Name: Himanshu Gupta

Course: Deep Learning Theory and Practice

Date: 22/02/2019

Assignment-1

1.

$\Rightarrow$ To find:

$$\frac{\partial}{\partial A} tr(AB)$$, A, B are matrices of shape n×m and m×n resp.

We can say that diagonal terms of A.B can be represented as

$$[AB]_i = \sum_{j=1}^{m} b_{ji} a_{ij}$$

where $i \in [1, n]$

$a_{ij} \in A$

$b_{ij} \in B$

Now, trace of matrix A·B can be represented as.

$$tr(AB) = \sum_{i=1}^{N} \sum_{j=1}^{M} a_{ij} b_{ji}$$

$$\Rightarrow \frac{\partial [tr(AB)]_{ij}}{\partial A_{ij}} = b_{ji}$$

$$\therefore \boxed{\frac{\partial\, tr(AB)}{\partial A} = B^T}$$

Logistic Regression

(Two Classes → C_1, C_2)

$$P(C_1|x) = \frac{P(x,C_1)\,P(C_1)}{\sum_{i=1}^{2} P(x,C_i)}$$

$$= \frac{p(x|C_1)\,p(C_1)}{p(x|C_1)\,P(C_1) + p(x|C_2)\,P(C_2)}$$

$$= \frac{1}{1+\exp(-a)} = \sigma(a)$$

where $a = \ln\left(\frac{p(x|C_1)\ p(C_1)}{p(x|C_2)\ p(C_2)}\right)$

To, simplify, lets assume that a is a linear function.

$$p(C_1|x) = \sigma(w^T x + w_0)$$

$$p(C_2|x) = 1 - \sigma(w^T x + w_0)$$

lets assume, given n training data

$(x_1\ t_1)\ (x_2\ t_2)\ \ldots\ (x_n\ t_n)$

$t_1 = 1$, if $x \in C_1$

$t_2 = 0$ if $x \in C_2$

<u>Likelihood</u> $= \prod_{n=1}^{N} (y_n^{t_n})\,(1-y_n)^{1-t_n}$

Now we have to maximize the likelihood

$$w^* = \underset{w}{\operatorname{argmax}} \left[p(t/w) \right]$$

$$w^* = \underset{w}{\operatorname{argmax}} \left[\underbrace{\sum_{n=1}^{N} t_n \ln(y_n) + (1-t_n)\ln(1-y_n)}_{L} \right]$$

$$w^{t+1} = w^t + \eta \frac{\partial L}{\partial w}\bigg|_{w=w^t}$$

We have to find $\frac{\partial L}{\partial w}$ to update w

As,

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

$$\frac{\partial L}{\partial w} = \sum_{n=1}^{N} t_n \cdot \frac{1}{y_n} \cdot \frac{\partial y_n}{\partial a_n} \cdot \frac{\partial a_n}{\partial w} + (1-t_n) \frac{-1}{(1-y_n)} \cdot \frac{\partial y_n}{\partial a} \cdot \frac{\partial a}{\partial w}$$

As we know that

$$\sigma(a) = \frac{1}{1+e^{-a}} = y$$

$$\therefore \frac{\partial \sigma(a)}{\partial a} = \frac{\partial y}{\partial a} = \frac{e^{-a}}{(1+e^{-a})^2} = (1-\sigma(a))\sigma(a)$$

$$\Rightarrow \frac{\partial L}{\partial w} = \sum_{n=1}^{N} t_n \frac{1}{y_n} \cdot y_n(1-y_n)x_n + (1-t_n) \frac{-1}{(1-y_n)} y_n(1-y_n)x_n$$

$$= \sum_{n=1}^{N} (t_n - y_n)x_n$$

$\therefore$ update eq^n:

$$\boxed{w^{t+1} = w^t + \eta \left(\sum_{n=1}^{N} (t_n - y_n)x_n \right)}$$

Now, Similarly for classes more that 2 ($k>2$)

$$\begin{bmatrix} \vdots \\ p(C_k|x) \\ \vdots \\ \vdots \end{bmatrix} = \frac{p(x|C_k)\, p(C_k)}{\sum_{j=1}^{K} p(x|C_j)\, p(C_j)} = \left[\frac{\exp(a_k)}{\sum_j \exp(a_j)}\right]$$

softmax a

$$a_j = \ln p(x|C_j)\, p(C_j)$$

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_K \end{bmatrix}$$

softmax is defined as, $\text{softmax}(a) = \begin{bmatrix} \frac{\exp(a_1)}{\sum \exp(a_i)} \\ \frac{\exp(a_2)}{\sum \exp(a_i)} \\ \vdots \\ \frac{\exp(a_K)}{\sum \exp(a_j)} \end{bmatrix}$

Assuming a_k to be linear in nature.

$$a_k = (w_k^T x + w_{k0})$$

$$a_{nk} = (w_k^T x + w_{k0}) \longrightarrow \text{for } n \text{ examples (training)}$$

Assuming there are n training examples

$$(x_1, t_1) \quad (x_2, t_2) \quad \cdots \quad (x_n, t_n)$$

$$t_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \longrightarrow \text{class that } x_n \text{ belongs to}$$

Maximizing the probability of class to which x belongs.

$$\tilde{\omega} = \underset{\tilde{\omega}}{\operatorname{argmax}} \text{ likelihood} = \prod_{n=1}^{N} \prod_{k=1}^{K} (y_{nk})^{t_{nk}}$$

$\downarrow$

k^{th} element of y_n vector.

taking log to simplify.

$$= \underset{\tilde{\omega}}{\operatorname{argmax}} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln(y_{nk})$$

Computing the gradient.

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

$$= \sum_{n=1}^{N} \left[\frac{\partial}{\partial w_j} \left(t_{nj} \ln(y_{nj}) \right) + \frac{\partial}{\partial w_j} \sum_{\substack{k=1 \\ k \neq j}}^{K} t_{nk} \ln(y_{nk}) \right]$$

$$= \sum_{n=1}^{N} \left[t_{nj} \cdot \frac{1}{y_{nj}} \cdot \frac{\partial y_{nj}}{\partial a_{nj}} \cdot \frac{\partial a_{nj}}{\partial w_j} + \sum_{\substack{k=1 \\ k \neq j}}^{K} t_{nk} \cdot \frac{1}{y_{nk}} \cdot \frac{\partial y_{nk}}{\partial a_{nj}} \cdot \frac{\partial a_{nj}}{\partial w_j} \right] \quad (1)$$

As we know y is a softmax matrix of dimension $n \times k$

$$\text{so, } y_{11} = \frac{e^{a_{11}}}{e^{a_{11}} + e^{a_{12}} + \dots + e^{a_{1n}}}$$

$$\frac{\partial y_{11}}{\partial a_{11}} = \frac{(e^{a_{11}} + e^{a_{12}} + \dots + e^{a_{1n}})(e^{a_{11}}) - e^{a_{11}} \cdot e^{a_{11}}}{(e^{a_{11}} + e^{a_{12}} + \dots + e^{a_{1n}})^2}$$

$$\frac{\partial y_{11}}{\partial a_{11}} = \frac{e^{a_{11}} \cdot \left[e^{a_{11}} + e^{a_{12}} + \cdots e^{a_{1m}} - e^{a_{11}}\right]}{\left(e^{a_{11}} + e^{a_{12}} + \cdots + e^{a_{1n}}\right)^2}$$

$$= \frac{e^{a_{11}}}{\left(e^{a_{11}} + e^{a_{12}} + \cdots + e^{a_{1n}}\right)} \cdot \left(1 - \frac{e^{a_{11}}}{e^{a_{11}} + e^{a_{12}} + \cdots e^{a_m}}\right)$$

$$= y_{11}(1 - y_{11}) \quad \text{— (2)}$$

From eq. 2.

$$\therefore \boxed{\frac{\partial y_{nj}}{\partial a_{nj}} = y_{nj}(1 - y_{nj})} \quad \text{— (3)}$$

Now finding $\frac{\partial y_{11}}{\partial a_{12}}$

$$\frac{\partial y_{11}}{\partial a_{12}} = \frac{\left(e^{a_{11}} + e^{a_{12}} + \cdots + e^{a_m}\right) \cdot 0 - e^{a_{11}}\left(e^{a_{12}}\right)}{\left(e^{a_{11}} + e^{a_{12}} + \cdots + e^{a_{1n}}\right)^2}$$

$$= -y_{11} \cdot y_{12} \quad \text{— (4)}$$

$\therefore$ from eq (4), we can generalize

$$\frac{\partial y_{nk}}{\partial a_{nj}} = -y_{nk} \cdot y_{nj} \quad \text{— (5)}$$

Putting values of eq (3) and (5) into eq. (1)

$$\frac{\partial L}{\partial w_j} = \sum_{n=1}^{N} \left[t_{nj} \cdot \frac{1}{y_{nj}} \cdot y_{nj}(1 - y_{nj}) \cdot x_n + \sum_{\substack{k=1 \\ k \neq j}}^{K} t_{nk} \cdot \frac{1}{y_{nk}} (-y_{nk} y_{nj}) x_n \right]$$

$$= \sum_{n=1}^{N} \left[t_{nj}(1 - y_{nj}) x_n + \left[- t_{n1} y_{nj} x_{n1} - t_{n2} y_{nj} x_n - \cdots - t_{nk} y_{nj} x_n \right] \right]$$

$$= \sum_{n=1}^{N} \left[t_{nj} x_n - t_{nj} y_{nj} x_n - t_{n1} y_{nj} x_n - t_{n2} y_{nj} x_n - \dots - t_{nk} \cdot y_{nj} \cdot x_n \right]$$

$$= \sum_{n=1}^{N} \left[t_{nj} x_n - y_{nj} x_n (t_{n1} + t_{n2} + \dots t_{nk}) \right]$$

We know that

$$t_{n1} + t_{n2} + \dots + t_{nk} = 1$$

$$\therefore \frac{\partial L}{\partial w_j} = \sum_{n=1}^{N} \left[(t_{nj} - y_{nj}) x_n \right]$$

$$\therefore \boxed{\frac{\partial L}{\partial w_{[1,2..k]}} = \sum_{n=1}^{N} (t_n - y_n) x_n}$$

So, the update equation will be:

$$w^{t+1} = w^t + \eta \left[\sum_{n=1}^{N} (t_n - y_n) x_n \right]$$