Date

CSE 253 ASSIGNMENT 2

Puneeth Bommi Reddy

A53043725

1. Solution is the same as that in Bishop's PRML chapter 3.

$t = y(\tilde{x}, \theta) + \epsilon$

where $\tilde{x} = [1, x]$ and $y = \sum x_k \theta_k = \theta^T x$

$\therefore\ p(t | x, \theta, \sigma^2) = \mathcal{N}(t | y(x, \theta), \sigma^2)$ σ^2 is assumed variance.

The joint distribution is given by:

$$p(\vec{t} | X, \theta, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(t^n | \theta^T \tilde{x}^n, \sigma^2) \quad \left[\vec{t} = t_1 \cdots t_n \right.$$

$$\Rightarrow \ln p = \sum_{n=1}^{N} \ln \mathcal{N}(t^n | \theta^T \tilde{x}^n, \sigma^2)$$

$$= -\frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln 2\pi - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (t^n - \theta^T \tilde{x}^n)^2$$

By the ML rule, $\theta^* = \arg\max_{\theta} \ln p$

$$= \arg\min \sum_{n=1}^{N} (t^n - \theta^T \tilde{x}^n)^2 \quad \left[\begin{array}{c} \text{Other terms} \\ \text{are independent} \\ \text{of } \theta \end{array} \right]$$

$$= \arg\min \sum_{i=1}^{N} (t^i - f((1, x^i), \theta))^2 \ //$$

↑

changed summation index to i, substituted $\tilde{x}^i = [1, x^i]$ and $f(x, \theta) = \theta^T x$.

Date

CSE 253 ASSIGNMENT 2

Puneeth BommiReddy

A53093725

2ⓐ $$E = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{\overset{\text{sample}}{n}k} \ln \underset{\uparrow\ k^{th}\text{ output}}{y_{nk}}$$

Output Layer:

$$-\frac{\partial E}{\partial a_i} = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\frac{\partial}{\partial a_i}\left(\ln y_{nk}\right)$$

$$= \sum\sum \frac{t_{nk}}{y_{nk}}\frac{\partial y_{nk}}{\partial a_i}$$

but $y_{nk} = \frac{e^{a_{nk}}}{\sum e^{a_{nk}}}$

$\delta_{ij} = 1$ iff $i=j$

$= 0$ if $i \neq j$

But $$\frac{\partial y_{nk}}{\partial a_i} = \frac{\partial}{\partial a_i}\left[\frac{e^{a_{nk}}}{\sum_k e^{a_{nk}}}\right] = \frac{e^{a_{nk}}}{\sum_k e^{a_{nk}}}\delta_{ki} - \frac{1\cdot e^{a_{nk}}e^{a_{ni}}}{\left(\sum e^{a_{nk}}\right)^2}$$

$$\therefore \frac{-\partial E}{\partial a_i} = \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{t_{nk}}{\cancel{y_{nk}}}\cancel{y_{nk}}\left[\delta_{ki} - \frac{e^{a_{ni}}}{\sum_k e^{a_{nk}}}\right]$$

$$= \sum_n\sum_k t_{nk}\left[\delta_{ki} - y_{ni}\right]$$

$$= \sum_n\left[t_{ni} - y_{ni}\sum_k t_{nk}\right]$$ but $\sum_k t_{nk} = 1$

$$= \sum_n\left(t_{ni} - y_{ni}\right)$$

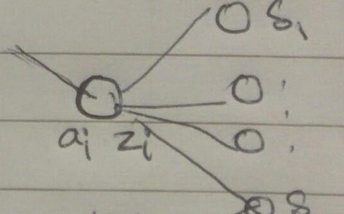
Per sample:

$$\delta_k = -\frac{\partial E_n}{\partial a_k} = \underline{\underline{(t_k - y_k)}}$$

Hidden Layer:

$$\delta_j = -\frac{\partial E_n}{\partial a_j} = -\sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial a_j}$$

a_k are from the output layer



But $$\frac{\partial a_k}{\partial a_j} = \frac{\partial a_k}{\partial z_j}\frac{\partial z_j}{\partial a_j}$$

$$= w_{jk}\, h'(a_j)$$

where $z_j = h(a_j)$
↑
activation

$$-\frac{\partial E_n}{\partial a_j} = +\sum_k \delta_k w_{jk} h'(a_j) = \underline{\underline{+h'(a_j)\sum \delta_k w_{jk}}}$$

Date

$$w_{jk} = w_{jk} + \alpha\,\delta_k z_j$$

ⓑ i. $$\frac{\partial E}{\partial w_{jk}} = \frac{\partial E}{\partial a_k}\frac{\partial a_k}{\partial w_{jk}} = -\delta_k z_j \;;\; w_{jk} = w_{jk} + \alpha\,\delta_k z_j$$

ii $$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial a_j}\frac{\partial a_j}{\partial w_{ij}} = -\delta_j x_i \;;\; w_{ij} = w_{ij} + \alpha\,\delta_j x_i$$

ⓒ (i) $w_{jk} = w_{jk} + \alpha\,\delta_k z_j$

$W_{HO} = W_{HO} + \alpha\, z\,(t-y)^T$ $\quad t = [t_1 .. t_O]^T ; y = [y_1 .. y_O]^t \; z = [z_1 .. z_H]^t$

but $a_j = \sum x_i w_{ij}$

$\vec{a} = (x^T W_{IH})^T = W_{IH}^T x \Rightarrow z = h(\vec{a}) = h(W_{IH}^T x)$

$\therefore W_{HO} = W_{HO} + \alpha\,[h(W_{IH}^T x)](t-y)^T$

(ii) $w_{ij} = w_{ij} + \alpha\,\delta_j x_i$

but $\delta_j = h'(a_j)\sum \delta_k w_{jk} = \overset{h'(a_j)}{\sum_k} (t_k - y_k)\, w_{jk} \;\exists\, \vec{\delta}_j = [W_{HO}(t-y)] .* h'(a_j)$

$\Rightarrow \vec{\delta}_j = h'(a_j) .* W_{HO}(t-y)$

$\therefore W_{IH} = W_{IH} + \alpha x\,[h'(W_{IH}^T x) .* W_{HO}(t-y)]^T$

Procedure: [For stochastic descent]

– Initialization:

$\vec{x} = [x_1 \dots x_I]^T ; \vec{t} = [t_1 \dots t_O]^T ; W_{IH} = (W_{IH})_{init} \; W_{HO} = (W_{HO})_{init}$

– Update:

$\vec{a}_{H1} = W_{IH}^T x \;;\; \vec{h}'_{H1} = h'(a_{H1}) \;;\; \vec{Z}_{H1} = h(a_{H1})$

$$\vec{y} = \frac{e^{W_{HO}^T \vec{Z}_{H1}}}{\sum_k e^{(W_{HO}^T Z_{H1})_k}}$$

$$(W_{IH})_{new} = W_{IH} + \alpha\vec{x}\,\overbrace{[\vec{h}' \cdot W_{HO}(\vec{t} - \vec{y})]}^{b}{}^T$$

$$(W_{HO})_{new} = W_{HO} + \alpha\vec{Z}\,(t - \vec{y})^T$$

For batch descent:

Init : $X = [\vec{x}_1 \dots \vec{x}_n]$; $T = [\vec{t}_1, \dots \vec{t}_n]$

$W_{IH} \sim U\left[\frac{\sqrt{6}}{fanin + fanout}\right] \sim W_{HO}$

Update : $A_{Hn} = W_{IH}^T X_{In}$ $\quad H'_{Hn} = h'(A_{Hn})$ $\quad Z_{Hn} = h(A_{Hn})$

$Y_{ON} = e^{W_{HO}^T Z_{Hn}}$ ← normalize.

$$W_{IH} = W_{IH} + \alpha X \left[H' \cdot W_{HO} (T - Y) \right]^T$$

$$W_{HO} = W_{HO} + \alpha Z \left[T - Y \right]^T$$

Note: Z will have additional bias terms appended as:

$$Z = \left[1_{N1}^T ; h(A) \right]$$

2g. i. $f(z) = 1/_{1+e^{-z}}$ $\quad \frac{d}{dz} f(z) = \frac{-1}{(1+e^{-z})^2} - e^{-z} = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}} \frac{e^{-z}}{1+e^{-z}}$

$$= f(z)(1 - f(z))$$

ii $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ $\quad \frac{d}{dx}\left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right) = \frac{e^x + e^{-x}}{e^x + e^{-x}} - \frac{(e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2}$

$$= 1 - f(x)^2$$

iii $f(z) = \begin{cases} z & z > 0 \\ 0 & \text{elsewhere} \end{cases}$ $\quad \frac{d}{dz} f(z) = \begin{cases} 1 & z > 0 \\ 1/2 & \text{numerical approx at } z = 0 \\ 0 & z < 0 \end{cases}$ //

$\lim_{z \to 0^+} f(z) = 1$ $\quad \lim_{z \to 0^-} f(z) = 0$

$$\therefore \frac{1}{2}\left[\lim_{z \to 0^+} f(z) + \lim_{z \to 0^-} f(z) \right] = 1/2 \ //$$