Algoritmo de retropropagación del error

Vamos a definir el algoritmo on-line, esto es, cada vez que tengamos un patrón de entrenamiento, si el error cometido es mayor que un determinado umbral, activamos el algoritmo.
Sea por tanto la función de error a minimizar del patrón $\mathbf{x_1}$ de la forma

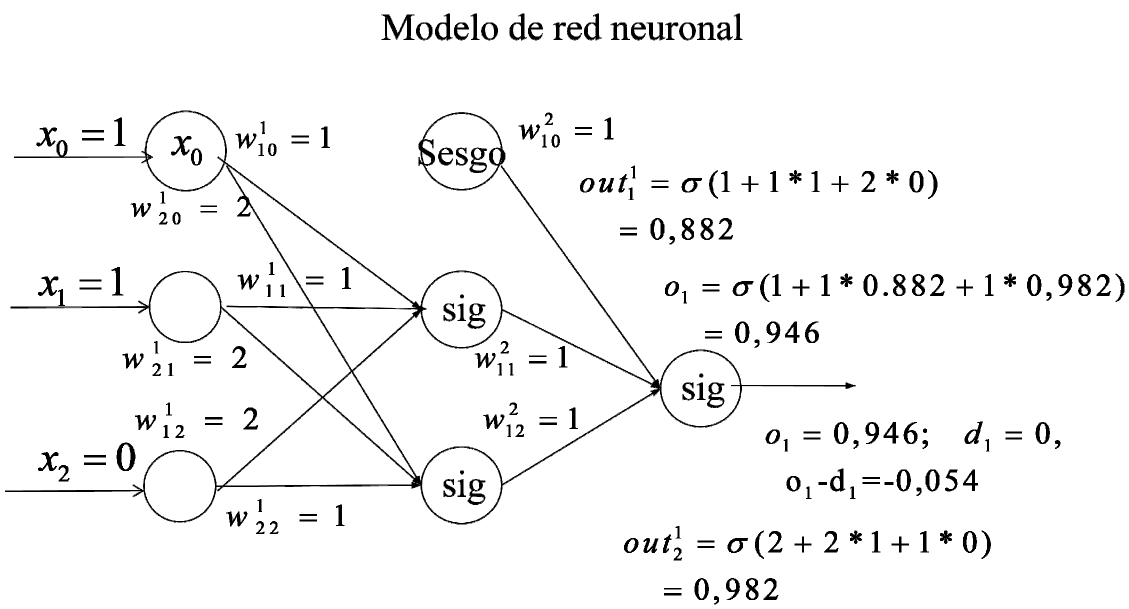
$$E = (d_1 - o_1)^2$$

donde d_1 es la salida deseada y o_1 la salida estimada por el modelo de red neuronal. Ahora derivamos la función de red con respecto a cada uno de los pesos de la red (ver Figura 1). Tenemos un modelo de red *feed-forward* con dos nodos en capa de entrada más el sesgo, dos nodos sigmoides en capa oculta mas el sesgo, y un nodo sigmoide en capa de salida.
La ecuación matemática del modelo es

$$f(\mathbf{x}_l) = \beta_0 + \sum_{i=1}^{2} \beta_i B_i(\mathbf{x}_l, \mathbf{w}_i) = w_{10}^2 + w_{11}^2 \frac{1}{1+e^{-(w_{10}^1 + w_{11}^1 x_1 + w_{12}^1 x_2)}} + w_{12}^2 \frac{1}{1+e^{-(w_{20}^1 + w_{21}^1 x_1 + w_{22}^1 x_2)}}$$

$$\mathbf{x}_l \in \mathbb{R}^k$$

Figura 1.-



Para todos los pesos, de capa oculta a capa de salida, se cumple que

$$\frac{\partial E}{\partial w_{ij}^2} = \frac{\partial E}{\partial o_1} \frac{\partial o_1}{\partial w_{ij}^2} = -2(d_1 - o_1) \frac{\partial o_1}{\partial w_{ij}^2}$$

siendo

$$o_1 = \sigma(\mathrm{w}_{10}^2 + w_{11}^2 * out_1^1 + w_{12}^2 * out_2^1) = \sigma(net_1^2)$$

donde

$$out_1^1 = \sigma(\mathrm{w}_{10}^1 * \mathrm{x}_0 + w_{11}^1 * \mathrm{x}_1 + w_{12}^1 * \mathrm{x}_2)$$

$$out_2^1 = \sigma(\mathrm{w}_{20}^1 * \mathrm{x}_0 + w_{21}^1 * \mathrm{x}_1 + w_{22}^1 * \mathrm{x}_2)$$

Por otra parte la derivada de la sigmoide es

$$\frac{\partial \sigma(x)}{\partial x} = \sigma(x)(1-\sigma(x))$$, luego

$$\frac{\partial o_1}{\partial w_{ij}^2} = \frac{\partial o_1}{\partial (net_1^2)} \frac{\partial (net_1^2)}{\partial w_{ij}^2} = o_1(1-o_1)\frac{\partial (net_1^2)}{\partial w_{ij}^2}$$

y como $net_1^2 = \mathrm{w}_{10}^2 + w_{11}^2 * out_1^1 + w_{12}^2 * out_2^1$

$$\frac{\partial (net_1^2)}{\partial w_{ij}^2} = \begin{cases} \dfrac{\partial (net_1^2)}{\partial w_{10}^2} = 1 \\ \dfrac{\partial (net_1^2)}{\partial w_{11}^2} = out_1^1 = \sigma(\mathrm{w}_{10}^1 * \mathrm{x}_0 + w_{11}^1 * \mathrm{x}_1 + w_{12}^1 * \mathrm{x}_2) \\ \dfrac{\partial (net_1^2)}{\partial w_{12}^2} = out_2^1 = \sigma(\mathrm{w}_{20}^1 * \mathrm{x}_0 + w_{21}^1 * \mathrm{x}_1 + w_{22}^1 * \mathrm{x}_2) \end{cases}$$

Así, las derivadas de los pesos de capa de salida a capa oculta son

$$\frac{\partial E}{\partial w_{10}^2} = \delta_1^2 * 1$$

$$\frac{\partial E}{\partial w_{11}^2} = \delta_1^2 out_1^1 = \delta_1^2 \sigma(\mathrm{w}_{10}^1 * \mathrm{x}_0 + w_{11}^1 * \mathrm{x}_1 + w_{12}^1 * \mathrm{x}_2)$$

$$\frac{\partial E}{\partial w_{12}^2} = \delta_1^2 out_2^1 = \delta_1^2 \sigma(\mathrm{w}_{20}^1 * \mathrm{x}_0 + w_{21}^1 * \mathrm{x}_1 + w_{22}^1 * \mathrm{x}_2)$$

Donde $\delta_1^2 = -2(d_1 - o_1)\mathrm{o}_1(1-\mathrm{o}_1)$

Para calcular las derivadas de los pesos de capa oculta a capa de entrada, $w_{10}^1, w_{11}^1, \mathrm{w}_{21}^1$, como estos pesos forman parte de out_1^1 y out_2^1, cuando llegamos a las derivadas $\frac{\partial(net_1^2)}{\partial w_{ij}^2}$ en

$$\frac{\partial E}{\partial w_{ij}^2} = \frac{\partial E}{\partial o_1}\frac{\partial o_1}{\partial w_{ij}^2} = -2(d_1 - o_1)\frac{\partial o_1}{\partial w_{ij}^2} = -2(d_1 - o_1)o_1(1 - o_1)\frac{\partial(net_1^2)}{\partial w_{ij}^2}$$

tenemos $net_1^2 = \mathrm{w}_{10}^2 + w_{11}^2 * out_1^1 + w_{12}^2 * out_2^1$

siendo

$$out_1^1 = \sigma(\mathrm{w}_{10}^1 * \mathrm{x}_0 + w_{11}^1 * \mathrm{x}_1 + w_{12}^1 * \mathrm{x}_2) = \sigma(\mathrm{net}_1^1)$$

y

$$out_2^1 = \sigma(\mathrm{w}_{20}^1 * \mathrm{x}_0 + w_{21}^1 * \mathrm{x}_1 + w_{22}^1 * \mathrm{x}_2) = \sigma(\mathrm{net}_2^1)$$

luego

$$\frac{\partial E}{\partial w_{10}^1} = \frac{\partial E}{\partial o_1}\frac{\partial o_1}{\partial w_{10}^1} = -2(d_1 - o_1)\frac{\partial o_1}{\partial w_{10}^1} = -2(d_1 - o_1)o_1(1 - o_1)\frac{\partial(net_1^2)}{\partial w_{10}^1} =$$

$$= \delta_1^2 w_{11}^2 \frac{\partial(out_1^1)}{\partial w_{10}^1} = \delta_1^2 w_{11}^2 out_1^1(1 - out_1^1) * \mathrm{x}_0$$

$$\frac{\partial E}{\partial w_{11}^1} = \frac{\partial E}{\partial o_1}\frac{\partial o_1}{\partial w_{11}^1} = \delta_1^2 w_{11}^2 \frac{\partial(out_1^1)}{\partial w_{11}^1} = \delta_1^2 w_{11}^2 out_1^1(1 - out_1^1) * \mathrm{x}_1$$

$$\frac{\partial E}{\partial w_{12}^1} = \frac{\partial E}{\partial o_1}\frac{\partial o_1}{\partial w_{11}^1} = \delta_1^2 w_{11}^2 \frac{\partial(out_1^1)}{\partial w_{11}^1} = \delta_1^2 w_{11}^2 out_1^1(1 - out_1^1) * \mathrm{x}_2$$

$$\frac{\partial E}{\partial w_{20}^1} = \frac{\partial E}{\partial o_1}\frac{\partial o_1}{\partial w_{20}^1} = -2(d_1 - o_1)\frac{\partial o_1}{\partial w_{20}^1} = -2(d_1 - o_1)o_1(1 - o_1)\frac{\partial(net_1^2)}{\partial w_{20}^1} =$$

$$= \delta_1^2 w_{12}^2 \frac{\partial(out_2^1)}{\partial w_{20}^1} = \delta_1^2 w_{12}^2 out_2^1(1 - out_2^1) * \mathrm{x}_0$$

$$\frac{\partial E}{\partial w_{21}^1} = \frac{\partial E}{\partial o_1}\frac{\partial o_1}{\partial w_{21}^1} = \delta_1^2 w_{12}^2 \frac{\partial(out_2^1)}{\partial w_{21}^1} = \delta_1^2 w_{12}^2 out_2^1(1 - out_2^1) * \mathrm{x}_1$$

$$\frac{\partial E}{\partial w_{22}^1} = \frac{\partial E}{\partial o_1}\frac{\partial o_1}{\partial w_{22}^1} = \delta_1^2 w_{12}^2 \frac{\partial (out_2^1)}{\partial w_{22}^1} = \delta_1^2 w_{12}^2 out_2^1 (1 - out_2^1) * \mathrm{x}_2$$

Si definimos

$$\delta_1^1 = \delta_1^2 w_{11}^2 out_1^1 (1 - out_1^1) \ \mathbf{y} \ \delta_2^1 = \delta_1^2 w_{12}^2 out_2^1 (1 - out_2^1)$$

Se pueden resumir las derivadas de la función de error con respecto a los pesos del modelo de red, para los pesos de capa de salida a capa oculta, en la forma

$$\frac{\partial E}{\partial w_{1j}^2} = \begin{cases} \delta_1^2 * 1 & \text{si j=0} \\ \delta_1^2 * \text{out}_j^1 & \text{si j} \neq 0 \end{cases}$$

y para los pesos de capa oculta a capa de entrada

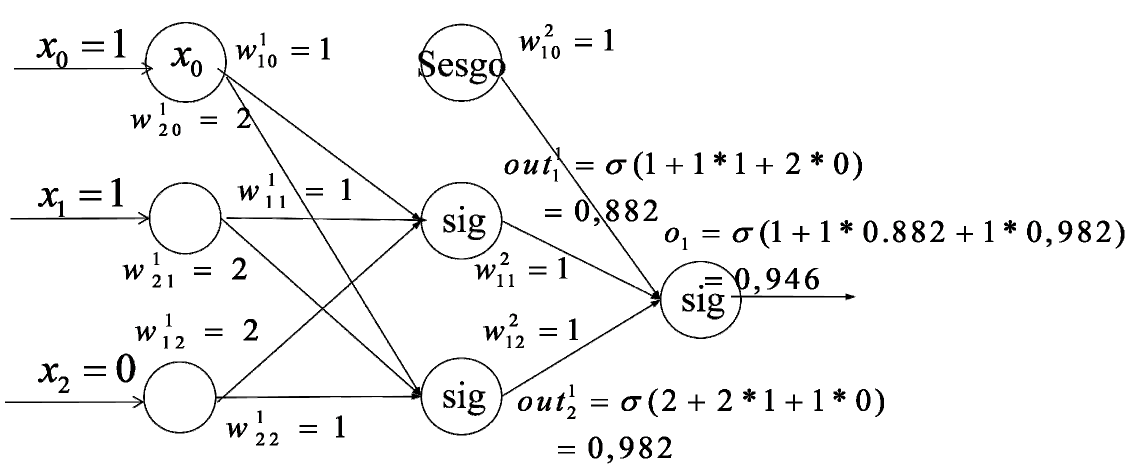
$$\frac{\partial E}{\partial w_{ij}^1} = \begin{cases} \delta_i^1 * 1 & \text{si j=0} \\ \delta_i^1 * x_j & \text{si j} \neq 0 \end{cases}$$

Ejemplo de aplicación del algoritmo de retropropagación del error

Tenemos un patrón de características de entrada x_1=1, x_2=0 y cuya etiqueta de salida es la clase 0, d_1=0. En la Figura 2 podemos ver la activación hacia adelante, dados unos pesos iniciales de la red.

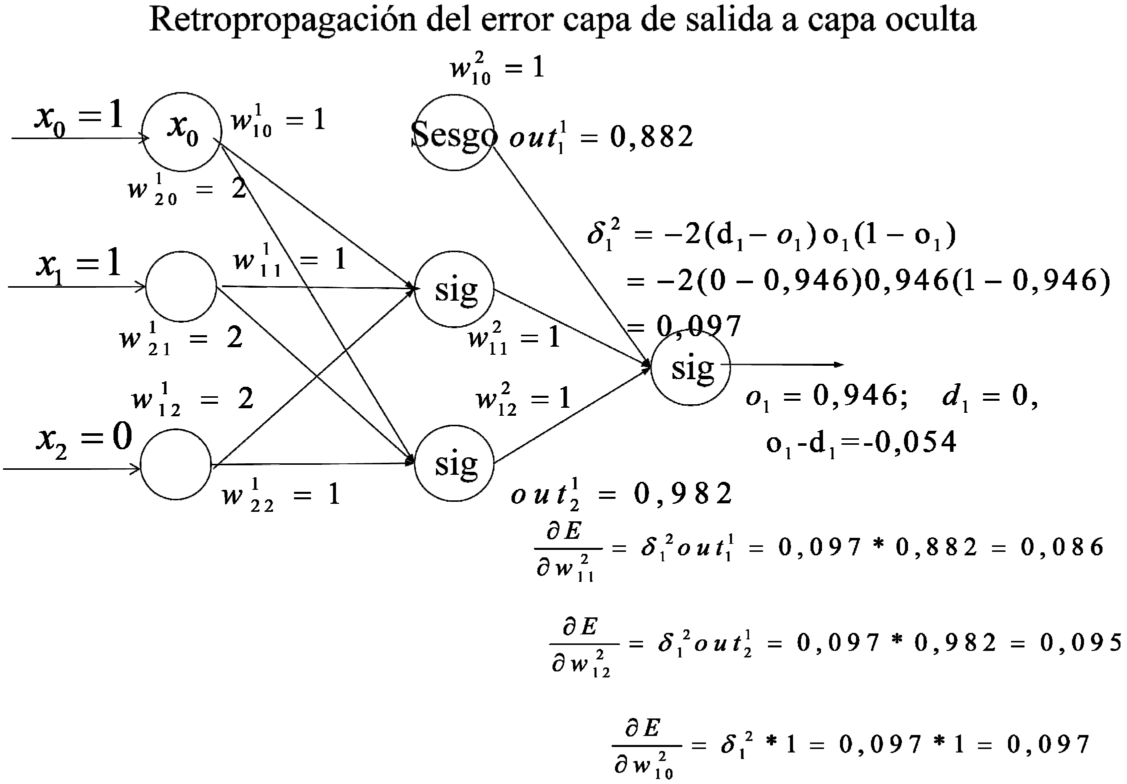
Figura 2.-

Activación hacia adelante



En la Figura 3 podemos ver las derivadas de la función de error con respecto a los pesos de capa de salida a capa oculta.

Figura 3.-



Por último, en la Figura 4 podemos ver las derivadas de la función de error con respecto a los pesos de capa oculta a capa de entrada

Figura 4.-