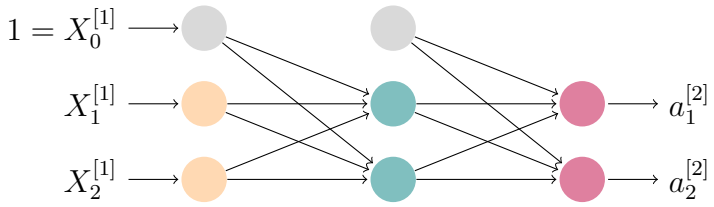
Rozważamy sieć neuronową z dwoma warstwami po dwa neurony. Poniżej szare kółka oznaczają sztucznie dodane wejścia stale równe 1.



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

Wagi początkowe (przykładowe):

$$W^{[1]} = \begin{bmatrix} 0.1 & -0.2 & 0.3 \\ -0.4 & 0.5 & -0.6 \end{bmatrix}, \qquad W^{[2]} = \begin{bmatrix} 0.15 & -0.25 & 0.35 \\ -0.45 & 0.55 & -0.65 \end{bmatrix}$$

Wejście (pierwsza współrzędna jest zawsze równa 1 i służy do kodowania stałego składnika):

$$X^{[1]} = \begin{bmatrix}1\\1\\0\end{bmatrix}.$$

Propagacja w przód. Warstwa pierwsza ($W^{[1]}$) daje

$$net^{[1]} = W^{[1]} \cdot X^{[1]} = \begin{bmatrix}-0.1\\0.1\end{bmatrix},$$

po nałożeniu na poszczególne elementy funkcji $\varphi(x) = (1+e^{-x})^{-1}$ dostajemy wyjście pierwszej wartswy,

$$a^{[1]} = \begin{bmatrix}\varphi(-0.1)\\ \varphi(0.1)\end{bmatrix} = \begin{bmatrix}0.47502081\\0.52497919\end{bmatrix},$$

a po dopisaniu 1 na początku otrzymujemy wejście drugiej warstwy:

$$X^{[2]} = \begin{bmatrix}1\\0.47502081\\0.52497919\end{bmatrix}$$

Warstwa druga ($W^{[2]}$) daje

$$net^{[2]} = W^{[2]} \cdot X^{[2]} = \begin{bmatrix}0.21498751\\-0.52997502\end{bmatrix}$$

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

$$a^{[2]} = \begin{bmatrix}\varphi(0.21498751)\\ \varphi(-0.52997502)\end{bmatrix} = \begin{bmatrix}0.55354082\\0.37052271\end{bmatrix}.$$

Propagacja wstecz. Oczekiwaliśmy odpowiedzi $y = \begin{bmatrix}1\\0\end{bmatrix}$. Dla funkcji kosztu $L(y, a^{[2]}) = \frac{1}{2}\|y - a^{[2]}\|_2^2$ mamy

$$\frac{\partial L}{\partial a^{[2]}} = a^{[2]} - y = a^{[2]} - \begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}-0.44645918\\0.37052271\end{bmatrix}. \tag{0.1}$$

Stąd dostajemy *sygnał delta* mnożąc poszczególne elementy przez wartość φ' w punktach $net^{[2]}$ (tutaj φ' obliczamy korzystając z konkretnej postaci funkcji φ),

$$\delta^{[2]} = \begin{bmatrix}-0.44645918 \cdot \varphi'(0.21498751)\\ 0.37052271 \cdot \varphi'(-0.52997502)\end{bmatrix} = \begin{bmatrix}-0.110334967\\0.086419099\end{bmatrix}.$$

Stąd

$$\frac{\partial L}{W^{[2]}} = \delta^{[2]} \cdot (X^{[2]})^T = \begin{bmatrix} -0.110335 & -0.0524114 & -0.0579236 \\ 0.0864191 & 0.0410509 & 0.0453682 \end{bmatrix}.$$

Zmieniamy wagi $W^{[2]}$, biorąc współczynnik uczenia $c = 0.1$,

$$\widetilde{W^{[2]}} = W^{[2]} - c\,\frac{\partial L}{W^{[2]}} = W^{[2]} - c \begin{bmatrix} -0.110335 & -0.0524114 & -0.0579236 \\ 0.0864191 & 0.0410509 & 0.0453682 \end{bmatrix}$$
$$= \begin{bmatrix} 0.1610335 & -0.24475886 & 0.35579236 \\ -0.45864191 & 0.54589491 & -0.65453682 \end{bmatrix}$$

Obliczamy

$$\frac{\partial L}{X^{[2]}} = (W^{[2]})^T \cdot \delta^{[2]} = \begin{bmatrix} -0.05543884 \\ 0.07511425 \\ -0.09478965 \end{bmatrix}.$$

Pierwsza współrzędna (-0.05543884) jest zbędna (odpowiada stałemu wejściu 1, które koduje stały składnik) – pomijamy ją i otrzymujemy

$$\frac{\partial L}{a^{[1]}} = \begin{bmatrix} 0.07511425 \\ -0.09478965 \end{bmatrix}. \tag{0.2}$$

Zauważmy, że otrzymaliśmy analogiczną pochodną jak w (0.1), tylko dla warstwy o jeden głębszej. Dalej postępujemy analogicznie. Konkretnie, domnażamy kolejne elementy $\frac{\partial L}{a^{[1]}}$ przez wartości φ' w punktach $net^{[1]}$, skąd otrzymujemy sygnał delta dla pierwszej warstwy,

$$\delta^{[1]} = \begin{bmatrix} 0.07511425 \cdot \varphi'(-0.1) \\ -0.09478965 \cdot \varphi'(0.1) \end{bmatrix} = \begin{bmatrix} 0.0187316942 \\ -0.023638267 \end{bmatrix}.$$

Stąd

$$\frac{\partial L}{W^{[1]}} = \delta^{[1]} \cdot (X^{[1]})^T = \begin{bmatrix} 0.0187317 & 0.0187317 & 0 \\ -0.0236383 & -0.0236383 & 0 \end{bmatrix}.$$

Zmieniamy wagi $W^{[1]}$, biorąc znowu współczynnik uczenia $c = 0.1$,

$$\widetilde{W^{[1]}} = W^{[1]} - c\frac{\partial L}{W^{[1]}} = \begin{bmatrix} 0.09812683 & -0.20187317 & 0.3 \\ -0.39763617 & 0.50236383 & -0.6 \end{bmatrix}.$$

Dostajemy sieć ze zmodyfikowanymi wagami $\widetilde{W^{[1]}}$ i $\widetilde{W^{[2]}}$, powtarzamy...

Uwagi:

(1) Gdybyśmy mieli głębszą sieć, to moglibyśmy dalej liczyć

$$\frac{\partial L}{X^{[1]}} = (W^{[1]})^T \cdot \delta^{[1]},$$

pominąć pierwszy element, aby obliczyć $\frac{\partial L}{a^{[0]}}$, i bylibyśmy znowu w sytuacji analogicznej do (0.1) i (0.2). Itd.

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

(3) Przy wyborze funkcji $\varphi(x) = (1 + e^{-x})^{-1}$ zachodzi, jak łatwo sprawdzić, wzór $\varphi'(x) = \varphi(x)(1 - \varphi(x))$. To pozwala wykonywać obliczenia nieco efektywniej, ponieważ $\varphi(x)$ obliczamy w przejściu w przód.

(4) Jeśli użyjemy innej funkcji aktywacji φ (ale nie *softmax*) i takiej samej funkcji kosztu, to w powyższych rachunkach zmienią się oczywiście wartości $\varphi(\ldots)$ oraz $\varphi'(\ldots)$, ale nie będzie innych zmian.

(5) Często w ostatniej warstwie używa się do aktywacji funkcji *softmax*:

$$\psi(\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}) = \begin{bmatrix} \frac{e^{x_1}}{\sum_{k=1}^n e^{x_k}} \\ \frac{e^{x_2}}{\sum_{k=1}^n e^{x_k}} \\ \dots \\ \frac{e^{x_n}}{\sum_{k=1}^n e^{x_k}} \end{bmatrix}.$$

Ponieważ zależy ona od całego wektora $net^{[L-1]}$, tzn. nie jest funkcją na $\mathbf{R}$, którą nakładamy tylko na kolejne współrzędne, więc sposób liczenia wyjścia sieci i wstecznej propagacji w ostatniej warstwie są nieco inne. W przykładzie powyżej, mielibyśmy

$$a^{[2]} = \psi(\begin{bmatrix} 0.21498751 \\ -0.52997502 \end{bmatrix}) = \begin{bmatrix} \frac{1.2398464}{1.828466} \\ \frac{0.58861967}{1.828466} \end{bmatrix} = \begin{bmatrix} 0.67808 \\ 0.32192 \end{bmatrix}$$

Jeśli dodatkowo użyjemy *entropii krzyżowej* jako funkcji kosztu (co zwykle się robi, jeśli używa się funkcji *softmax*) – w przykładzie jak wyżej byłaby to funkcja

$$L_e(y, a^{[2]}) = \sum_{k=1}^{2} -y_k \log(a_k^{[2]}),$$

to wówczas powyższe wzory na uaktualnianie wag pozostaną w mocy, jeśli zmienimy definicję $\delta^{[2]}$ w następujący sposób (ale tylko jej, nie delt dla głębszych warstw):

$$\delta^{[2]} := \left(\sum_{j=1}^{2} y_j\right) a^{[2]} - y.$$

W porównaniu do poprzednej definicji, nie przemnażamy przez pochodne φ'. Uzasadnienie tego faktu wymaga powtórzenia obliczeń $\frac{\partial L}{\partial W^{[2]}}$ oraz $\frac{\partial L}{\partial X^{[2]}}$; rachunki są dość żmudne, ponieważ teraz każde $a_j^{[2]}$ zależy od wszystkich współczynników macierzy $W^{[2]}$.

W naszym przykładzie mielibyśmy

$$\delta^{[2]} = (1+0)a^{[2]} - y = \begin{bmatrix} -0.32192 \\ 0.32192 \end{bmatrix}.$$

Dalej rachunki przebiegają tak samo jak poprzednio, jednak ponieważ mamy inną wartość $\delta^{[2]}$, więc będziemy otrzymywali inne liczby. Konkretnie, otrzymujemy

$$\frac{\partial L}{W^{[2]}} = \delta^{[2]} \cdot (X^{[2]})^T = \begin{bmatrix} -0.32192 & -0.1529187 & -0.1690013 \\ 0.32192 & 0.1529187 & 0.1690013 \end{bmatrix}.$$

Zmieniamy wagi $W^{[2]}$, biorąc współczynnik uczenia $c = 0.1$,

$$\widetilde{W^{[2]}} = W^{[2]} - c\frac{\partial L}{W^{[2]}} = W^{[2]} - c\begin{bmatrix} -0.32192 & -0.1529187 & -0.1690013 \\ 0.32192 & 0.1529187 & 0.1690013 \end{bmatrix}$$
$$= \begin{bmatrix} 0.182192 & -0.23470813 & 0.36690013 \\ -0.482192 & 0.53470813 & -0.66690013 \end{bmatrix}$$

Obliczamy

$$\frac{\partial L}{X^{[2]}} = (W^{[2]})^T \cdot \delta^{[2]} = \begin{bmatrix} -0.193152 \\ 0.257536 \\ -0.32192 \end{bmatrix}.$$

Pierwsza współrzędna (-0.193152) jest zbędna (odpowiada stałemu wejściu 1, które koduje stały składnik) – pomijamy ją i otrzymujemy

$$\frac{\partial L}{a^{[1]}} = \begin{bmatrix} 0.257536 \\ -0.32192 \end{bmatrix}.$$

Zatem

$$\delta^{[1]} = \begin{bmatrix} 0.257536 \cdot \varphi'(-0.1) \\ -0.32192 \cdot \varphi'(0.1) \end{bmatrix} = \begin{bmatrix} 0.06422331 \\ -0.08027913 \end{bmatrix}.$$

Stąd

$$\frac{\partial L}{W^{[1]}} = \delta^{[1]} \cdot (X^{[1]})^T = \begin{bmatrix} 0.06422331 & 0.06422331 & 0 \\ -0.08027913 & -0.08027913 & 0 \end{bmatrix}.$$

Zmieniamy wagi $W^{[1]}$, biorąc znowu współczynnik uczenia $c = 0.1$,

$$\widetilde{W^{[1]}} = W^{[1]} - c\frac{\partial L}{W^{[1]}} = \begin{bmatrix} 0.09357767 & -0.20642233 & 0.3 \\ -0.39197209 & 0.50802791 & -0.6 \end{bmatrix}.$$

przygotował Bartek Dyda