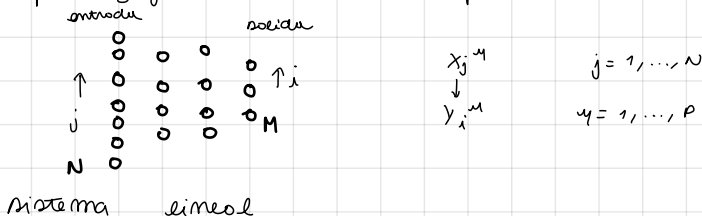


Aprendizaje en redes multicapa



sistema lineal

$$g(x) = x$$

$$O_i^m = \sum_{j=1}^N w_{ij}^m x_j^m$$

Annotations:
 - w_{ij}^m is labeled "SALIDA ENTADA"
 - x_j^m is labeled "salida real"

$$w_{ij}^m$$

Poderemos encontrar los que $O_i^m =$

Annotations:
 - y_i^m is circled and labeled "salida deseada"
 - w_{ij}^m is labeled "w"

Veamos que w_{ij} debe ser de la forma:

$$w_{ij} = \frac{1}{N} \sum_{m,v=1}^M y_i^m (Q^{-1})_{mv} x_j^v$$

Annotations:
 - $Q_{mv} = \frac{1}{N} \sum_j x_j^m x_j^v$
 - Q is labeled "debe ser de la forma"
 - Q is labeled "debe ser de la forma"
 - Q is labeled "debe ser de la forma"

es el que cumple con $O_i^m = y_i^m$

$$\sum_{j=1}^N w_{ij} x_j^m = \frac{1}{N} \sum_{m,v,j} y_i^m (Q^{-1})_{mv} x_j^v x_j^m$$

$$= \sum_{m,v} y_i^m \underbrace{(Q^{-1})_{mv} Q_{mv}}_{\delta_{mv}} = y_i^m$$

Aprendizaje por gradiente

$$E = \frac{1}{2} \sum_{m,v} (y_i^m - O_i^m)^2 \quad E \geq 0$$

$$E = 0 \Leftrightarrow y_i^m = O_i^m \quad i = 1, \dots, M \quad m = 1, \dots, P$$

Resolver el problema de aprendizaje \Leftrightarrow mínimo de E

$$\bar{w}_0$$

$$\bar{w}_0 + \Delta \bar{w}$$

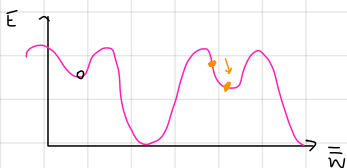
$$\Rightarrow E(\bar{w}_0 + \Delta \bar{w}) = E(\bar{w}) + \overline{\nabla} E(\bar{w}) \cdot \Delta \bar{w} + \mathcal{O}(\Delta \bar{w}^2)$$

tomamos $\Delta \bar{w} = -\eta \overline{\nabla} E(\bar{w})$ para que (*) sea negativo

$$= E(\bar{w}) - \eta |\overline{\nabla} E(\bar{w})|^2 + \mathcal{O}(\Delta \bar{w}^2) \leq E(\bar{w})$$

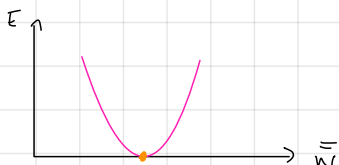
para asegurar esto pedimos η pequeño

grafiquemos una proyección de \bar{w}



el método converge al mínimo local más cercano a los C.I

Para sistemas lineales E es cuadrática



\Rightarrow el sistema es globalmente convergente

Para estos sist. tenemos:

$$E(\bar{w}) = \frac{1}{2} \sum_{i, \mu} \left(y_i^\mu - \sum_j w_{ij} x_j^\mu \right)^2$$

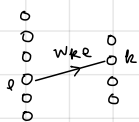
$$\frac{\partial E}{\partial w_{kl}} = - \sum_{\mu} \underbrace{\left(y_k^\mu - \sum_j w_{kj} x_j^\mu \right)}_{\delta l^\mu} x_l^\mu$$

son continuos (mo 0,1 como en el perceptron)

$$\Delta w_{kl} = \eta \sum_{\mu=1}^P \delta l^\mu x_l^\mu$$

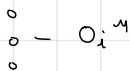
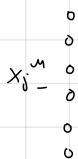
$$\delta l^\mu = y_k^\mu - o_k^\mu$$

regla de Adeline



conexión efectiva entre neuronas l de entrada y k de salida : w_{kl}

una capa, no lineal



$$o_i^u = g \left(\sum_{j=1}^N w_{ij} x_j^u \right)$$

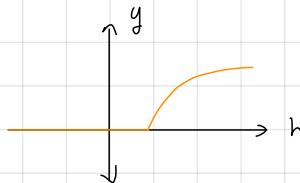
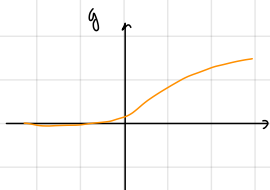
$$E = \frac{1}{2} \sum_{i,u} \left(y_i^u - g \left(\sum_{j=1}^N w_{ij} x_j^u \right) \right)^2$$

$$\frac{\partial E}{\partial w_{kl}} = \sum_u \left(y_k^u - g \left(\sum_{j=1}^N w_{kj} x_j^u \right) \right) g' \left(\sum_{j=1}^N w_{kj} x_j^u \right) x_l^u$$

$\underbrace{\hspace{10em}}_{\delta_k^u} \quad \underbrace{\hspace{10em}}_{h_k^u}$

$$= - \sum_u \delta_k^u g'(h_k^u) x_l^u$$

posibles g :



si $g(h) = \tanh(h) \Rightarrow g'(h) = 1 - g^2(h)$

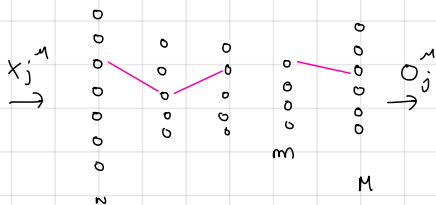
Podríamos definir la esperanza de otra manera:

$$E = \sum_{i,u} \left[\frac{1 + y_i^u}{2} \ln \left(\frac{1 + y_i^u}{1 + o_i^u} \right) + \frac{1 - y_i^u}{2} \ln \left(\frac{1 - y_i^u}{1 - o_i^u} \right) \right]$$

$$\frac{\partial E}{\partial w_{kl}} = \sum_u \frac{g'(h_k^u)}{(1 - o_k^u)^2} [y_k^u - o_k^u] x_l^u$$

se redefinió δ pero la forma funcional es la misma

Caso general: redes multicapa no-lineales



O_i^u : funciones de los
entrados ($\bar{w}_{ij} x_j^u$)

$$E = \frac{1}{2} \sum_{i,j,u} (y_i^u - O_i^u(\bar{w}_{ij} x_j^u))^2$$

$$\Delta w_{kl} = -\eta \frac{\partial E}{\partial w_{kl}}$$

numéricamente:

$$\frac{\partial E}{\partial w_{kl}} \approx \frac{E(w_{kl} + \Delta) - E(w_{kl})}{\Delta}$$

tiempo de cálculo d n° de
parámetros (P)
x
n° total
de conexiones

Back-propagation

1-D $f(g(x))' = f'(g(x)) g'(x)$

N-D $f(g_1(x), \dots, g_n(x))' = \sum_i \frac{\partial f}{\partial g_i} (g_1(x), \dots, g_n(x)) g_i'(x)$

N-N-D
$$\frac{\partial}{\partial x_k} f(g_1(x_1, \dots, x_n), g_2(x_1, \dots, x_n), \dots, g_n(x_1, \dots, x_n))$$

$$= \sum_i \frac{\partial f}{\partial g_i} (g_1, \dots, g_n) \cdot \frac{\partial g_i}{\partial x_k} (x_1, \dots, x_n)$$

tenemos una función E , sobre la cual queremos
calcular el gradiente $\frac{\partial E}{\partial w_{kl}}$

E : función de los neuronas de alguna capa
salida de la capa

$$E(\{O_i^u\}) = E\left(g\left(\sum_{h,i} w_{ih} O_h^u\right)\right), \frac{\partial E}{\partial w_{kl}} = \frac{\partial E}{\partial h_k^u} \frac{\partial h_k^u}{\partial w_{kl}} = \frac{\partial E}{\partial h_k^u} g'(h_k^u) O_l^u$$

- δ_k^u
 ~
 matri-
 que
 el
 índice
 de esta
 neurona

al comenzar la salida, se comienza a introducir de las neuronas de la capa siguiente m .

$$\frac{\partial E}{\partial h_k^m} = \sum_m \frac{\partial E}{\partial h_m^m} \underbrace{\frac{\partial h_m^m}{\partial h_k^m}}_{\text{como varia la salida de la capa respecto a una variación en la salida de la capa anterior}} \quad h_m^m = \sum_s w_{ms} o_s^m = \sum_s w_{ms} g(h_s^m)$$

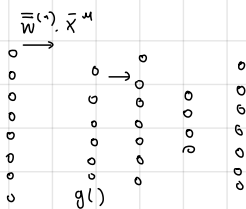
$$\rightarrow \frac{\partial h_m^m}{\partial h_k^m} = w_{mk} g'(h_k^m)$$

$$\frac{\partial E}{\partial h_k^m} = \sum_m \left(\frac{\partial E}{\partial h_m^m} \right) w_{mk} g'(h_k^m)$$

↓ debemos calcular esto → en columnas para la última capa y vamos para atrás

$$E = \frac{1}{2} \sum_{k,m} (y_k^m - o_k^m)^2 = \frac{1}{2} \sum (y_k^m - g(h_k^m))^2$$

$$\Rightarrow \frac{\partial E}{\partial h_m^m} = - \sum_k \underbrace{(y_k^m - g(h_k^m))}_{\text{errores}} g'(h_k^m) \rightarrow \text{para la última capa}$$



$$(\bar{y} - \bar{o}) \otimes g'(\bar{h})$$

↓
prod.
pto
pto

Paso:

introduce \bar{x}^m : multiplicamos por $\bar{w}^{(1)}$ y obtenemos la salida parcial $\bar{o}^1 = \bar{w}^{(1)} \bar{x}^m$

