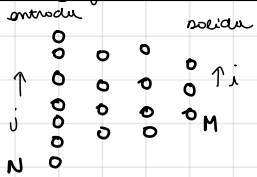
Aprendizaje en redes multicapa



x_j^μ $\quad j = 1, \dots, N$

$\downarrow$

$y_i^\mu \quad \mu = 1, \dots, p$

sistema lineal

$g(x) = x$ — función de activación de una neurona

$O_i^\mu = \sum_{j=1}^{N} w_{ij} x_j^\mu$ — salida real (SALIDA, ENTRADA)

w_{ij} ( )

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

Veamos que w_{ij} dada por: (debe ser $\exists$ si $p < N$)

$$w_{ij} = \frac{1}{N} \sum_{\mu,\nu=1} y_i^\mu (Q^{-1})_{\mu\nu} x_j^\nu \qquad Q_{\mu\nu} = \frac{1}{N} \sum_j x_j^\mu x_j^\nu \quad (P \times P)$$

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 $\iff$ mínimo abs de E

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

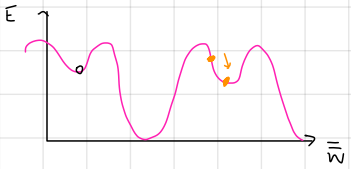
$$\Rightarrow E(\bar{\bar{w}}_0 + \Delta\bar{\bar{w}}) = E(\bar{\bar{w}}) + \underbrace{\bar{\nabla} E(\bar{\bar{w}}) \cdot \Delta\bar{\bar{w}}}_{(*)} + O(\Delta\bar{\bar{w}}^2)$$

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

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

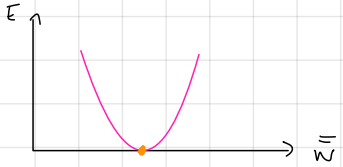
↓ para asegurar esto pedimos η pequeño

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



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

Para sistemas lineales E es cuadrática



⇒ el sistema es globalmente convergente

Calculemos de manera explícita $E(\bar{\bar{w}})$ para sist lineales $(g(x) = x)$

$$E(\bar{\bar{w}}) = \frac{1}{2} \sum_{i,\mu} \left(Y_i^\mu - \sum_j W_{ij} x_j^\mu \right)^2 = \dots + \underbrace{\left(Y_k^\mu - \sum_j W_{kj} x_j^\mu \right)^2}_{O_k^\mu} + \dots$$

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

↓ solo el sumando con $i = k$ contiene a W_{ke}

δ_k^μ término de error

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

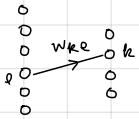
$$\Delta W_{ke} = \eta \sum_{\mu=1}^{P} \delta_k^\mu x_e^\mu$$

entrada

regla δ Adeline

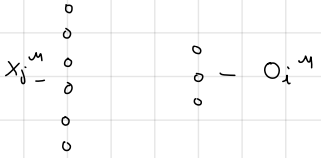
$$\delta_{k\mu} = Y_k^\mu - O_k^\mu$$

↓ dif salida deseada y real → error en la salida



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

una capa, no lineal (entrada - salida)



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

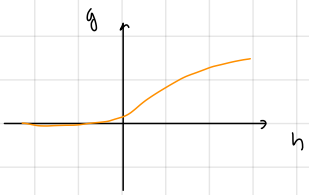
$$E = \frac{1}{2}\sum_{i,\mu}\left(Y_i^\mu - g\left(\sum_{j=1}^{N} w_{ij}\, x_j^\mu\right)\right)^2$$

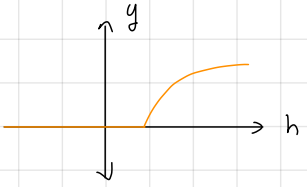
$$\frac{\partial E}{\partial w_{ke}} = \sum_\mu \left(Y_k^\mu - g\left(\sum_{j=1}^{N} w_{kj}\, x_j^\mu\right)\right) g'\left(\sum_j w_{kj}\, x_j^\mu\right) x_e^\mu$$

O_k^μ ; δ_k^μ ; h_k^μ = entrada a neurona k en la última capa

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

posibles g: (## integrate and fin)





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

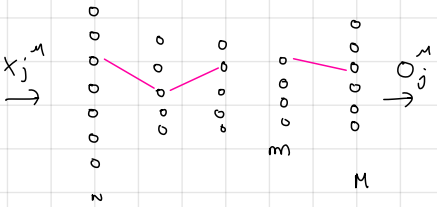
Podríamos definir la esperanza de otra manera:

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

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

se redefinió δ pero la forma funcional es la misma

Caso general: redes multicapa no - lineales



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

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

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

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-propogation

regla de la cadena

calcula los gradientes para todos los conexiones a la vez

impracticable!

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

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

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

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

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

E: función de los neuronas de alguna capa

salida de la capa

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

notar que el índice l está repetido

al comienzo la salida, se comienza de, entradas de las neuronas de la capa siguiente m.

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

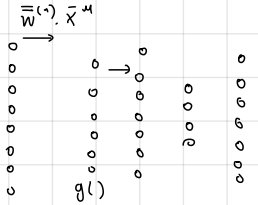
$\frac{\partial h_m^m}{\partial h_k^{m-1}}$ → como varía la salida de la capa respecto a una variación en la salida de la capa anterior → $\frac{\partial h_m^m}{\partial h_k^{m-1}} = w_{mk}\, g'(h_k^{m-1})$

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

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

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

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



$(\bar{y} - \bar{o}) \otimes g'(\bar{h})$ → prod. pto a pto

- aplico los w a los entrados
- calculo $w_1 x$
- aplico $g_1 \rightarrow g(w_1 x)$
- calculo $w_2\, g_1(w_1 x)$
- aplico $g_2 \rightarrow g_2(w_2\, g_1(w_1 x))$
- llego a la última capa y obtengo los salidos $\bar{o}$
- calculo $\bar{y} - \bar{o}$
- luego prod pto a pto con la $g'(\bar{h})$: $(\bar{y} - \bar{o}) \otimes g'(\bar{h})$

$\left((a,b) \cdot (c,d) = (ac, bd) \right)$

Para la ult capa

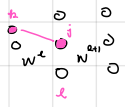
$$\frac{\partial E}{\partial h_m^{m-1}} = -\sum_m \left[(y_m^m - g(h_m^m)) \right] g'(h_m^{m-1})$$

$$\frac{\partial E}{\partial h_{m-1}} = \frac{\partial E}{\partial h_m} \frac{\partial h_m}{\partial h_{m-1}}$$

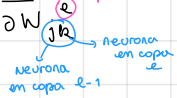
$$\frac{\partial E}{\partial h_{m-1}} = \frac{\partial E}{\partial h_m} w_{m,k} g'(h_k)$$

$$\frac{\partial E}{\partial h_{m-2}} = \frac{\partial E}{\partial h_{m-1}} \frac{\partial h_{m-1}}{\partial h_{m-2}} = \frac{\partial E}{\partial h_{m-1}} w_{m-1,k} g'$$

Queremos minimizar la función de coste E



$$\frac{\partial E}{\partial w^{\ell}_{jk}} = \frac{\partial E}{\partial h^{\ell}_{j}} \frac{\partial h^{\ell}_{j}}{\partial w^{\ell}_{jk}} \qquad \text{con} \quad \boxed{h^{\ell}_{j} = \sum_{k=1}^{m} w^{\ell}_{jk} O^{\ell-1}_{k} + b^{\ell}_{j}}$$



m: # de neuronas en la capa $\ell-1$

($\frac{\partial E}{\partial w^{\ell}_{jk}}$: como varía E al cambiar una conexión $j-k$ en capa ℓ

$\frac{\partial E}{\partial h^{\ell}_{j}}$ como varía E con el input de la neurona j en capa ℓ

$\frac{\partial h^{\ell}_{j}}{\partial w^{\ell}_{jk}}$ como varía el input de la neurona j en ℓ con la conexión w^{ℓ}_{jk})

$$\frac{\partial h^{\ell}_{j}}{\partial w^{\ell}_{kj}} = O^{\ell-1}_{k}$$

gradiente local

$$\boxed{\frac{\partial E}{\partial w^{\ell}_{jk}} = \frac{\partial E}{\partial h^{\ell}_{j}} O^{\ell-1}_{k} = \delta^{\ell}_{j} O^{\ell-1}_{k}}$$

Al mismo tiempo, para los pesos

$$\frac{\partial E}{\partial b^{\ell}_{j}} = \frac{\partial E}{\partial h^{\ell}_{j}} \underbrace{\frac{\partial h^{\ell}_{j}}{\partial b^{\ell}_{j}}}_{1} \Rightarrow \boxed{\frac{\partial E}{\partial b^{\ell}_{j}} = \frac{\partial E}{\partial h^{\ell}_{j}} = \delta^{\ell}_{j}}$$

E puede ser MSE : $E = \frac{1}{2} \sum_{j\ell} (y^{\ell}_{j} - O^{\ell}_{j})^2$

activación

En ese caso: $\frac{\partial E}{\partial h^{\ell}_{j}} = \frac{\partial E}{\partial O^{\ell}_{j}} \frac{\partial O^{\ell}_{j}}{\partial h^{\ell}_{j}}$ $\rightarrow O^{\ell}_{j} = g(h^{\ell}_{j})$

$$= \left[(y^{\ell}_{j} - O^{\ell}_{j}) \right] g'(h^{\ell}_{j})$$

$$\Rightarrow \quad \delta^{\ell}_{j} = (y^{\ell}_{j} - O^{\ell}_{j})\, g'(h^{\ell}_{j})$$

$$\Rightarrow \quad \frac{\partial E}{\partial w^{\ell}_{jk}} = \underbrace{(y^{\ell}_{j} - O^{\ell}_{j})}_{\text{conocido para la última capa}} \cdot O^{\ell-1}_{k}\, g'(h^{\ell}_{j})$$

$$\Rightarrow \quad \Delta \bar{W} = -\eta \frac{\partial E}{\partial \bar{W}} \qquad \Rightarrow \qquad \Delta W_{jk}^{\ell} = -\eta \frac{\partial E}{\partial W_{jk}^{\ell}}$$

$$\boxed{\Delta W_{jk}^{\ell} = -\eta \left(y_j^{\ell} - o_j^{\ell}\right) o_k^{\ell-1} g'(h_j^{\ell})}$$

Para la antepenúltima capa

$$\frac{\partial E}{\partial h_j^{\ell-1}} = \sum_m \frac{\partial E}{\partial h_m^{\ell}} \frac{\partial h_m^{\ell}}{\partial h_j^{\ell-1}}$$

$$\boxed{h_j^{\ell} = \sum_{k=1}^{m} w_{jk}^{\ell} o_k^{\ell-1} + b_j^{\ell} = \sum_k w_{jk} g(h_k^{\ell}) + b_j^{\ell}}$$

$$\frac{\partial h_m^{\ell}}{\partial h_j^{\ell-1}} = w_{mj} \, g'(h_j^{\ell})$$

$$\Rightarrow \quad \frac{\partial E}{\partial h_j^{\ell-1}} = \sum_m \frac{\partial E}{\partial h_m^{\ell}} w_{jm} \, g'(h_j^{\ell-1}) \qquad \frac{\partial E}{\partial h_j^{\ell}}$$

$$\Delta W_{jk}^{\ell-1} = -\eta \frac{\partial E}{\partial W_{jk}^{\ell}} = -\eta \left(\frac{\partial E}{\partial h_j^{\ell}} o_k^{\ell-2} \right)$$

$$= -\eta \left(\sum_m w_{jm}^{\ell} \, g'(h_j^{\ell-1}) \frac{\partial E}{\partial h_m^{j}} \right) o_k^{\ell-2}$$

en notación matricial

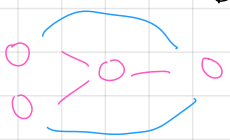
PRODUCTO ESCALAR (np.dot)

última capa $\underbrace{\Delta W_{jk}^{\ell}}_{\text{vector fila}} = -\eta \underbrace{\left(y_j^{\ell} - o_j^{\ell} \right) g'(h_j^{\ell})}_{\text{PRODUCTO PUNTO A PUNTO } (*)} \underbrace{o_k^{\ell-1}}_{\text{fila}}$

nota: vectores

$j \to \begin{pmatrix} & \overset{k}{\downarrow} & \\ & \bar{W} & \end{pmatrix} \qquad \begin{pmatrix} & j \times 1 & \\ & y, o & \end{pmatrix} \qquad \begin{pmatrix} & 1 \times k & \\ & o^T & \end{pmatrix}$

$$\Rightarrow \Delta \bar{W} = -\eta \underbrace{(\bar{y} - \bar{o}) g(h)}_{\bar{D}} (o^{\ell-1})^T$$



anteúltima capa

$$\Delta w_{jk}^{l-1} = -\eta \left(\sum_m w_{jm}^{l} \; g'(h_j^{l-1}) \; \frac{\partial E}{\partial h_m^{l}} \right) O_k^{l-2}$$

$$\sum_m w_{jm}^{l} \frac{\partial E}{\partial h_m^{l}} = W_j^T \bar{D}$$ → error output

es un prod matriz × vector con W^T

$$\Rightarrow \Delta w_{jh}^{l-1} = -\eta \; W_j^{l+} \bar{D} \; g'(h_j^{l-1}) \; O_k^{l-2}$$

matricial

$$\Rightarrow \Delta w_{jh}^{l-1} = -\eta \left[\left(W_j^{l+} \bar{D} \right) \cdot g' \right] O_k^{l-2}$$

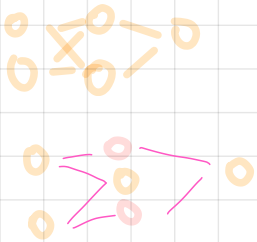
prod pto a pto

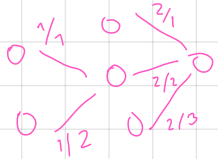
$$\Delta \bar{W}^{l-1} = -\eta \left[\left(W^{l+} \bar{D} \right) \cdot g'(h) \right] (O^{l-2})^T$$

matricial

$$h_m^m = \sum_s w_{ms} O_s^m = \sum_s w_{ms} g(h_s^m)$$

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





$$
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{matrix} 0 \\ 0 \end{matrix}
$$

$$
\begin{matrix} w_1 \\ 0 \end{matrix}
$$

$$
\begin{matrix} w_2 \\ 0 \end{matrix}
$$