Nhat Nguyen Comp 5600

Neural Network

Part 1

$$p(C_k \mid \phi) = y_k(\phi) = \frac{e^{a_k}}{\sum e^{a_i}}$$

$$\Rightarrow \frac{d y_k}{d a_j} = \frac{d \frac{e^{a_k}}{\sum e^{a_i}}}{d a_j},$$

Let $M = \sum_i e^{a_i}$, using quotient rule with $k = j$

$$\Leftrightarrow \frac{d \frac{e^{a_k}}{M}}{d a_j} = \frac{M e^{a_k} - e^{a_k} \cdot e^{a_k}}{M^2} = \frac{M e^{a_k}}{M^2} - \frac{e^{a_k^2}}{M^2}$$

$$\Leftrightarrow \frac{d \frac{e^{a_k}}{M}}{d a_j} = \frac{e^{a_k}}{M} - \frac{e^{a_k^2}}{M^2}$$

$$= y_k - (y_k)^2$$

$$= y_k(1 - y_k) \text{ or } y_k(1 - y_j) \text{ since } k = j$$

Similarly, let $k \neq j$

$$\Leftrightarrow \frac{d \frac{e^{a_k}}{M}}{d a_j} = \frac{0 - e^{a_k} \cdot e^{a_j}}{M^2} = -\frac{(e^{a_k})}{M} \cdot \frac{e^{a_j}}{M}$$

$$= -y_k y_j$$

let $I_{kj} = \begin{cases} 1, & k = j \\ 0, & k \neq j \end{cases}$

$$\boxed{\frac{d y_k}{d a_j} = y_k(I_{kj} - y_j)}$$

Part 2

$E(w_1, \ldots, w_K) = -\ln P(T \mid w_1, \ldots, w_K) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln(y_{nk})$

Using the chain rule with respect to w_j

$\frac{dE}{dw_j} = -\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} \cdot \frac{dy_{nk}}{dw_j}$. let $a_j = w_j^T \varphi_n$

$(=) \frac{dE}{dw_j} = -\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} \cdot \frac{dy_{nk}}{dw_j} \cdot \frac{da_j}{da_j}$

$(=) \frac{dE}{dw_j} = -\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} \cdot \frac{dy_{nk}}{da_j} \cdot \frac{da_j}{dw_j}$

According to the first problem, the derivative of $\frac{dy_{nk}}{da_j}$ is :

$\frac{dy_{nk}}{da_j} = y_{nk}(I_{kj} - y_{nj})$

In addition, the derivative of $\frac{da_j}{dw_j}$ according to problem one ;

$da_j = dw_j^T \varphi_n \Rightarrow \frac{da_j}{dw_j} = \varphi_n$

$\Rightarrow \frac{dE}{dw_j} = -\sum_{n=1}^{N} \sum_{k=1}^{K} \cancel{y_{nk}} (I_{kj} - y_{nj}) \frac{t_{nk}}{\cancel{y_{nk}}} \varphi_n$

$(=) \frac{dE}{dw_j} = -\sum_{n=1}^{N} \sum_{k=1}^{K} (I_{kj} - y_{nj}) t_{nk} \cdot \varphi_n$

Consider both $k = j$ and $k \neq j$

At $k = j$, $I_{kj} = 1$ (1st problem), and $t_{nk} = t_{nj}$

At $k \neq j$, $I_{kj} = t_{nk} = 0$ and $t_{nj} = 1$ (according to the one-hot encoded vector, t_{nk} is 0 for any incorrect class and since t_{nj} already represents the true class, t_{nj} is 1.)

This means that,

$$\sum_{k=1}^{K} (I_{kj} - y_{nj})\, t_{nk} = (\underset{I_{kj}=1}{1} - y_{nj})\, \underset{t_{nk}=t_{nj}}{t_{nj}} + \sum_{\substack{k=1 \\ k \neq j}}^{K} (\underset{I_{kj}=0}{0} - y_{nj}) \cdot \underset{t_{nk}=0}{0}$$

$\Rightarrow t_{nj}(1 - y_{nj}) = (t_{nj} - y_{nj})$

$\Rightarrow$ Therefore,

$$\frac{dE}{dw_j} = -\sum_{n=1}^{N} (t_{nj} - y_{nj})\, \varphi_n$$