Assignment I

Ans 1:-

$$A = \begin{bmatrix} \vec{a}_{11} & \vec{a}_{12} & \cdots & \vec{a}_{1m} \\ \vec{a}_{21} & \vec{a}_{22} & \cdots & \vec{a}_{2m} \\ \vdots & \vdots & & \vdots \\ \vec{a}_{n1} & \vec{a}_{n2} & \cdots & \vec{a}_{nm} \end{bmatrix}_{n \times m}$$

$$B = \begin{bmatrix} \vec{b}_{11} & \vec{b}_{12} & \cdots & \vec{b}_{1n} \\ \vec{b}_{21} & \vec{b}_{22} & \cdots & \vec{b}_{2n} \\ \vdots & \vdots & & \vdots \\ \vec{b}_{m1} & \vec{b}_{m2} & \cdots & \vec{b}_{mn} \end{bmatrix}_{m \times n}$$

$$AB = \begin{bmatrix} \sum_{i=1}^{m} \vec{a}_{1i} \vec{b}_{i1} & \sum_{i=1}^{m} \vec{a}_{1i} \vec{b}_{i2} & \cdots & \sum_{i=1}^{m} \vec{a}_{1i} \vec{b}_{in} \\ \sum_{i=1}^{m} \vec{a}_{2i} \vec{b}_{i1} & \sum_{i=1}^{m} \vec{a}_{2i} \vec{b}_{i2} & \cdots & \sum_{i=1}^{m} \vec{a}_{2i} \vec{b}_{in} \\ \vdots & & & \\ \sum_{i=1}^{m} \vec{a}_{ni} \vec{b}_{i1} & \sum_{i=1}^{m} \vec{a}_{ni} \vec{b}_{i2} & \cdots & \sum_{i=1}^{m} \vec{a}_{ni} \vec{b}_{in} \end{bmatrix}$$

Now

$$tr(AB) = \sum_{i=1}^{m} a_{1i} b_{i1} + \sum_{i=1}^{m} a_{2i} b_{i2} + \cdots + \sum_{i=1}^{m} a_{ni} b_{in}$$

P.T.O.

So,

$$\frac{\partial \, tr(AB)}{\partial a_{ij}} = \sum_{i=1}^{m} b_{i1} + \sum_{i=1}^{m} b_{i2} + \cdots + \sum_{i=1}^{m} b_{in}$$

$$= b_{ji}$$

i.e.

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

Ans (2) :

Binary classes :- ($K=2$)

Suppose we have two classes C_1 and C_2 and the training data sets given as

$$(x_1, t_1) \; (x_2, t_2) \cdots (x_N, t_N)$$

So, the posteriors of class C_1

$$p(C_1/x) = \frac{p(x, C_1)}{p(x)}$$

$$= \frac{p(x/C_1) \, p(C_1)}{\sum_{i=1}^{N} p(x_i)}$$

$$= \frac{p(\underline{x}/C_1) \cdot p(C_1)}{\sum_{i=1}^{2} p(\underline{x}, C_i)}$$

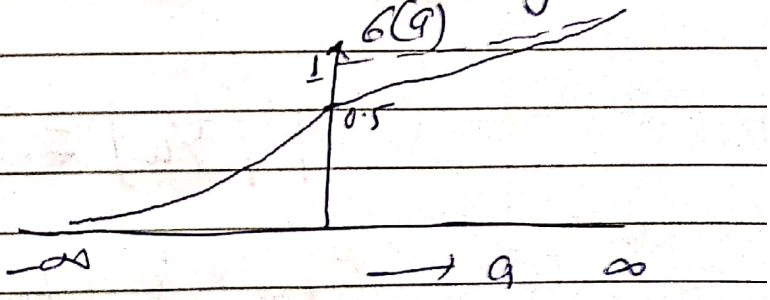
$$= \frac{p(\underline{x}/C_1) \cdot p(C_1)}{p(\underline{x}/C_1)\,p(C_1) + p(\underline{x}/C_2)\,p(C_2)}$$

Say,

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

where $a = \ln\left\{\frac{p(\underline{x}/C_1)\,p(C_1)}{p(\underline{x}/C_2) \cdot p(C_2)}\right.$

and $\sigma(a)$ is also called as sigmoid fun^n.



$$\Rightarrow p(C_1/\underline{x}) = \sigma(w^T \underline{x} + w_0) \quad \text{--- (1)}$$

Assuming the probabilities are sigmoid fun^n

So, $p(C_2/\underline{x}) = 1 - p(C_1/\underline{x})$

So, if we know the weights, then we can get the probabilities of $p(C_1/\underline{x})$ & $p(C_2/\underline{x})$

P.T.O

Now,

When $(W^T x + W_0) > 0$ then the x belongs to C_1 (say)

and when $(W^T x + W_0) < 0$ then x belongs to C_2 (say).

Also when

$W^T x + W_0 = 0$ this means x is ~~right~~ on the decision surface.

eqn (1) can also be written as,

$$y_n = \sigma(W^T x_n + W_0) = P[C_1/x_n]$$

So, as per likelihood

$$P[t/W] = \prod_{n=1}^{N} (y_n)^{t_n} (1-y_n)^{1-t_n} \quad \text{--- (ii)}$$

i.e. if $t_n = 1$ then Maximize y_n

$t_n = 0$ then Maximize $(1-y_n)$

For maximum likelihood,

$$W^* = \underset{W}{\operatorname{argmax}} \left(P[t/W] \right) \quad \text{--- (iii)}$$

Since we know that the "log" of likelihood is also the likelihood.

So,

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

$$= \underset{w}{\operatorname{argmax}} \left\{ \ln \left[\prod_{n=1}^{N} y_n^{t_n} \cdot (1-y_n)^{(1-t_n)} \right] \right\}$$

$$= \underset{w}{\operatorname{argmax}} \left\{ \sum_{n=1}^{N} t_n \ln(y_n) + \sum_{n=1}^{N} (1-t_n) \ln(1-y_n) \right\}$$

$$= \underset{w}{\operatorname{argmax}} \left\{ \sum_{n=1}^{N} t_n \ln\left[\sigma(w^T x_n + w_0) \right] + \sum_{n=1}^{N} (1-t_n) \ln\left[1 - \sigma(w^T x_n + w_0) \right] \right\}$$

This is called L [loss] also called log likelihood loss.

Now,

$$\frac{\partial}{\partial a} \sigma(a) = \frac{\partial}{\partial a} \left[\frac{1}{1+e^{-a}} \right]$$

$$= \frac{\partial}{\partial a} \left[\frac{e^a}{1+e^a} \right]$$

$$= \frac{e^{-a}}{(1-e^{-a})^2}$$

P.T.O.

$$\Rightarrow \boxed{\frac{\partial \sigma(a)}{\partial a} = \{1-\sigma(a)\}\,\sigma(a)} \quad \text{(iv)}$$

$$\Rightarrow \frac{\partial L}{\partial w} = \sum_{n=1}^{N} t_n \frac{\partial}{\partial w}[\ln(y_w)] + \sum_{n=1}^{N}(1-t_n)\frac{\partial}{\partial w}[\ln(1-y_w)]$$

Since $\frac{\partial}{\partial x}\log x = \frac{1}{x}$

So,

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

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

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

$$\Rightarrow \boxed{\frac{\partial L}{\partial w} = \sum_{n=1}^{N}(t_n - y_n)\,x_n} \quad \text{(v)}$$

Now, As per gradient ascent method, for moving from one point to another we have a rule

$$\boxed{w^{t+1} = w^t + \eta \left.\frac{\partial L}{\partial w}\right|_{w=w^t}}$$

So, putting eq^n (v)th value in above eq^n we get,

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

Multi- class ($K > 2$) :-

$$\begin{bmatrix} \vdots \\ p(C_K/x_K) \\ \vdots \end{bmatrix}_{K\times 1} = \frac{p\left[x/C_K\right) p(C_K)}{\sum_{j=1}^{K} p\left[x/C_j\right) p(C_j)}$$

$$= \frac{\exp(a_K)}{\sum_{j=1}^{K} \exp(a_j)} \quad \text{---(1)}$$

Where.

$$\exp(a_k) = p[x/c_k)\ p(c_k)$$

$$\exp(a_j) = p[x/c_j) \cdot p(c_j)$$

And

$$a_j = \ln\left[p[x/c_j) \cdot p[c_j)\right]$$

$$\underline{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_K \end{bmatrix}_{K\times 1}$$

eq^n ① is called the softmax of "$\underline{a}$".

Logistic regression for K-classes

$$a_{nk} = w_k^T x_n + w_{k0}$$

where $\underline{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_K \end{bmatrix}_{K\times 1}$

output $\boxed{\underline{y} = \text{Softmax}(\underline{a})}$

So, for training data $(x_n, t_n)_{n=1}^{N}$ Assuming t_n is one hot encoded.

then the likelihood model is defined as,

$$p(t/w) = \prod_{n=1}^{N} \prod_{k=1}^{K} (y_{nk})^{t_{nk}}$$

Maximum likelihood model is,

$$\tilde{w}^* = \underset{\tilde{w}}{\operatorname{argmax}} \left[\prod_{n=1}^{N} \prod_{k=1}^{K} (y_{nk})^{t_{nk}} \right]$$

Where $\tilde{w} = \begin{bmatrix} \tilde{w}_1^T \\ \tilde{w}_2^T \\ \vdots \\ \tilde{w}_K^T \end{bmatrix}$ $\tilde{w} = \begin{bmatrix} w \\ 1 \end{bmatrix}$.

Since we know that the "log" of likelihood is also a likelihood hence,

$$\tilde{w}^* = \underset{\tilde{w}}{\operatorname{argmax}} \left[\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln(y_{nk}) \right]$$

P.T.O.

Now, $\frac{\partial}{\partial a} \text{Softmax}(a) = \frac{\partial}{\partial a} \frac{e^a}{\sum_{j=1}^{K} e^{a_j}}$

$$= \frac{\partial}{\partial a} \left\{ \frac{e^a}{e^{a_1} + e^{a_2} + e^{a_3} + \cdots e^{a_K}} \right\}$$

Using quotient rule

$$\frac{\partial}{\partial a} \text{Softmax}(a) = \frac{e^a \times \sum_{j=1}^{K} e^{a_j} - (e^a)^2}{\left(\sum_{j=1}^{K} e^{a_j} \right)^2}$$

$$= \frac{e^a}{\sum_{j=1}^{K} e^{a_j}} - \frac{(e^a)^2}{\left(\sum_{j=1}^{K} e^{a_j} \right)^2}$$

$$= \text{Softmax}(a) - \text{Softmax}^2(a)$$

$$= \text{Softmax}(a) \{ 1 - \text{Softmax}(a) \}$$

Using this

$$w^* = \underset{w}{\text{argmax}} \left\{ \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{\partial}{\partial w} (\ln(y_{nk})) \right.$$

$$\Rightarrow \tilde{w}^* = \underset{\tilde{w}}{\operatorname{argmax}} \left[\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \cdot \frac{1}{y_{nk}} \frac{\partial y_{nk}}{\partial \tilde{w}} \right]$$

$$\Rightarrow \tilde{w}^* = \underset{\tilde{w}}{\operatorname{argmax}} \left[\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \cdot \frac{1}{y_{nk}} (1 - y_{nk}) \cdot y_{nk} \right]$$

$$\tilde{w}^* = \underset{\tilde{w}}{\operatorname{argmax}} \left[\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \cdot (1 - y_{nk}) \right]$$

This is $\frac{\partial L}{\partial \tilde{w}}$.

So, as per gradient ascent method, for moving from one point to another we have a rule,

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

So,

$$w^{t+1} = w^t + \eta \left[\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} (1 - y_{nk}) \right]_{w = w^t}$$