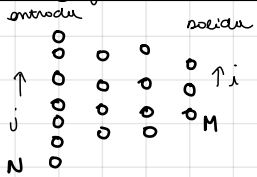
Aprendizaje en redes multicapa



x_j^μ $\quad j = 1, \dots, N$

$\downarrow$

y_i^μ $\quad \mu = 1, \dots, p$

sistema lineal

$g(x) = x$

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

(w_{ij}: SALIDA; x_j^μ: ENTRADA)

O_i^μ: salida real

w_{ij} ( )

queremos encontrar w_{ij} tales que $O_i^\mu = y_i^\mu$ (salida deseada)

Veamos que w_{ij} dada por:

$$w_{ij} = \frac{1}{N} \sum_{\mu,\nu=1} y_i^\mu (Q^{-1})_{\mu\nu} x_j^\nu$$

debe ser bi (si $p < N$)

$$Q_{\mu\nu} = \frac{1}{N} \sum_j x_j^\mu x_j^\nu \quad (p \times p)$$

es tal que cumple con $O_i^\mu = y_i^\mu$

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

$$= \sum_{\mu,\nu} y_i^\mu \underbrace{(Q^{-1})_{\mu\nu} Q_{\mu\nu}}_{\delta_{\mu\nu}} = y_i^\mu$$

aprendizaje por gradiente

$$E = \frac{1}{2} \sum_{\nu,\mu} (y_i^\mu - O_i^\mu)^2 \qquad E \geq 0$$

$E = 0 \iff y_i^\mu = O_i^\mu \quad i = 1, \dots \mu \quad \mu = 1, \dots, p$

Resolver el problema de aprendizaje $\Leftrightarrow$ mínimo abs de E

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

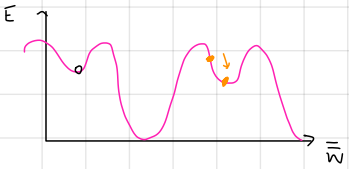
$$\Rightarrow \quad E(\bar{\bar{w}}_0 + \Delta\bar{\bar{w}}) = E(\bar{\bar{w}}) + \underbrace{\bar{\nabla} E(\bar{\bar{w}}) \cdot \Delta\bar{\bar{w}}}_{(*)} + \mathcal{O}(\Delta\bar{\bar{w}}^2)$$

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

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

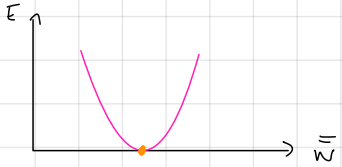
para asegurar esto pedimos η pequeño

grafiquemos una proyección de $\bar{\bar{w}}$



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:

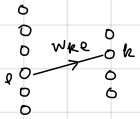
$$E(\bar{\bar{w}}) = \frac{1}{2} \sum_{i,\mu} \left(Y_i^\mu - \sum_j W_{ij} x_j^\mu \right)^2$$

$$\frac{\partial E}{\partial W_{kl}} = -\sum_\mu \underbrace{\left(Y_k^\mu - \underbrace{\sum_j W_{kj} x_j^\mu}_{O_k^\mu} \right)}_{\delta_k^\mu} x_l^\mu$$

son continuos (no 0, 1 como en el perceptrón)

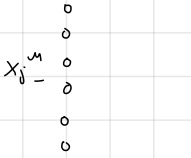
$$\boxed{\Delta W_{kl} = \eta \sum_{\mu=1}^{P} \delta_k^\mu x_l^\mu \qquad \delta_{k\mu} = Y_k^\mu - O_k^\mu}$$

regla δ Adeline



conexión efectiva entre neuronas e de entrada y k de salida : w_{ke}

una capa, no lineal



$$O_i^\mu = g\left(\sum_{j=1}^{N} w_{ij}\, x_j^\mu\right) \quad \text{cont.}$$

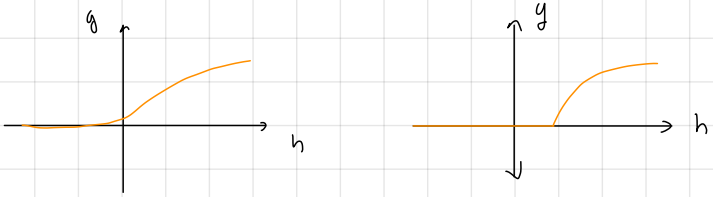
$$E = \frac{1}{2}\sum_{i,\mu}\left(Y_i^\mu - g\left(\sum_{j=1}^{N} w_{ij}\, x_j^\mu\right)\right)^2$$

$$\frac{\partial E}{\partial w_{ke}} = \sum_\mu \left(Y_k^\mu - g\left(\sum_{j=1}^{N} w_{kj}\, x_j^\mu\right)\right) g'\left(\sum_j w_{kj}\, x_j^\mu\right) x_e^\mu$$

(O_k^μ : $g\left(\sum_{j=1}^{N} w_{kj}\, x_j^\mu\right)$; $h_{k\mu}$: $\sum_j w_{kj}\, x_j^\mu$; δ_k^μ : $\left(Y_k^\mu - g\left(\sum_{j=1}^{N} w_{kj}\, x_j^\mu\right)\right)$)

$$= -\sum_\mu \delta_k^\mu\, g'(h_k^\mu)\, x_e^\mu$$

posibles g:



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

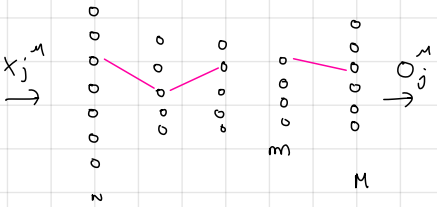
Podríamos definir la esperanza de otra manera:

$$E = \sum_{i,\mu}\left[\frac{1+Y_i^\mu}{2}\ln\left(\frac{1+Y_i^\mu}{1+O_i^\mu}\right) + \frac{1-Y_i^\mu}{2}\ln\left(\frac{1-Y_i^\mu}{1-O_i^\mu}\right)\right]$$

$$\frac{\partial E}{\partial w_{ke}} = \sum_\mu \frac{g'(h_k^\mu)}{\left(1 - O_k^\mu\right)^2}\left[Y_k^\mu - O_k^\mu\right] x_e^\mu$$

se redefinió δ pero la forma funcional es la misma

Caso general: redes multicapa no - lineales



O_i^μ : funciones de los entrados $(\overline{\overline{w}}_j x_j^\mu)$

$$E = \frac{1}{2} \sum_{i,\mu} \left(y_i^\mu - O_i^\mu (\overline{\overline{w}}_j \overline{x}^\mu) \right)^2$$

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

numericamente: $O(P)$ $O(P)$

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

tiempo de cálculo $\propto$ nº de patrones (P) $\times$ nº total de conexiones

Back-propagation

$(1-D)$ $\quad f(g(x))' = f'(g(x))\, g'(x)$

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

$(N-N)-D$ $\quad \frac{\partial}{\partial x_k} f(g_1(x_1, \dots, x_N), g_2(x_1, \dots, x_N), \dots, g_N(x_1, \dots, x_N))$

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

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

E_i: función de los neuronas de alguna capa

$$E\left(\left| O_i^\mu \right|\right) = E\left(g\left(\underbrace{\sum w_{il} O_l^\mu}_{h_i^\mu} \right)\right), \quad \frac{\partial E}{\partial w_{kl}} = \frac{\partial E}{\partial h_k^\mu} \frac{\partial h_k^\mu}{\partial w_{kl}} = \underbrace{\frac{\partial E}{\partial h_k^\mu}}_{-\delta_k^\mu} g'(h_k^\mu) O_l^\mu$$

salida de la capa

notar que el índice l está repetido

al cambiar la salida, se cambia de entradas de las neuronas de la capa siguiente m.

$$\frac{\partial E}{\partial h_k^{m-1}} = \sum_m \frac{\partial E}{\partial h_m^m} \frac{\partial h_m^m}{\partial h_k^{m-1}} \qquad h_m^m = \sum_s w_{ms} o_s^{m-1} = \sum_s w_{ms} g(h_s^{m-1})$$

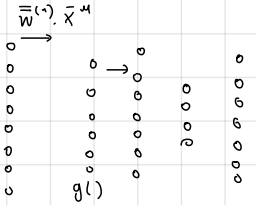
$\frac{\partial h_m^m}{\partial h_k^{m-1}}$: como varía la salida de la capa respecto a una variación en la salida de la capa anterior $\rightarrow \frac{\partial h_m^m}{\partial h_k^{m-1}} = w_{mk} g'(h_k^{m-1})$

$$\frac{\partial E}{\partial h_k^{m-1}} = \sum_m \frac{\partial E}{\partial h_m^m} w_{mk} g'(h_k^{m-1})$$

↓ debemos calcular esto → lo calculamos para la última capa y vamos para atrás

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

$$\Rightarrow \frac{\partial E}{\partial h_m^m} = - \sum_m \underbrace{(y_m^m - \overbrace{g(h_m^m)}^{o_m^m})}_{\text{errores}} g'(h_m^m) \rightarrow \text{para la última capa}$$



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

↓ producto a producto