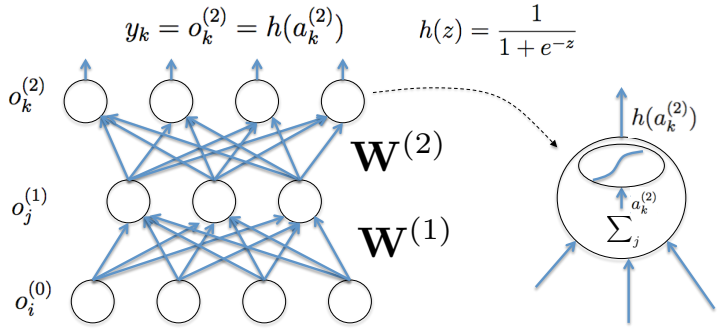
Backpropagation Algorithm

Man-Wai MAK

July 2015

1 Loss Function: Mean Square Error



Output Layer:

$$y_k = o_k^{(2)} = h(a_k^{(2)}) \qquad h(z) = \frac{1}{1+e^{-z}}$$
$$a_k^{(2)} = \sum_j w_{kj}^{(2)} o_j^{(1)}$$

Hidden Layer:

$$o_j^{(1)} = h(a_j^{(1)})$$
$$a_j^{(1)} = \sum_i w_{ji}^{(1)} o_i^{(0)} = \sum_i w_{ji}^{(1)} x_i$$

where x_i is the i-th elements of the input vector $\mathbf{x}$.

Error for a training vector:

$$E = \frac{1}{2}\sum_k (y_k - t_k)^2 = \frac{1}{2}\sum_k (o_k^{(2)} - t_k)^2$$

$$\begin{aligned}\frac{\partial E}{\partial w_{kj}^{(2)}} &= \frac{\partial E}{\partial a_k^{(2)}}\frac{\partial a_k^{(2)}}{\partial w_{kj}^{(2)}} \\ &= \delta_k^{(2)} o_j^{(1)}\end{aligned}$$

where

$$\begin{aligned}\delta_k^{(2)} &= \frac{\partial E}{\partial a_k^{(2)}} = \frac{\partial E}{\partial o_k^{(2)}}\frac{\partial o_k^{(2)}}{\partial a_k^{(2)}} \\ &= (o_k^{(2)} - t_k)\frac{\partial h(a_k^{(2)})}{\partial a_k^{(2)}}\end{aligned}$$

$$= \quad (o_k^{(2)} - t_k) h'(a_k^{(2)})$$

$$\begin{aligned} \frac{\partial E}{\partial w_{ji}^{(1)}} &= \frac{\partial E}{\partial a_j^{(1)}} \frac{\partial a_j^{(1)}}{\partial w_{ji}^{(1)}} \\ &= \delta_j^{(1)} x_i \end{aligned}$$

where

$$\begin{aligned} \delta_j^{(1)} &= \frac{\partial E}{\partial a_j^{(1)}} = \sum_k \frac{\partial E}{\partial a_k^{(2)}} \frac{\partial a_k^{(2)}}{\partial a_j^{(1)}} \\ &= \sum_k \delta_k^{(2)} \frac{\partial a_k^{(2)}}{\partial o_j^{(1)}} \frac{\partial o_j^{(2)}}{\partial a_j^{(1)}} \\ &= \sum_k \delta_k^{(2)} w_{kj}^{(2)} h'(a_j^{(1)}) \\ &= h'(a_j^{(1)}) \sum_k \delta_k^{(2)} w_{kj}^{(2)} \end{aligned}$$

The above derivative is based on the fact that E depends on $a_k^{(2)}$ $\forall k$ and that each of the $a_k^{(2)}$ depends on $a_j^{(1)}$.

Weight update formula

$$\begin{aligned} w_{kj}^{(2)} &\leftarrow w_{kj}^{(2)} - \eta \frac{\partial E}{\partial w_{kj}^{(2)}} \\ \Rightarrow w_{kj}^{(2)} &\leftarrow w_{kj}^{(2)} - \eta \delta_k^{(2)} o_j^{(1)} \\ \Rightarrow w_{kj}^{(2)} &\leftarrow w_{kj}^{(2)} - \eta (y_k - t_k) h'(a_k^{(2)}) o_j^{(1)} \end{aligned}$$

where $h'(a_k^{(2)}) = y_k(1 - y_k)$ for non-linear sigmoid output.

$$\begin{aligned} w_{ji}^{(1)} &\leftarrow w_{ji}^{(1)} - \eta \frac{\partial E}{\partial w_{ji}^{(1)}} \\ \Rightarrow w_{ji}^{(1)} &\leftarrow w_{ji}^{(1)} - \eta \delta_k^{(1)} x_i^{(1)} \\ \Rightarrow w_{ji}^{(1)} &\leftarrow w_{ji}^{(1)} - \eta \left[h'(a_j^{(1)}) \sum_k \delta_k^{(2)} w_{kj}^{(2)} \right] x_i \end{aligned}$$

where $h'(a_j^{(1)}) = o_j(1 - o_j^{(1)})$.

2 Cross-Entropy Loss Function (2-class, one output)

Assume that the sigmoid function is used for a single-output BP network. Using Eq. 4.90 of Bishop's book,

$$E_{Total} = -\sum_n \{t_n \log y_n + (1 - t_n) \log(1 - y_n)\}$$

where n is the sample index.

The instantaneous error is

$$E_n = -t_n \log y_n - (1 - t_n) \log(1 - y_n)$$

To simplify notation, we drop the subscript n

$$E = -t \log y - (1 - t) \log(1 - y)$$
$$\frac{\partial E}{\partial w_{ij}^{(2)}} = \frac{\partial E}{\partial a_1^{(2)}} \frac{\partial a_1^{(2)}}{\partial w_{1j}^{(2)}} = \delta_1^{(2)} o_j^{(1)}$$

where

$$\begin{aligned}
\delta_1^{(2)} &= \frac{\partial E}{\partial a_1^{(2)}} = \frac{\partial E}{\partial o_1^{(2)}} \frac{\partial o_1^{(2)}}{\partial a_1^{(2)}} = \frac{\partial E}{\partial y} \frac{\partial y}{\partial a_1^{(2)}} \\
&= \left[-\frac{t}{y} + \frac{1 - t}{1 - y} \right] h'(a_1^{(2)}) \qquad\qquad h(z) = \frac{1}{1 + e^{-z}} \\
&= \frac{-t(1 - y) + (1 - t)y}{y(1 - y)} h(a_1^{(2)})(1 - h(a_1^{(2)})) \\
&= y - t
\end{aligned}$$

$$\Rightarrow \frac{\partial E}{\partial w_{ij}^{(2)}} = (y - t) o_j^{(1)}$$

Similarly,

$$\frac{\partial E}{\partial w_{ji}^{(1)}} = \frac{\partial E}{\partial a_j^{(1)}} \frac{\partial a_j^{(1)}}{\partial w_{ji}^{(1)}} = \delta_j^{(1)} x_i$$

where

$$\begin{aligned}
\delta_j^{(1)} &= \frac{\partial E}{\partial a_j^{(1)}} = \frac{\partial E}{\partial a_1^{(2)}} \frac{\partial a_1^{(2)}}{\partial a_j^{(1)}} \\
&= \delta_1^{(2)} \frac{\partial a_1^{(2)}}{\partial o_j^{(1)}} \frac{\partial o_j^{(1)}}{\partial a_j^{(1)}} \\
&= (y - t) w_{kj}^{(2)} o_j^{(1)} (1 - o_j^{(1)})
\end{aligned}$$

Update Formula

$$w_{1j}^{(2)} \leftarrow w_{1j}^{(2)} - \eta (y - t) o_j^{(1)}$$
$$w_{ji}^{(1)} \leftarrow w_{ji}^{(1)} - \eta \left[(y - t) w_{1j}^{(2)} o_j^{(1)} (1 - o_j^{(1)}) \right] x_i$$

3 Cross-Entropy Loss Function (Multi-class)

Assume that softmax function is used for the output nodes:

$$y_k = o_k^{(2)} = \frac{\exp(a_k^{(2)})}{\sum_j \exp(a_j^{(2)})} \tag{1}$$

Because y_k depends on $a_j^{(2)}$ where $j = 1, \ldots, K$, we need to compute the derivative of y_k w.r.t. $a_j^{(2)}$, i.e.,

$$\begin{aligned}\frac{\partial y_k}{\partial a_j^{(2)}} &= \frac{I_{kj}\exp(a_k^{(2)})\sum_j \exp(a_j^{(2)}) - \exp(a_k^{(2)})\exp(a_j^{(2)})}{(\sum_j \exp(a_j^{(2)}))^2} \\ &= y_k I_{kj} - y_k y_j\end{aligned}$$

where

$$I_{kj} = \begin{cases} 1 & k = j; \\ 0 & \text{otherwise.} \end{cases}$$

Using Eq. 4.108 of Bishop's book, the instantaneous cross-entropy is

$$E = -\sum_{k=1}^{K} t_k \log y_k. \tag{2}$$

Eq. 2 suggests that E is a function of $y_k \; \forall \; k = 1, \ldots, K$, and according to Eq. 1, y_k depends on $a_r^{(2)} \; \forall \; r = 1, \ldots, K$. Therefore, we need to use the chain rule to compute $\frac{\partial E}{\partial w_{kj}^{(2)}}$:

$$\frac{\partial E}{\partial w_{kj}^{(2)}} = \sum_{r=1}^{K} \frac{\partial E}{\partial a_r^{(2)}} \frac{\partial a_r^{(2)}}{\partial w_{kj}^{(2)}}$$

where

$$\begin{aligned}\frac{\partial E}{\partial a_r^{(2)}} &= \sum_{k=1}^{K} \frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial a_r^{(2)}} \\ &= \frac{\partial E}{\partial y_r}\frac{\partial y_r}{\partial a_r^{(2)}} + \sum_{k\neq r} \frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial a_r^{(2)}} \\ &= -\frac{t_r}{y_r} y_r(1-y_r) - \sum_{k\neq r} \frac{t_k}{y_k} y_k(I_{kr} - 1)y_r \qquad \text{using Eq. 2} \\ &= -t_r(1-y_r) - \sum_{k\neq r} t_k(0-1)y_r \\ &= -t_r(1-y_r) + y_r \sum_{k\neq r} t_k \\ &= -t_r + y_r \sum_{k\neq r}^{K} t_k + t_r y_r \\ &= -t_r + y_r \sum_{k=1}^{K} t_k \\ &= y_r - t_r\end{aligned}$$

Therefore, we have

$$\frac{\partial E}{\partial w_{kj}^{(2)}} = \sum_{r} \frac{\partial E}{\partial a_r^{(2)}} \frac{\partial a_r^{(2)}}{\partial w_{kj}^{(2)}}$$

$$= \quad (y_k - t_k) o_j^{(1)} \qquad\qquad \because \frac{\partial a_r^{(2)}}{\partial w_{kj}^{(2)}} = 0 \ \ \forall r \neq k.$$

Similarly,

$$\begin{aligned}
\frac{\partial E}{\partial a_i^{(1)}} &= \sum_{k=1}^{K} \frac{\partial E}{\partial a_k^{(2)}} \frac{\partial a_k^{(2)}}{\partial o_i^{(1)}} \frac{\partial o_i^{(1)}}{\partial a_i^{(1)}} \\
&= \sum_{k=1}^{K} (y_k - t_k) w_{kj}^{(2)} o_i^{(1)} (1 - o_i^{(1)})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial E}{\partial w_{ji}^{(1)}} &= \sum_{s} \frac{\partial E}{\partial a_s^{(1)}} \frac{\partial a_s^{(1)}}{\partial w_{ji}^{(1)}} \\
&= \sum_{s} \sum_{k} (y_k - t_k) w_{ks}^{(2)} o_s^{(1)} (1 - o_s^{(1)}) \frac{\partial a_s^{(1)}}{\partial w_{ji}^{(1)}} \\
&= \sum_{k} (y_k - t_k) w_{kj}^{(2)} o_j^{(1)} (1 - o_j^{(1)}) x_i
\end{aligned}$$

Update Formula

$$\begin{aligned}
w_{kj}^{(2)} &\leftarrow w_{kj}^{(2)} - \eta (y_k - t_k) o_j^{(1)} \\
w_{ji}^{(1)} &\leftarrow w_{ji}^{(1)} - \eta \sum_{k=1}^{K} (y_k - t_k) w_{kj}^{(2)} o_j^{(1)} (1 - o_j^{(1)}) x_i
\end{aligned}$$