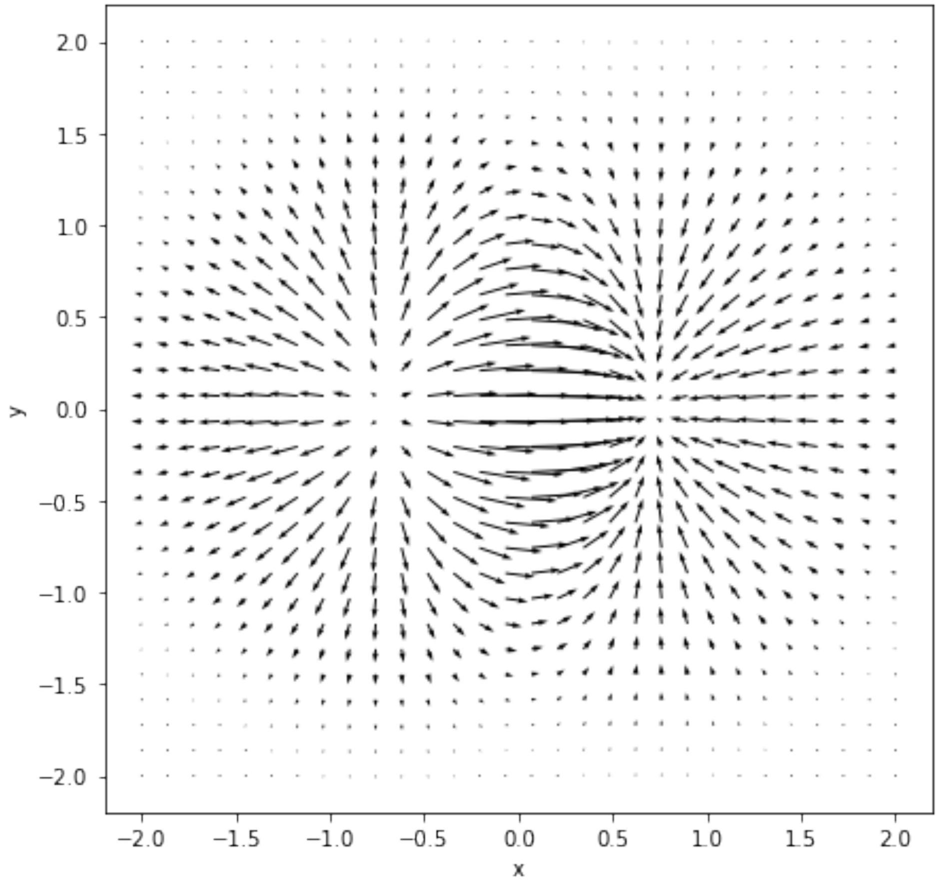
sea mínima.

¿Cómo?

Gradiente de una Función

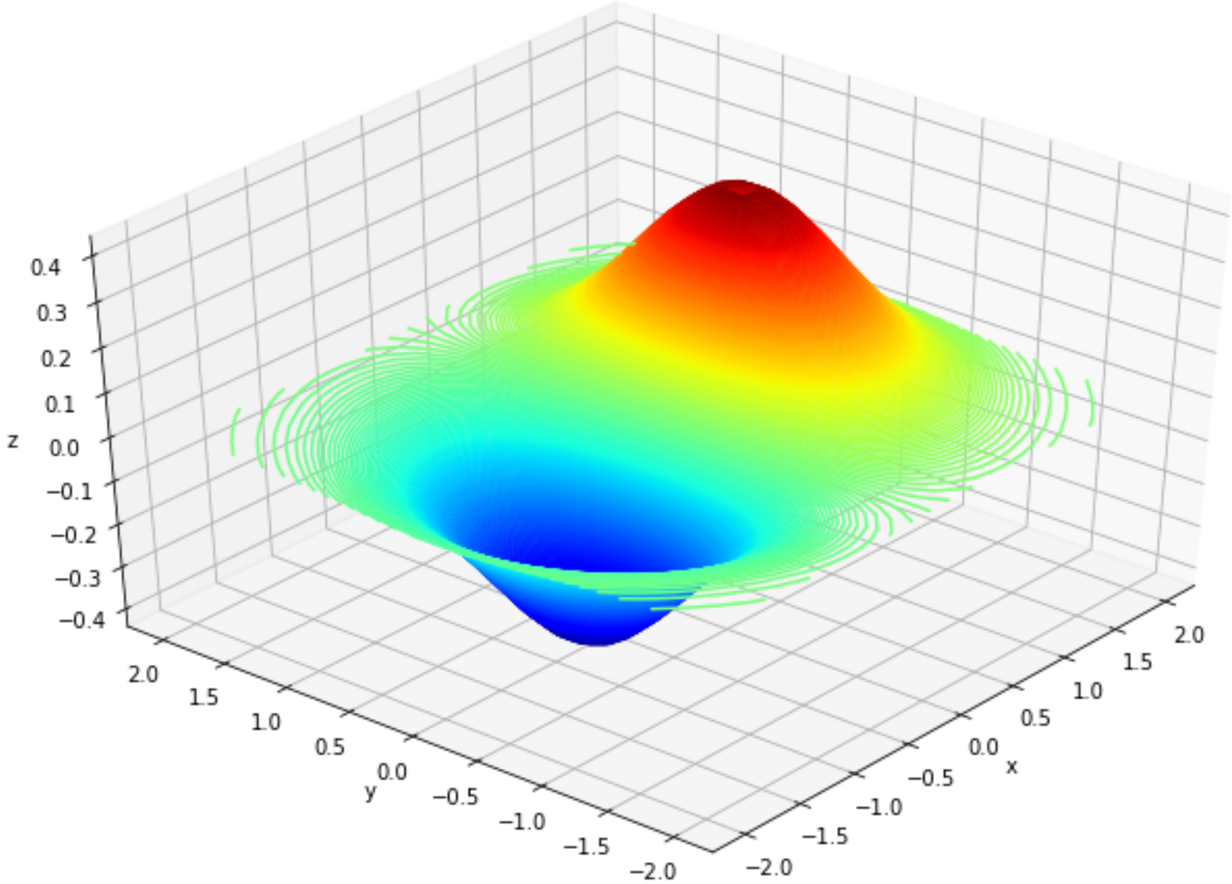
$$\nabla f(x_0, x_1, x_2, \ldots) = \frac{\partial f}{\partial x_i}$$

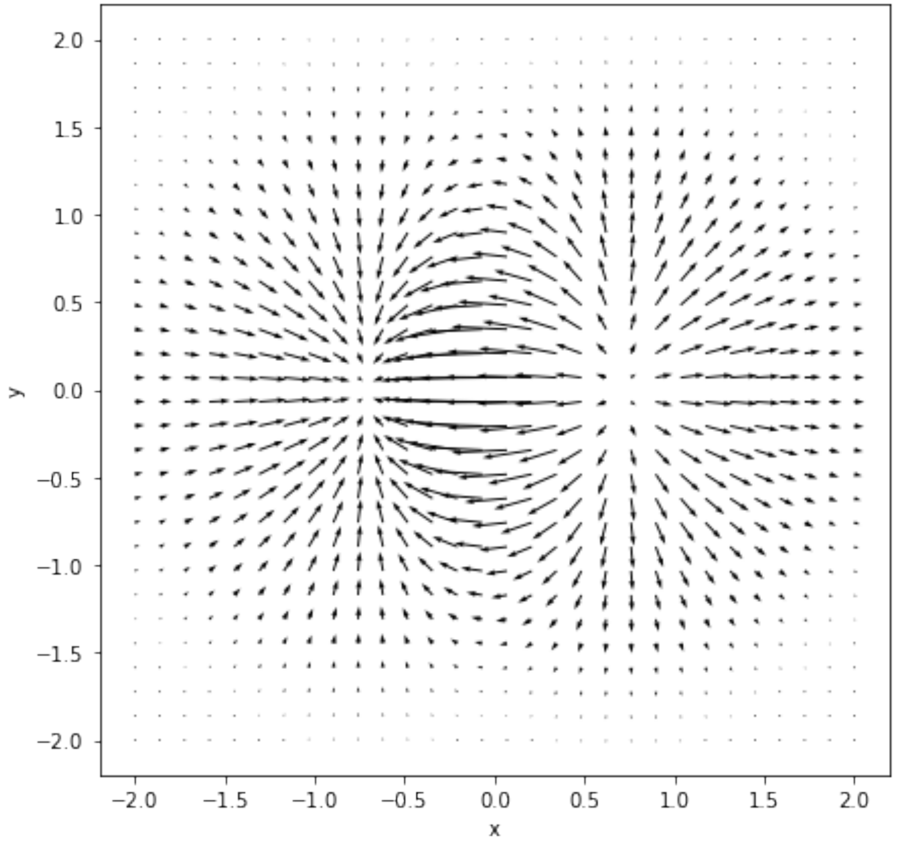




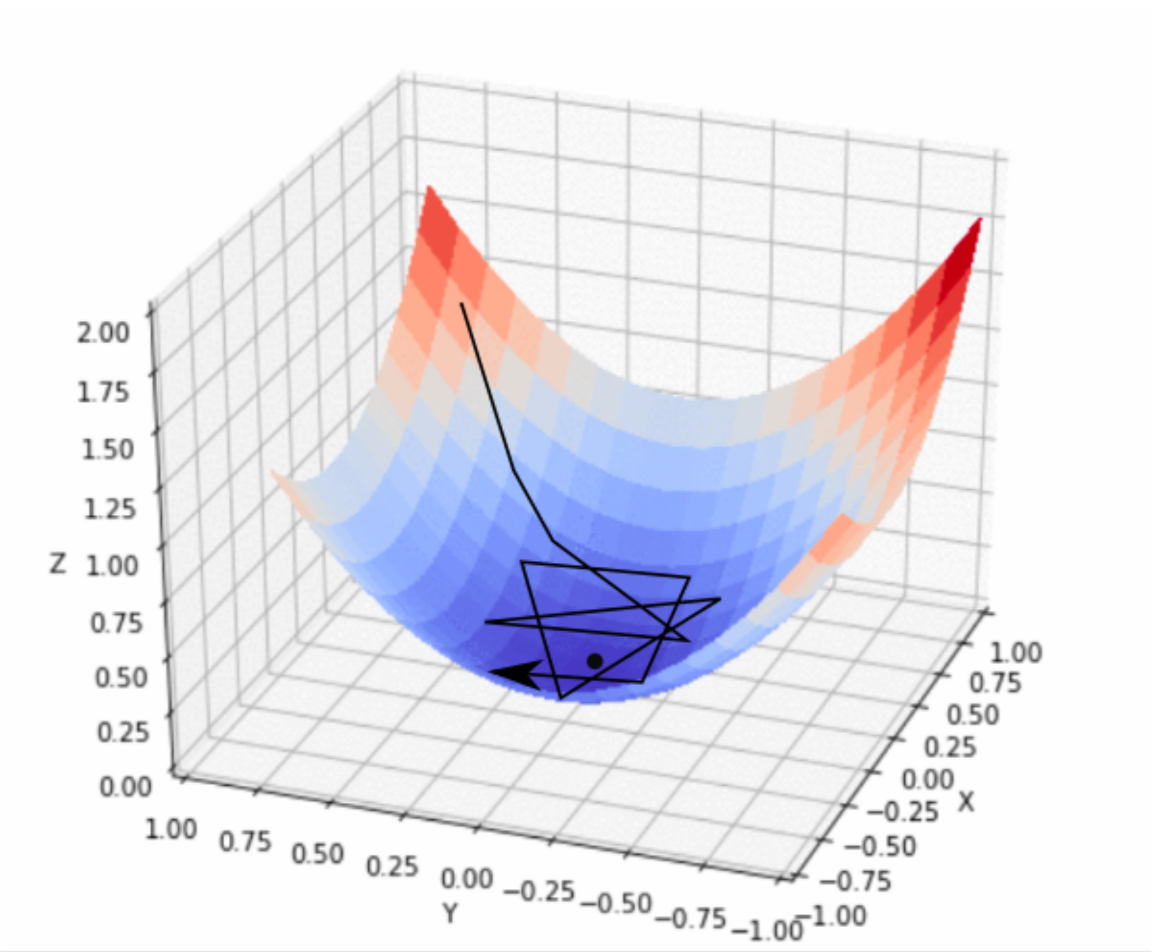
Gradiente Descendente

$$\nabla f(x_0, x_1, x_2, \ldots) = -\frac{\partial f}{\partial x_i}$$





Gradiente descendente



$$E(w) = \frac{1}{2}\sum_{i\mu}\left\|\lambda_i^{\mu} - O_i^{\mu}\right\|^2$$

$$\Delta W_{ij} = -\eta\,\frac{\partial E}{\partial W_{ij}} \qquad \Delta w_{jk} = -\eta\,\frac{\partial E}{\partial w_{jk}}$$

$$W_{ij}(n+1) = W_{ij}(n) + \Delta W_{ij}$$

$$w_{jk}(n+1) = w_{jk}(n) + \Delta w_{jk}$$

Backpropagation

$$E(w) = \frac{1}{2}\sum_{i\mu}\left\|\lambda_i^\mu - O_i^\mu\right\|^2 \qquad O_i = f\left(\sum_j W_{ij}V_j\right) \qquad V_j = f\left(\sum_k w_{jk}\xi_k\right)$$

$$\Delta W_{ij} = -\eta\frac{\partial E}{\partial W_{ij}} \qquad \Delta w_{jk} = -\eta\frac{\partial E}{\partial w_{jk}} = -\eta\sum_\mu\frac{\partial E}{\partial V_j}\frac{\partial V_j}{\partial w_{jk}}$$

$$\Delta W_{ij} = \eta\sum_\mu\left(\lambda_i^\mu - O_i^\mu\right)f'(\cdot)V_j \qquad \Delta w_{jk} = \eta\sum_\mu\sum_i\left(\lambda_i^\mu - O_i^\mu\right)f'(\cdot_i)W_{ij}f'(\cdot_j)\xi_j$$

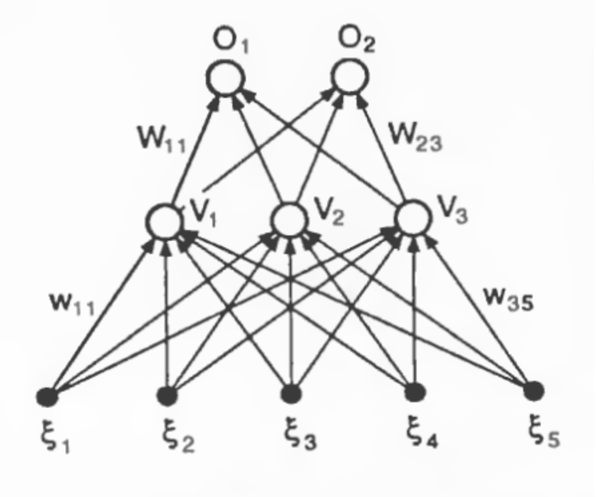
$$E(w) = \frac{1}{2}\sum_{i\mu}\left\|\lambda_i^\mu - O_i^\mu\right\|^2 \qquad O_i = f\left(\sum_j W_{ij}V_j\right) \qquad V_j = f\left(\sum_k w_{jk}\xi_k\right)$$

$$\Delta W_{ij} = -\eta\frac{\partial E}{\partial W_{ij}} \qquad \Delta w_{jk} = -\eta\frac{\partial E}{\partial w_{jk}} = -\eta\sum_\mu \boxed{\frac{\partial E}{\partial V_j}}\frac{\partial V_j}{\partial w_{jk}}$$

$$\Delta W_{ij} = \eta\sum_\mu(\lambda_i^\mu - O_i^\mu)f'(\cdot)V_j \longrightarrow \Delta w_{jk} = \eta\sum_\mu \boxed{\sum_i(\lambda_i^\mu - O_i^\mu)f'(\cdot_i)W_{ij}}f'(\cdot_j)\xi_j$$

$$\Delta W_{ij} = \eta\sum_\mu(\lambda_i^\mu - O_i^\mu)f'(\cdot_i)V_j$$

$$\Delta w_{jk} = \eta\sum_\mu f'(\cdot_j)\xi_j\sum_i(\lambda_i^\mu - O_i^\mu)f'(\cdot_i)W_{ij}$$



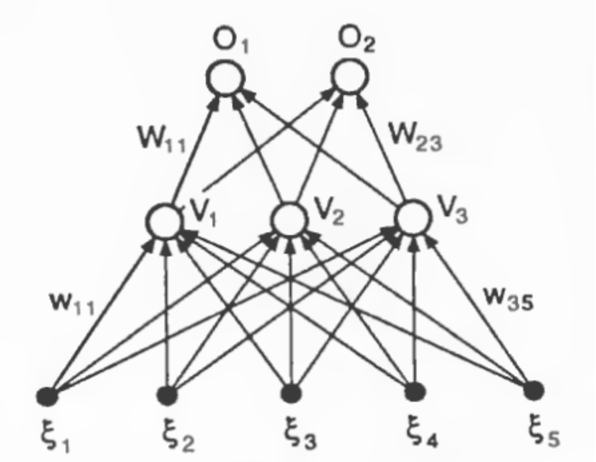
$$E(w) = \frac{1}{2}\sum_{i\mu}\left\|\lambda_i^\mu - O_i^\mu\right\|^2 \qquad O_i = f\left(\sum_j W_{ij}V_j\right) \qquad V_j = f\left(\sum_k w_{jk}\xi_k\right)$$

$$\Delta W_{ij} = -\eta\frac{\partial E}{\partial W_{ij}} \qquad \Delta w_{jk} = -\eta\frac{\partial E}{\partial w_{jk}} = -\eta\sum_\mu\frac{\partial E}{\partial V_j}\frac{\partial V_j}{\partial w_{jk}}$$

$$\Delta W_{ij} = \eta\sum_\mu(\lambda_i^\mu - O_i^\mu)f'(\cdot)V_j \qquad \Delta w_{jk} = \eta\sum_\mu\sum_i(\lambda_i^\mu - O_i^\mu)f'(\cdot_i)W_{ij}f'(\cdot_j)\xi_j$$

$$\Delta W_{ij} = \eta\sum_\mu\boxed{(\lambda_i^\mu - O_i^\mu)f'(\cdot_i)}V_j$$

$$\Delta w_{jk} = \eta\sum_\mu f'(\cdot_j)\xi_j\sum_i\boxed{(\lambda_i^\mu - O_i^\mu)f'(\cdot_i)}W_{ij}$$

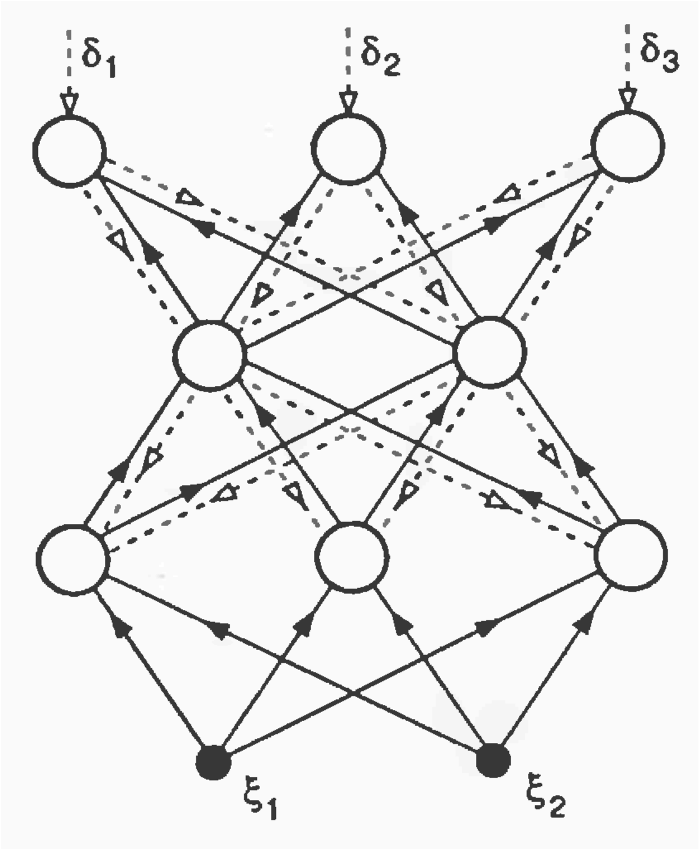


$$\Delta W_{ij} = \eta \sum_{\mu} \boxed{(\lambda_i^{\mu} - O_i^{\mu}) f'(\cdot_i)} V_j$$

$$\delta_i = (\lambda_i^{\mu} - O_i^{\mu}) f'(\cdot_i)$$

$$\Delta w_{jk} = \eta \sum_{\mu} f'(\cdot_j) \xi_k \sum_{i} \boxed{(\lambda_i^{\mu} - O_i^{\mu}) f'(\cdot_i)} W_{ij}$$

$$\delta_j = f'(\cdot_j) \sum_{i} \delta_i W_{ij}$$

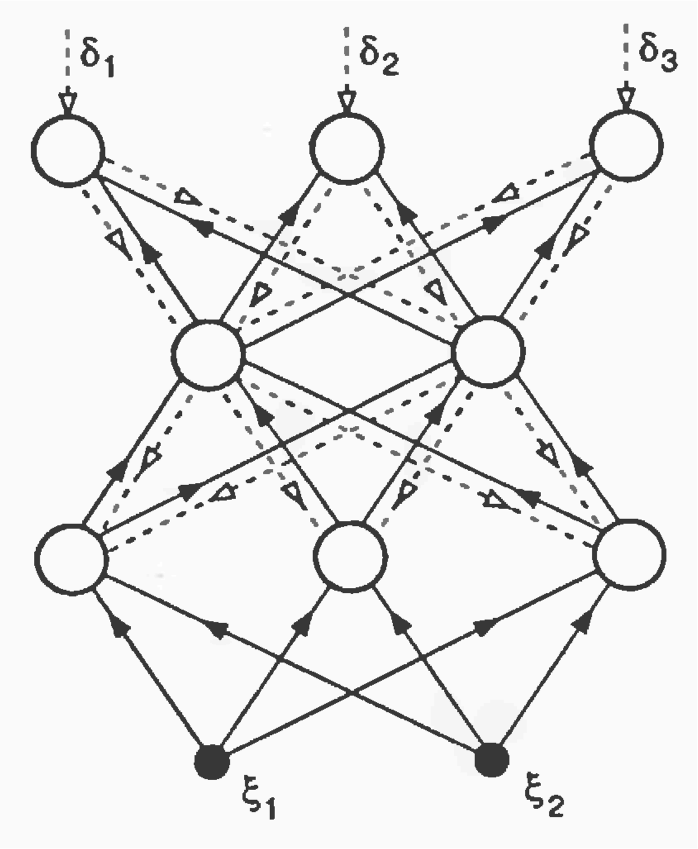


$$\Delta W_{ij} = \eta \sum_{\mu} \delta_i V_j$$

$$\Delta w_{jk} = \eta \sum_{\mu} \delta_j \xi_k$$

Back - Propagation

1986 – David Rumelhart: Back-Propagation aplicado a redes de neuronas artificiales



$$\delta_i = (\lambda_i^\mu - O_i^\mu) f'(\cdot_i)$$

$$\delta_j = f'(\cdot_j) \sum_i \delta_i W_{ij}$$

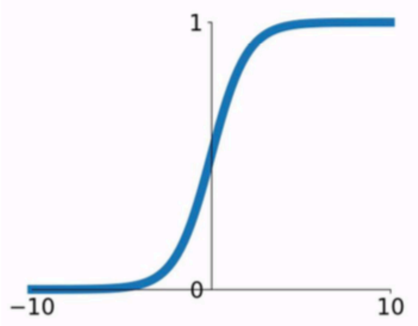
$$\Delta W_{ij} = \eta \sum_\mu \delta_i V_j$$

$$\Delta w_{jk} = \eta \sum_\mu \delta_j \xi_k$$

Activation functions $f: y = f(w^T x + b)$

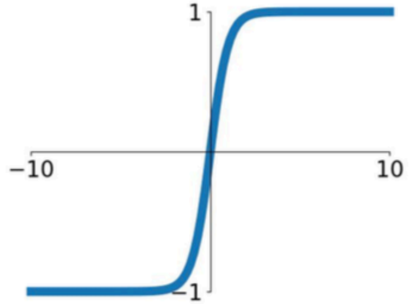
Sigmoid

$\sigma(x) = \frac{1}{1+e^{-x}}$



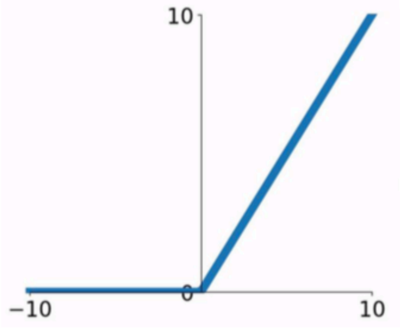
tanh

$\tanh(x)$



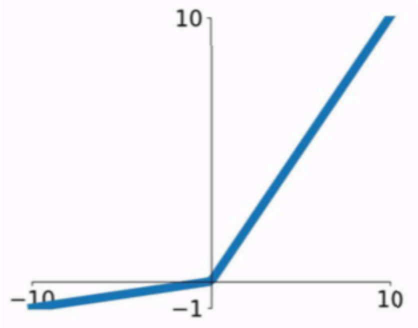
ReLU

$\max(0, x)$



Leaky ReLU

$\max(0.1x, x)$

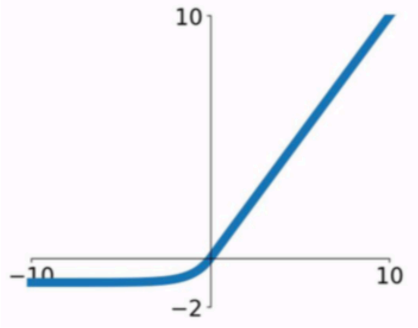


Maxout

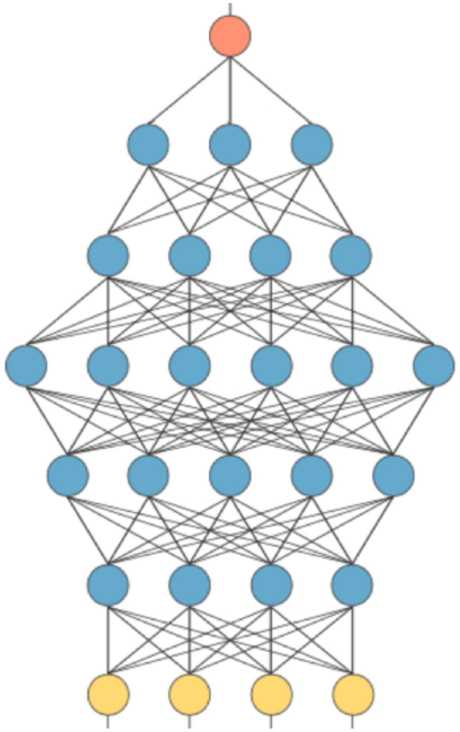
$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU

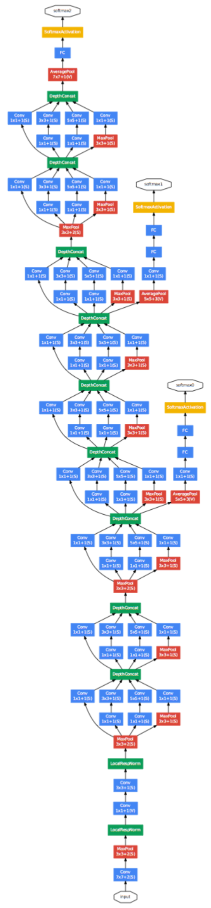
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Backpropagation y …



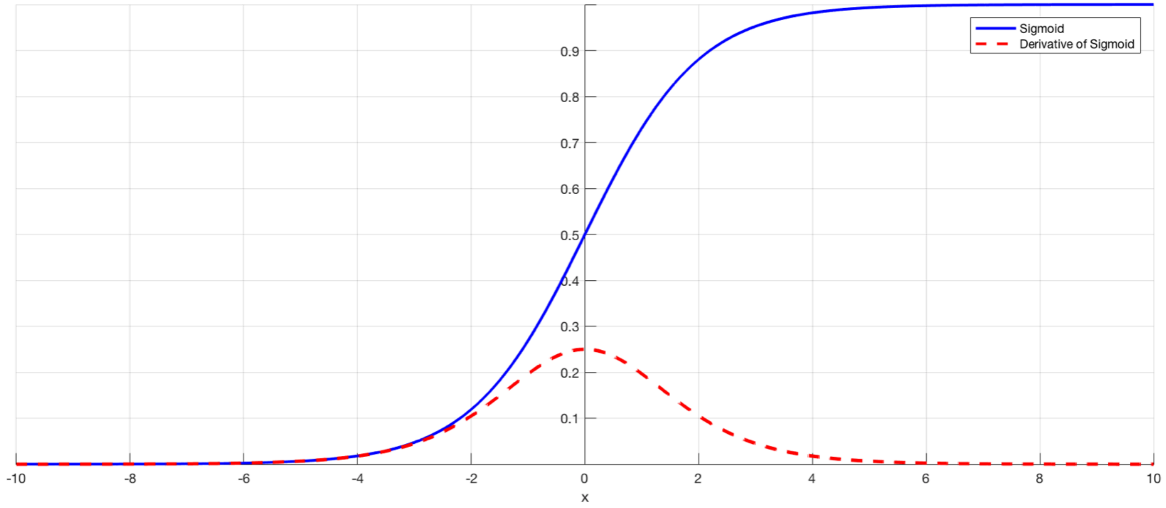
6 layers



30 layers

SIGMOIDEA

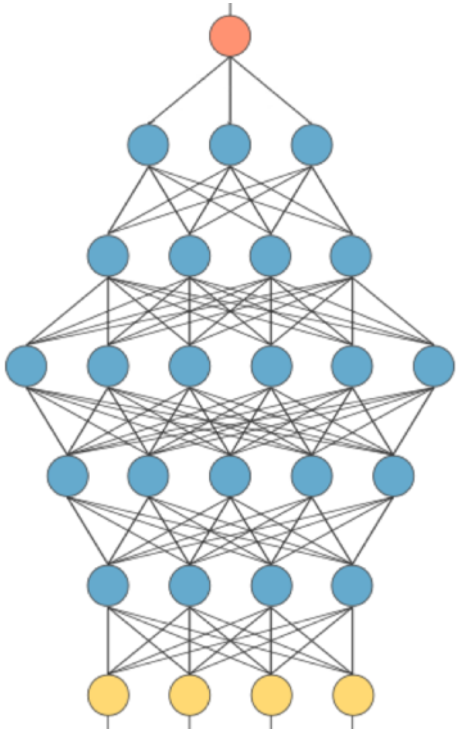
$$f(x) = \frac{1}{1+e^{-x}} \qquad f'(x) = f(x) * (1 - f(x))$$



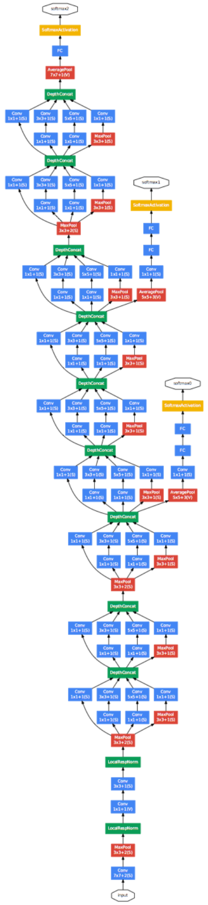
$$\delta_i = (\lambda_i^{\mu} - O_i^{\mu}) f'(\cdot_i)$$

$$\delta_j = f'(\cdot_j) \sum_i \delta_i W_{ij} = \sum_i \underline{f'(\cdot_j)\, f'(\cdot_i)}\, (\lambda_i^{\mu} - O_i^{\mu}) W_{ij}$$

Backpropagation y … ReLU – ELU las más usadas en Aprendizaje Profundo



6 layers

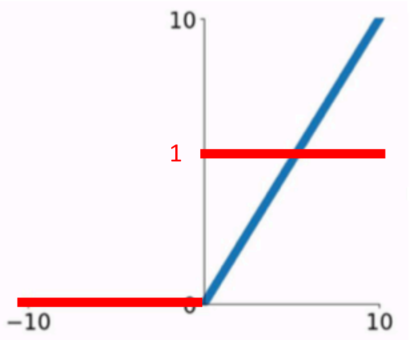


30 layers

$$f_{ReLU}(x) = \max(0, x) \qquad f'_{RELU}(x) = \begin{cases} 1 & x > 0 \\ 0 & x < 0 \end{cases}$$

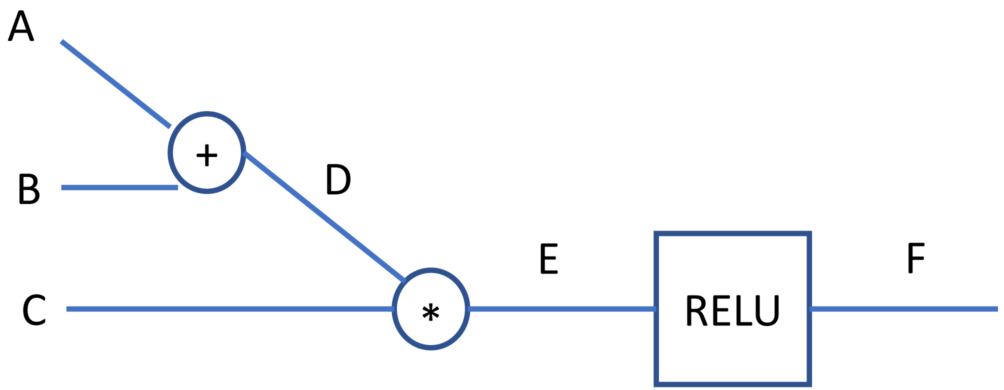
ReLU

$\max(0, x)$



$$\delta_i = (\lambda_i^\mu - O_i^\mu) f'(\cdot_i)$$

$$\delta_j = f'(\cdot_j) \sum_i \delta_i W_{ij} = \sum_i \underline{f'(\cdot_j)\, f'(\cdot_i)}\, (\lambda_i^\mu - O_i^\mu) W_{ij}$$



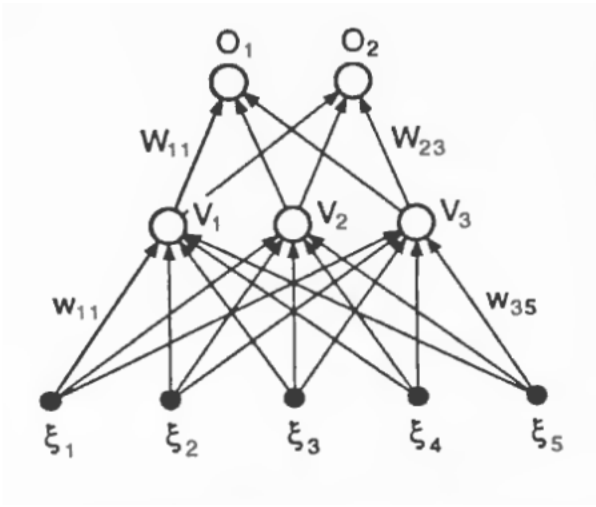
Calcular:

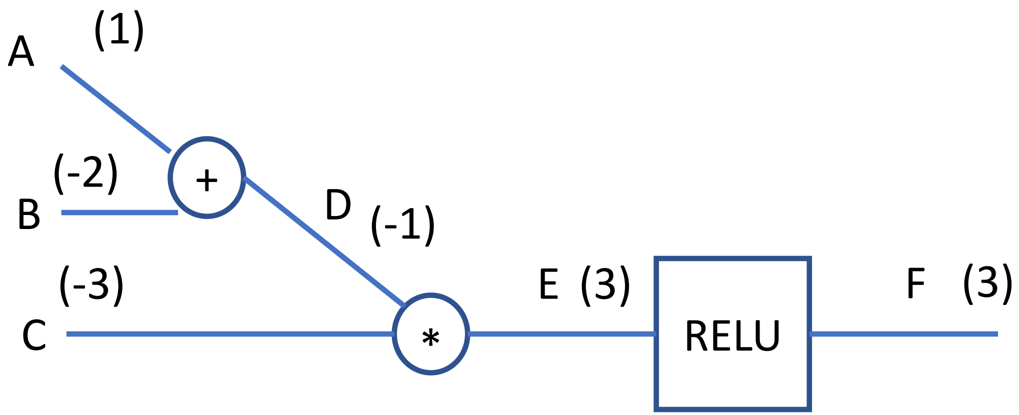
$$\frac{\partial F}{\partial A} \qquad \frac{\partial F}{\partial B} \qquad \frac{\partial F}{\partial C}$$

A=1   B=-2   C = -3

$$\Delta W_{ij} = \eta \sum_{\mu} \delta_i V_j \qquad \Delta w_{jk} = \eta \sum_{\mu} \delta_j \xi_k$$

$$\delta_i = (\lambda_i^{\mu} - O_i^{\mu}) f'(\cdot_i) \qquad \delta_j = f'(\cdot_j) \sum_i \delta_i W_{ij}$$





Calcular:

$$\frac{\partial F}{\partial A} \qquad \frac{\partial F}{\partial B} \qquad \frac{\partial F}{\partial C}$$

A=1   B=-2   C = -3

$$\frac{\partial F}{\partial E} = 1$$

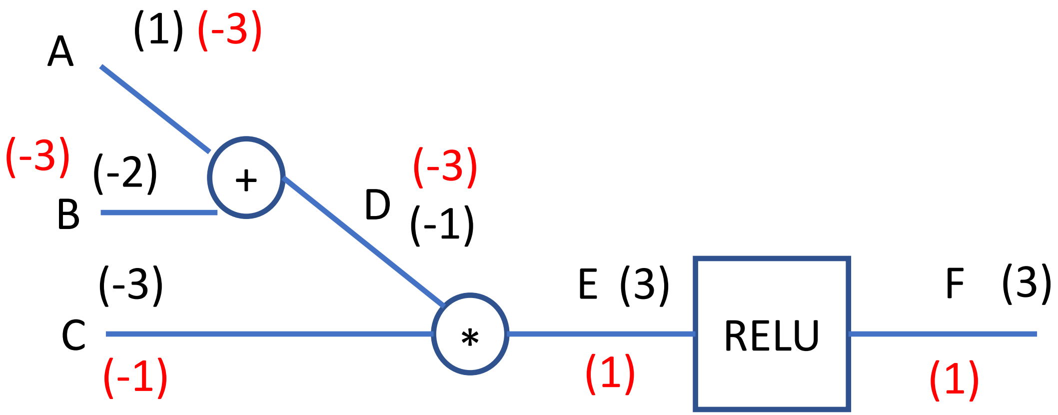
$$\frac{\partial E}{\partial C} = D \qquad \frac{\partial E}{\partial D} = C$$

$$\frac{\partial D}{\partial A} = 1 \qquad \frac{\partial D}{\partial B} = 1$$

$$\frac{\partial F}{\partial A} = \frac{\partial F}{\partial E}\frac{\partial E}{\partial D}\frac{\partial D}{\partial A}$$

$$\frac{\partial F}{\partial B} = \frac{\partial F}{\partial E}\frac{\partial E}{\partial D}\frac{\partial D}{\partial B}$$

$$\frac{\partial F}{\partial C} = \frac{\partial F}{\partial E}\frac{\partial E}{\partial C}$$



Calcular:

$$\frac{\partial F}{\partial A} \qquad \frac{\partial F}{\partial B} \qquad \frac{\partial F}{\partial C}$$

A=1 B=-2 C = -3

$$\frac{\partial F}{\partial E} = 1$$

$$\frac{\partial E}{\partial C} = D \qquad \frac{\partial E}{\partial D} = C$$

$$\frac{\partial D}{\partial A} = 1 \qquad \frac{\partial D}{\partial B} = 1$$

$$\frac{\partial F}{\partial A} = \frac{\partial F}{\partial E}\frac{\partial E}{\partial D}\frac{\partial D}{\partial A} = C = -3$$

$$\frac{\partial F}{\partial B} = \frac{\partial F}{\partial E}\frac{\partial E}{\partial D}\frac{\partial D}{\partial B} = C = -3$$

$$\frac{\partial F}{\partial C} = \frac{\partial F}{\partial E}\frac{\partial E}{\partial C} = D = -1$$

Gracias