CSE253 Homework 2
Problems from Bishop

Xiyun Liu
A53099348
xil429@eng.ucsd.edu

1

From the question, we know that

$$t = y(x) + \epsilon \tag{1}$$

and $\epsilon \sim N(0, \sigma^2),\ t \sim N(y(x), \sigma^2)$ Thus, the distribution of ϵ is given by

$$p(\epsilon) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp(-\frac{\epsilon^2}{2\sigma^2}) \tag{2}$$

Since $\epsilon = t - h(x)$ and $h(x) = y(x; w)$

$$p(t^n|x^n) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp(-\frac{(t^n - y(x^n; w))^2}{2\sigma^2}) \tag{3}$$

Since the likelihood is $\Pi_{n=1}^N P(t^n|x^n)$, we could take the negative of log of equation (4) to get the following error function for this problem

$$E = -\Sigma_{n=1}^N [\log \frac{1}{(2\pi\sigma^2)^{1/2}} - \frac{(t^n - y(x^n; w))^2}{2\sigma^2}] \tag{4}$$

$$= \frac{1}{2\sigma^2} \Sigma_{n=1}^N [(t^n - y(x^n; w))^2] + \frac{N}{2} \log(2\pi\sigma^2) \tag{5}$$

Since the coefficient of the first part and the whole second part of equation(6) is independent of w,

$$E = \Sigma_{n=1}^N (t^n - y(x^n; w))^2 \tag{6}$$

The optimal parameter w can be obtained from the principle of maximum likelihood.

$$w^* = argmin_w E \tag{7}$$

$$= argmin_w \Sigma_{n=1}^N (t^n - y^n)^2 \tag{8}$$

proved.

2

2.1 Derivation

(1) The derivative for output layer δ_k

From the problem we know that

$$\delta_k = -\frac{\partial E}{\partial a_k} \tag{9}$$

$$= -\Sigma_{k'} \frac{\partial E}{\partial y_{k'}^{(n)}} \frac{\partial y_{k'}^{(n)}}{\partial a_k^{(n)}} \tag{10}$$

The cross-entropy cost function and the prediction rule for softmax regression is

$$E = -\Sigma_n \Sigma_{k'=1}^{c} t_{k'}^{(n)} ln y_{k'}^{(n)} \tag{11}$$

$$y_k^{(n)} = \frac{exp(a_k^n)}{\Sigma_{k'} exp(a_{k'}^n)} \tag{12}$$

Thus

$$\frac{\partial E}{\partial y_{k'}^{(n)}} = -\frac{t_{k'}^{(n)}}{y_{k'}^{(n)}} \tag{13}$$

$$\frac{\partial y_{k'}^{(n)}}{\partial a_k} = \begin{cases} y_k^{(n)}(1-y_k^{(n)}), & \text{if } k = k'. \\ -y_k^{(n)} y_{k'}^{(n)}, & \text{otherwise.} \end{cases} \tag{14}$$

Then equation 10 can be written as

$$\delta_k = \frac{t_k^{(n)}}{y_k^{(n)}} y_k^{(n)}(1-y_k^{(n)}) - \Sigma_{k' \neq k} \frac{t_{k'}^{(n)}}{y_{k'}^{(n)}} y_k^{(n)} y_{k'}^{(n)} \tag{15}$$

$$= t_k^{(n)} - \Sigma_{k'} t_{k'}^{(n)} y_k^{(n)} \tag{16}$$

Since $\Sigma_{k'} t_k^{(n)} = 1$, we can write Equation (16) as

$$\delta_k = t_k - y_k \tag{17}$$

(2) The derivative for hidden layer δ_j

$$\delta_j = \frac{\partial E}{\partial a_j} = -\Sigma_k \frac{\partial E}{\partial a_k} \frac{\partial a_k}{\partial a_j} = -\Sigma_k \frac{\partial E}{\partial a_k} \frac{\partial a_k}{\partial z_j} \frac{\partial z_j}{\partial a_j} \tag{18}$$

where z is the hidden unit and $a_k = \Sigma_j w_{jk} z_j$, $\frac{\partial a_k}{\partial z_j} = w_{jk}$
Let $g(a_j)$ be the hidden unit function mapping a_j to z_j. With $\delta_k = -\frac{\partial E}{\partial a_k} = t_k - y_k$, Equation (18) can be written as

$$\delta_j = \Sigma_k (t_k - y_k) w_{jk} g(a_j)' \tag{19}$$

$$= g(a_j)(1 - g(a_j)) \Sigma_k (t_k - y_k) w_{jk} \tag{20}$$

2.2 Update rule

$$
\begin{align}
w_{jk} &= w_{jk} - \eta \frac{\partial E}{\partial w_{jk}} \tag{21}\\
&= w_{jk} - \eta \frac{\partial E}{\partial a_k} \frac{\partial a_k}{\partial w_{jk}} \tag{22}\\
&= w_{jk} + \eta \delta_k z_j \tag{23}\\
&= w_{jk} + \eta (t_k - y_k) z_j \tag{24}\\
w_{ij} &= w_{ij} - \eta \frac{\partial E}{\partial w_{ij}} \tag{25}\\
&= w_{ij} - \eta \frac{\partial E}{\partial a_j} \frac{\partial a_j}{\partial w_{ij}} \tag{26}\\
&= w_{ij} + \eta \delta_j x_i \tag{27}\\
&= w_{ij} + \eta g(a_j)(1 - g(a_j)) \Sigma_k (t_k - y_k) w_{jk} x_i \tag{28}
\end{align}
$$

3 Vectorize computation

Let w_{HO} be a $j \times k$ metric represents the weights from hidden units to output, $W_{input\ to\ hidden}$ be a $i \times j$ metric represents the weights from input to hidden units, $\vec{x}$ is the input column vector , $\vec{t}$ is the k target column vector, $\vec{y}$ is the k predict column vector, $\vec{G}$ is the element-wise product of $g(a_j)$ and $1 - g(a_j)$, then

$$
\begin{align}
w_{HO} &= w_{HO} + \eta [(\vec{t} - \vec{y}) \cdot \vec{z}^T]^T \tag{29}\\
w_{IH} &= W_{IH} + \eta \vec{x} \cdot (G * (\vec{t} - \vec{y}) \cdot w_{HO}^T) \tag{30}
\end{align}
$$

where * means element-wise product