Homework

SA17011027 侯旻

Question:

> Given overall network function:
>
> $$y_k(x,m) = f(\sum_{j=0}^{M} w_{kj}^{(2)} h(\sum_{i=0}^{D} w_{ji}^{(1)} x_i))$$
>
> and cross-entropy error function for the multiclass classification problem:
>
> $$E(w) = -\sum_{k=1}^{K} t_k \ln y_k$$
>
> $t_k = (0,1)^K$ and $\sum_{k=1}^{K} t_k = 1$, $f(z_k) = \frac{\exp(z_k)}{\sum_{l=1}^{K} \exp(z_l)}$ is softmax function, $h(a) = \tanh(a) = \frac{\exp(a)-\exp(-a)}{\exp(a)+exp(-a)}$.
> Compute the derivatives of error function w.r.t the first-layer and second-layer weights with back propagation.
>
> $$\frac{\partial E}{\partial w_{ji}^{(1)}} and \frac{\partial E}{\partial w_{kj}^{(2)}}$$

Answer :

- Derivative of Softmax

$$\frac{\partial f(z_j)}{\partial z_k} = \frac{\partial \frac{\exp(z_j)}{\sum_{l=1}^{K} \exp{(z_l)}}}{\partial z_k}$$

If $k = j$,

$$\begin{aligned}
\frac{\partial f(z_j)}{\partial z_k} &= \frac{\partial \frac{\exp(z_k)}{\sum_{l=1}^{K} \exp{(z_l)}}}{\partial z_k} \\
&= \frac{\exp(z_k) \cdot \sum_{l=1}^{K} \exp{(z_l)} - \exp^2{(z_k)}}{(\sum_{l=1}^{K} \exp{(z_l)})^2} \\
&= \frac{\exp{(z_k)}}{\sum_{l=1}^{K} \exp{(z_l)}} \cdot (1 - \frac{\exp{(z_k)}}{\sum_{l=1}^{K} \exp{(z_l)}}) \\
&= f(z_k) \cdot (1 - f(z_k))
\end{aligned}$$

If $k \neq j$,

$$
\begin{aligned}
\frac{\partial f(z_j)}{\partial z_k} &= \frac{\partial \frac{\exp(z_j)}{\sum_{l=1}^{K}\exp(z_l)}}{\partial z_k} \\
&= \frac{-\exp(z_j)\cdot\exp(z_k)}{(\sum_{l=1}^{K}\exp(z_l))^2} \\
&= -f(z_j)\cdot f(z_k)
\end{aligned}
$$

So the derivative of the softmax function is given as,

$$
\frac{\partial f(z_j)}{\partial z_k} = \begin{cases} f(z_k)\cdot(1-f(z_k)), & \text{if } \mathrm{k}=\mathrm{j} \\ -f(z_k)\cdot f(z_j), & \text{if } \mathrm{k}\neq\mathrm{j} \end{cases}
$$

- Derivative of Cross Entropy Loss with Softmax

$$
\begin{aligned}
\frac{\partial E}{\partial z_k} &= \sum_{l=1}^{K}\frac{\partial E}{\partial f(z_l)}\times\frac{\partial f(z_l)}{\partial z_k} \\
&= \sum_{l=1}^{K}\frac{\partial(-\sum_{k=1}^{K}t_k\ln(f(z_l)))}{\partial f(z_l)}\times\frac{\partial f(z_l)}{\partial z_k} \\
&= -\sum_{l=1}^{K}t_l\frac{1}{f(z_l)}\times\frac{\partial f(z_l)}{\partial z_k} \\
&= -t_k\cdot\frac{1}{f(z_k)}\cdot f(z_k)(1-f(z_k)) - \sum_{l\neq k}^{K}t_l\frac{1}{f(z_l)}(-f(z_k)\cdot f(z_l)) \\
&= -t_k(1-f(z_k)) + \sum_{l\neq k}^{K}t_l\cdot f(z_k) \\
&= \sum_{l=1}^{K}t_l\cdot f(z_l) - t_k
\end{aligned}
$$

t is a one hot encoded vector for the labels, so $\sum_l t_l = 1$. So we have,

$$
\frac{\partial E}{\partial z_k} = f(z_k) - t_k
$$

- Derivative of $w_{kj}^{(2)}$

$$
\begin{aligned}
\frac{\partial E}{\partial w_{kj}^{(2)}} &= \frac{\partial E}{\partial z_k}\times\frac{\partial z_k}{\partial w_{kj}^{(2)}} \\
&= (f(z_k)-t_k)\cdot h(\sum_{i=0}^{D}w_{ji}^{(1)}x_i) \\
&= (y_k-t_k)\cdot h(\sum_{i=0}^{D}w_{ji}^{(1)}x_i)
\end{aligned}
$$

- Derivative of $w_{ji}^{(1)}$

$$
\begin{aligned}
\frac{\partial E}{\partial w_{ji}^{(1)}} &= \sum_{k}^{K} \frac{\partial E}{\partial z_k} \times \frac{\partial z_k}{\partial h(\sum_{i=0}^{D} w_{ji}^{(1)} x_i)} \times \frac{\partial h(\sum_{i=0}^{D} w_{ji}^{(1)} x_i)}{\partial w_{ji}^{(1)}} \\
&= \sum_{k}^{K} (f(z_k) - t_k) \cdot w_{kj}^{(2)} \cdot (1 - h^2(\sum_{i=0}^{D} w_{ji}^{(1)} x_i))) \cdot x_i \\
&= \sum_{k}^{K} (y_k - t_k) \cdot w_{kj}^{(2)} \cdot (1 - h^2(\sum_{i=0}^{D} w_{ji}^{(1)} x_i))) \cdot x_i
\end{aligned}
$$