Assisgnment 5 part 1

$$P(C_K|\phi) = y_K(\phi) = \frac{\exp(a_K)}{\sum_j \exp(a_j)} \qquad a_K = w_K^T \phi$$

derivates y_K with respect of a_j $\qquad \log \text{Softmax}(a)_i = a_i - \log \sum_j \exp(a_j)$

$$\frac{\partial y_K}{\partial a_j} = y_K(I_{kj} - y_j)$$

Next we need likihood function using the 1-of k coding scheme

$$P(T|w_1 \dots w_K) = \prod_{n=1}^{n} \prod_{K=1}^{K} P(C_K|\phi_n)^{t_{nK}} = \prod_{n=1}^{n} \prod_{K=1}^{K} y_K(\phi_n)^{t_{nK}}$$

where $y_{nK} = y_K(\phi_n)$ and T is an $N \times K$ matrix of target variables with element t_{nK}

taking negative logarithm gives the cross entropy error function:

$$E(w_1 \dots w_K) = -\ln p(T|w_1 \dots w_K) = -\sum_{n=1}^{n} \sum_{K=1}^{K} t_{nK} \ln(y_{nK})$$

gradient of the error function wrt to one parameter vector w_j

$$\nabla_{w_j} E(w_1 \dots w_K) = \sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n$$

$$\nabla_{w_K} \nabla_{w_j} E(w_1 \dots w_N) = -\sum_{n=1}^{N} y_{nK}(I_{kj} - y_{nj}) \phi_n \phi_n^T$$

Differentiating the softmax function First we differentiate $y_K = h(a_K) = \text{Softmax}(a_K)$ with respect to a_K

$$\frac{\partial y_K}{\partial a_K} = \frac{\partial}{\partial a_K} \text{Softmax}(a_K) = \frac{\partial}{\partial a_K}\left(\frac{\exp a_K}{\sum_j \exp a_j}\right) = \frac{\exp a_K \sum_{j \neq K} \exp a_j}{\left(\sum_j \exp a_j\right)^2} \quad \text{(product rule)}$$

$$= y_K \frac{\sum_{j \neq K} \exp a_j}{\sum_j \exp a_j} = y_K\left(1 - \frac{\exp a_K}{\sum_j \exp a_j}\right) = y_K(1 - y_K)$$

we differiate $y_K = \text{Softmax}(a_K)$ with respect to a_i to obtain

$$\frac{\partial y_K}{\partial a_i} = \frac{\partial}{\partial a_i} \text{softmax}(a_K) = \frac{-\exp(a_K)\exp(a_i)}{\left(\sum_j \exp a_j\right)^2} = -y_i y_K$$

These two results may be summarized using the Kronecker delta function δ_{Kj} as

$$\frac{\partial y_K}{\partial a_i} = \frac{\partial}{\partial a_i} \text{softmax}(a_K) = y_K(\delta_{Kj} - y_i)$$

Given $E(w_1 \dots w_K) = -\ln P(T|w_1 \dots w_K) = -\sum_{n=1}^{N} \sum_{K=1}^{K} t_{nK} \ln y_{nK}$ mathematically show that the derivitive E will take the following form:

$$\nabla_{w_j} E(w_1, \dots w_K) = \sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n$$

minimize the negative log likelihood function

$$E(\omega) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

Taking the gradient of the error function:

$$\nabla_w E(\omega) = \sum_{n=1}^{N} (y_n - t_n) \phi_n = \Phi^T(Y - T)$$

minimize error function efficient

$$E(\omega) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

$$\nabla_w E(\omega) = \Phi^T(Y - T)$$

$$\nabla_w \nabla_w E(\omega) = H = \Phi^T R \Phi$$

Using Newton Raphson

$$w^{t+1} = (\Phi^T R \Phi)^{-1} \Phi^T R z$$

define z as

$$z = \Phi w - R^{-1}(y - t)$$

The likelihood of an example (x, t) is

$$P(t|x) = \prod_{k=1}^{K} y_k^{t_k}$$

where $y_K = y_K(x) = \frac{\exp(a_K(x))}{\sum_{j=1}^{K} \exp(a_j(x))}$

The likihood of a data set $D = \{(x_n, t_n)\}_{n=1}^{N}$ is

$$P(D|w) = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

maximizing the data likelihood is equivalent to minimizing the negative log data likelihood

$$E(\omega) = -\log \prod_{n=1}^{N} p(t_n|x_n)$$

$$= -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log y_{nk}$$

$$= \sum_{n=1}^{N} E_n(\omega)$$

note that

$$E_n = -\sum_{k=1}^{K} t_{nk} \log y_{nk}$$ is the cross entropy between t_n and y_n

From $a_j = w_j^T \phi$ and $y_K = \frac{\exp(a_K)}{\sum_j \exp(a_j)}$ we have

$$\nabla_{w_j} a_l = \delta_{jl} \phi \text{ and } \frac{\partial y_K}{\partial a_j} = y_K(\delta_{Kj} - y_j)$$

It follows that

$$\nabla_{w_j} y_K = \sum_l \frac{\partial y_K}{\partial a_l}(\nabla_{w_j} a_l)$$

$$= \sum_l y_K(\delta_{Kl} - y_l)\delta_{jl}\phi$$

$$= y_K(\delta_{Kj} - y_j)\phi$$

Consider the first-order derivatives of the error $E = -\sum_{k=1}^{K} t_k \log y_k$ of point (t, x)

$$\nabla_{w_j}\left(-\sum_{k=1}^{K} t_k \log y_k\right) = -\sum_{k=1}^{K} t_k \left(\frac{1}{y_k} \nabla_{w_j} y_k\right)$$

$$= -\sum_{k=1}^{K} t_k((\delta_{kj} - y_j)\phi)$$

$$= -t_j \phi + \sum_{k=1}^{K} t_k y_j \phi$$

$$= -t_j \phi + y_j \phi$$

$$= (y_j - t_j)\phi$$

gradient of $E(\omega)$ of data set D is

$$\nabla_{w_j} E(\omega) = \nabla_{w_j}\left(\sum_{n=1}^{N} E_n(\omega)\right) = \sum_{n=1}^{N} (y_{nj} - t_{nj})\phi_n = \Phi^T(y_j - t_j)$$