Powiedzmy, że przy wejściu $\begin{bmatrix}1\\0\end{bmatrix}$ oczekujemy od sieci odpowiedzi $\begin{bmatrix}1\\0\end{bmatrix}$.

Wagi początkowe (przykładowe):

$$W = \begin{bmatrix} 0.1 & -0.2 & 0.3 \\ -0.4 & 0.5 & -0.6 \end{bmatrix}, \qquad V = \begin{bmatrix} 0.15 & -0.25 & 0.35 \\ -0.45 & 0.55 & -0.65 \end{bmatrix}$$

Wejście (ostatnia współrzędna jest zawsze równa -1 i służy do kodowania progu aktywacji):

$$X^{(1)} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

Warstwa pierwsza (W) daje

$$net_1 = W \cdot X^{(1)} = \begin{bmatrix} -0.2 \\ 0.2 \end{bmatrix}$$

po nałożeniu na poszczególne elementy funkcji $f(x) = (1 + e^{-x})^{-1}$

$$Y^{(1)} = \begin{bmatrix} f(-0.2) \\ f(0.2) \end{bmatrix} = \begin{bmatrix} 0.450166 \\ 0.549834 \end{bmatrix},$$

a po dopisaniu -1 na trzeciej współrzędnej otrzymujemy wejście drugiej warstwy:

$$X^{(2)} = \begin{bmatrix} 0.450166 \\ 0.549834 \\ -1 \end{bmatrix}$$

Warstwa druga (V) daje

$$net_2 = V \cdot X^{(2)} = \begin{bmatrix} -0.4199336 \\ 0.749834 \end{bmatrix}$$

po nałożeniu na poszczególne elementy funkcji $f(x) = (1 + e^{-x})^{-1}$ dostajemy wyjście sieci:

$$Y^{(2)} = \begin{bmatrix} f(-0.4199336) \\ f(0.749834) \end{bmatrix} = \begin{bmatrix} 0.39653264 \\ 0.67914253 \end{bmatrix}.$$

Oczekiwaliśmy odpowiedzi $\begin{bmatrix}1\\0\end{bmatrix}$, dlatego błąd wynosi $b^{(2)} = \begin{bmatrix}1\\0\end{bmatrix} - Y^{(2)} = \begin{bmatrix} 0.60346736 \\ -0.67914253 \end{bmatrix}$ Stąd dostajemy *sygnał delta* mnożąc poszczególne elementy przez wartość f' w punktach net_2; ponieważ $f'(x) = f(x)(1 - f(x))$, więc

$$\delta^{(2)} = \begin{bmatrix} 0.60346736 \cdot f'(-0.4199336) \\ -0.67914253 \cdot f'(0.749834) \end{bmatrix} = \begin{bmatrix} 0.14440642 \\ -0.147990559 \end{bmatrix}$$

Zmieniamy wagi V, biorąc współczynnik uczenia $c = 0.1$,

$$\tilde{V} = V + c\,\delta^{(2)} \cdot (X^{(2)})^T = V + c \begin{bmatrix} 0.06500686 & 0.07939956 & -0.1444 \\ -0.06662 & -0.08137 & 0.14799 \end{bmatrix} = \begin{bmatrix} 0.156500686 & -0.24206 & 0.33556 \\ -0.456662 & 0.541863 & -0.6352 \end{bmatrix}$$

Obliczamy błąd dla pierwszej warstwy następująco

$$b^{(1)} = V^T \cdot \delta^{(2)} = \begin{bmatrix} 0.08825672 \\ -0.11749641 \\ 0.14673611 \end{bmatrix}.$$

Ostatnia współrzędna (0.14673611) jest zbędna (odpowiada stałemu wejściu -1, które koduje poziom aktywacji) – pomijamy ją, zaś pozostałe domnażamy przez wartości f' w punktach net_1, skąd otrzymujemy sygnał delta dla pierwszej warstwy,

$$\delta^{(1)} = \begin{bmatrix} 0.08825672 \cdot f'(-0.2) \\ -0.11749641 \cdot f'(0.2) \end{bmatrix} = \begin{bmatrix} 0.021845 \\ -0.0290823 \end{bmatrix}.$$

Zmieniamy wagi W, biorąc znowu współczynnik uczenia $c = 0.1$,

$$\tilde{W} = W + c\,\delta^{(1)} \cdot (X^{(1)})^T = \begin{bmatrix} 0.1021845 & -0.2 & 0.2978155 \\ -0.40290823 & 0.5 & -0.59709177 \end{bmatrix}.$$

Dostajemy sieć ze zmodyfikowanymi wagami $\tilde{W}$ i $\tilde{V}$, powtarzamy...

Uwagi:

(1) Powyższe równości są przybliżone, mogą też występować błędy z różnych zaokrągleń (w obliczeniach zwykle używane były dokładniejsze wartości niż wypisane).

(2) Przy wyborze funkcji $f(x) = (1 + e^{-x})^{-1}$ zachodzi, jak łatwo sprawdzić, wzór $f'(x) = f(x)(1 - f(x))$. Jeśli wzięlibyśmy $\tilde{f}(x) = f(\lambda x)$, to $\tilde{f}'(x) = \lambda \tilde{f}(x)(1 - \tilde{f}(x))$. Mimo to można czynnik λ pominąć we wzorach, włączając go do stałej uczenia c.

przygotował B.D.