HW-2

② 2- Class Support Vector Machine

Let the seperating hyperplane be given by the equation

$\text{sign}\{W^T x^{(i)} + W_0\} = \hat{y}$ for $i \in N$ i.e, there are N samples

Let the labels be $y = \{-1, 1\}$ and W, W_0 are such that

when $W^T x^{(i)} + W_0 < 0 \Rightarrow \hat{y} = -1$
$W^T x^{(i)} + W_0 > 0 \Rightarrow \hat{y} = +1$

$\Rightarrow$ Error occurs when $y^{(i)}(W^T x^{(i)} + W_0) < 0$

Let the loss function be $L(y, \hat{y})$

$$L(y,\hat{y}) = \sum_{i \in N} y^{(i)} \left(\vec{w}^T x^{(i)} + W_0\right)$$

Maximizing $L(y,\hat{y})$ reduces the errors

W^* S.T $L(y,\hat{y})$ is minimized

But we need an initial W & W_0 for this

Note :- $\dfrac{W^T x^{(i)} + W_0}{\|W\|}$ is the distance of the points $x^{(i)}$ from W plane

$\therefore$ Let $y^{(i)}(W^T x^{(i)} + W_0) \geq \mu$ $\quad \mu > 0$

$\Rightarrow$ (Grnd truth $\times$ notion of distance with sign) $\geq$ pure distance measure

Letting the distance to be thresholded to be greater than μ reduces the chances of errors

$$\Rightarrow \max_{w, w_0, \|w\|=1} \mu \quad \left\{ \begin{array}{l} \text{Subject to} \quad y^{(i)}(w^T x^i + w_0) \geq \mu \\ \mu \geq 0 \end{array} \right\}.$$

the contraint $y^{(i)}[w^T x^{(i)} + w_0] \geq \mu$

can be changed to $\dfrac{y^{(i)}[w^T x^{(i)} + w_0]}{\|w\|} \geq \mu$

Since μ is arbitrary

& Let $\|w\| = \dfrac{1}{\mu}$ since arbitrary

$$\Rightarrow \{ y^{(i)} [w^T x^{(i)} + w_0] \geq 1 \}$$

& $\max\limits_{w, w_0, \|w\|=1} \mu$ changes to $\left[\max\limits_{w_0, \vec{w}} \frac{1}{2} \|\vec{w}\|^2 \right\}$

<u>Note</u>:- $\max\limits_{w, w_0} \|w\|$, $\max\limits_{w, w_0} \|w\|^2$ & $\max\limits_{w, w_0} \frac{1}{2}\|w\|^2$

yields the same results.

$\therefore$ Finally

$$\max_{w, w_0} \frac{1}{2}\|w\|^2 \quad \text{S.T} \quad y^{(i)}[w^T x^{(i)} + w_0] \geq 1.$$

by introducing Lagrange multipliers, we can change it into an unconstrained problem

$$L_p = \left[\frac{\|w\|^2}{2} - \sum_{i=1}^{N} \alpha_i \left\{ y^{(i)} [w^T x^{(i)} + w_0) - 1 \right\} \right] \quad \alpha_i \geq 0.$$

minimizing L_p gives us the required weights

$\Rightarrow \nabla_{w^*} L_p = 0$ gives us w^*

$$\nabla_w L_p = w - \nabla \sum_{i=1}^{N} \alpha_i \left\{ y^i (w^T x^{(i)} + w_0) - 1 \right\} = 0$$

$$\Rightarrow w - \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)} = 0$$

$$\Rightarrow \boxed{w^* = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}} \quad — ①$$

$$\nabla_{w_0} L_p = 0 \Rightarrow \nabla_{w_0} \sum_{i=1}^{N} \alpha_i \left(y^{(i)} [w^T x^{(i)} + w_0] - 1 \right) = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{N} \alpha_i y^{(i)} = 0} \quad — ②$$

Plugging ① and ② in primal form yields the dual form

$$L_D = \frac{1}{2} \left\| \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)} \right\|^2 - \sum_{i=1}^{N} \alpha_i \left\{ y^{(i)} \left\{ \sum_{j=1}^{N} \alpha_j y^{(j)} x^{(j)T} x^{i} + w_0 \right\} - 1 \right\}$$

$$= \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i y^{(i)} w_0 - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i y^{(i)} x^{(i)T} \alpha_j y^{(j)} x^{(j)}$$

$$\sum_{i=1}^{N} \alpha_i y^{(i)} = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y^{(i)} w_0 = 0.$$

$$\Rightarrow L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)'} x^{(j)} \right)$$

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)} \right\|^2$$

$$\boxed{L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \left\| \operatorname{diag}(\alpha) . Y . X \right\|^2}$$

$\left\{ \operatorname{diag}(\alpha) : \text{diagonal matrix with diagonal elements as } \alpha \right\}$.

Convex optimizing L_D yields values for α

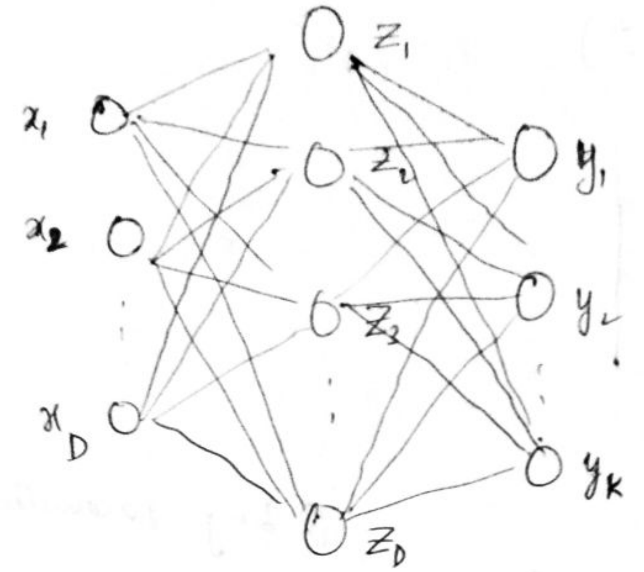
Subst. α in $W = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}$ yields W^*

Additionally from KKT condition

$$\alpha_i \left[y^{(i)} \left(\vec{w}^T x^{(i)} + w_0 \right) - 1 \right] = 0$$

Subst. $\begin{pmatrix} \text{non zero} \\ \alpha_i \end{pmatrix}$ in KKT gives W_0.

(3) Let each data point be of dimension D & number of hidden nodes ~~layer~~ be M. Let output be of dimension K



Each node is connected to every other node in the immediate next layer. Let $A_{11}, A_{12}, \ldots A_{1M}$ be the weights associated with x_1 to $\vec{Z}$ or in general

$\vec{A}_m = \{A_{d1}, A_{d2}, \ldots A_{dM}\}$ be weights from x_d to $\vec{Z}$

in precise A_{dm} be weight from x_d to Z_m. Let its bias be A_{0m}. ~~Bea~~ More like $\{A_{01}, A_{02}, \ldots A_{0M}\}$ be bias' for each hidden node.

Let the activation function be sigmoid.

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

$$Z_{dm} = \sigma\left(A_{m0} + \vec{A}_{dM}^{T} x_d\right)$$

$$\vec{Z}_m = \sigma\left(\vec{A}_{m0} + \vec{A}_M^{T} \vec{x}\right).$$

for the last layer, assigning weights and biases similarly, we ~~g~~ have

$$\hat{y}_k = g_k(\beta_{k0} + \vec{\beta}_k Z)$$

$$\boxed{g_k(\vec{x}) = \frac{e^{x_k}}{\sum_{i=1}^{k} e^{x_i}}}$$

let the ~~lost~~ function be . θ being parameters

$$R(\theta) = \sum_{i=1}^{N} \| \vec{y}^{(i)} - \hat{\vec{y}}(x^{(i)}) \|^2$$

$$R^{(i)}(\theta) = \sum_{k=1}^{K} (y_k^{(i)} - \hat{y}_k(\vec{x}^{(i)}))^2$$

~~find~~ $\frac{\partial R^{(i)}(\theta)}{\partial A_{dm}} = 0$ $\frac{\partial R^{(i)}(\theta)}{\partial \beta_{mk}} = 0$ for locally optimal ~~point~~ params θ

$$\frac{\partial}{\partial \beta_{mk}} \sum (y_k^{(i)} - \hat{y}_k(x^{(i)}))^2 = 0$$

$$\Rightarrow 2(y_k^{(i)} - \hat{y}_k(x^{(i)}))\left(-\frac{\partial}{\partial \beta_{mk}}(\hat{y}_k(x^{(i)})\right) = 0$$

$$\hat{y}_k^{(i)} = g_k(\beta_{k0} + \vec{\beta}_{mk} x^{(i)})$$

$$\Rightarrow \frac{\partial R^{(i)}(\theta)}{\partial \beta_{mk}} = 0 \Rightarrow -2(y_i - \hat{y}_i)\left(g'(\beta_{k0} + \vec{\beta}_{mk} z^{(i)})\right)(z_m^{(i)})$$

$$= \delta_k^{(i)} z_m^{(i)}$$

$$\delta_k^{(i)} = -2(y_i - \hat{y})\left(\frac{\partial}{\partial \beta_{km}} \hat{y}\right)$$

III^{ly} $$\frac{\partial R^{(i)}(\theta)}{\partial A_{dm}} = \frac{\partial}{\partial A_{dm}}\left(\sum_{i=1}^{K}(y-\hat{y})^2\right)$$

$$= -2(y-\hat{y})\left(\frac{\partial}{\partial A_{dm}}(\hat{y})\right)$$

$$= \sum_{k=1}^{K} -2(y_i - \hat{y}_i)\left(g'(\beta_0 + \beta_m Z)\right)\left(\frac{\partial}{\partial A_{dm}} Z\right)(\beta_{km})$$

$$= \sum_{k=1}^{K} \beta_{km}\,\delta_k^{(i)}\,\sigma'\left(\alpha_{mo} + \alpha_m^T x^{(i)}\right) x^{(i)}.$$

$$\therefore \boxed{\frac{\partial R^{(i)}(\theta)}{\partial A_{dm}} = S_m^{(i)}\,\vec{x}^{(i)} \qquad S_m^{(i)} = \sum_{k=1}^{K}\left(\delta_k^{(i)}\beta_{km}\right)\sigma'\left(A_{mo} + A_m^T x^{(i)}\right)}$$

Similar to Newton-Raphson approach, we can find optimal $\vec{\beta}$ and $\vec{A}$ by back propagation with some learning Rate

Learning Rate:- By what fraction of the gradient were correcting our weights and biases

Back propagation :- for each epoch, we tune our weights by sending back information of gradients of cost functions at current datapoint and correct our current weights.

$$\left\{ \beta_{mk}^{r+1} = \beta_{mk}^{r} - (Lr) \sum_{i=1}^{N} \frac{\partial R^{(i)}(\theta)}{\partial \beta_{mk}} \right\}$$

$$\left\{ A_{dm}^{r+1} = A_{dm}^{r} - (Lr) \sum_{i=1}^{N} \frac{\partial R^{(i)}(\theta)}{\partial A_{dm}} \right\}$$

$$\left\{ \begin{array}{l} r = \text{epoch number} \\ Lr = \text{Learning rate} \end{array} \right\}.$$

(4) Given the cost function is crossentropy loss function

$$R^{(i)}(\theta) = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

$$y_{o,c} = \begin{cases} 1 & \text{if } x \in \text{class } c \\ 0 & \text{else} \end{cases}$$

(or) more precisely

for $R_K^{(i)}(\theta)$ i.e., $x^{(i)} \longrightarrow y^{(k)}$ (or) class k

$$R_k^{(i)}(\theta) = -y_k \log\left(Pr\{x^{(i)} \in C_k\}\right)$$

generally $\hat{y} = Pr\{x^{(i)} \in C_k\}$

$$\boxed{R^{(i)}(\theta) = \sum_{k=1}^{K} -y_{i,k} \log(\hat{y}_{i,k})}$$

for optimal solution consider

$$\frac{\partial R^{(i)}(\theta)}{\partial A_{dm}} = \sum_{k=1}^{K} \frac{\partial}{\partial A_{dm}} \left(-y_{i,k} \left(\log\left(\hat{y}_{i,k}\right)\right)\right)$$

$$\frac{\partial R^{(i)}(\theta)}{\partial A_{dm}} = \sum_{k=1}^{K} -\frac{y_{i,k}}{\hat{y}_{i,k}} \cancel{\log} \frac{\partial}{\partial A_{dm}}\left(\hat{y}_{i,k}\right)$$

$$\frac{\partial R^{(i)}(\theta)}{\partial A_{dm}} = \sum_{k=1}^{K} \left(\frac{-y_{i,k}}{\hat{y}_{i,k}}\right)\left(g'\left(\beta_{0k} + \beta_{mk} Z_k\right)\right)\left(\sigma'\left(A_{0d} + A_{dm} x^{(i)}\right)\right) x^{(i)}$$

$$\frac{\partial R^{(i)}(\theta)}{\partial \beta_{mk}} = \sum_{k=1}^{K} \frac{-y_{i,k}}{\hat{y}_{i,k}} g'\left(\beta_{0k} + \beta_{mk} z_k\right) z_k$$

$$\boxed{\frac{\partial R^{(i)}(\theta)}{\partial \beta_{mk}} = \delta_k^{(i)} z_k}$$

$$\boxed{\delta_k^{(i)} = \sum_{k=1}^{K} \frac{-y_{i,k}}{\hat{y}_{i,k}} g'\left(\beta_{0k} + \beta_{mk} z_k\right)}$$

$$\frac{\partial R^{(i)}(\theta)}{\partial A_{dm}} = \left\{\delta_k^{(i)} \cancel{\ } \sigma'\left(A_{0d} + A_{dm} x^{(i)}\right)\right\} x^{(i)} \cdot \beta_{mk}$$

$$\boxed{\frac{\partial R^{(i)}(\theta)}{\partial A_{dm}} = \delta_m^{(i)} x^{(i)}}$$

$$\boxed{\delta_{(m)}^{(i)} = \sum_{k=1}^{K} \frac{-y_{i,k}}{\hat{y}_{i,k}} g'(\beta)\, \sigma'(A)\, \beta_{mk}}$$

① Let the decision hyperplane b/w class j & k be

$$W_k^T x^i + W_{k0} = \hat{y}_k$$

Let $W^T x^j + W_0 = \hat{y}_j > 0$ if $x_i \in j$

$\hat{y}_k > 0$ $\quad x_i \in k$

now consider two points in j

x_1 & x_2

any point x_λ b/w x_1, x_2 is given by

$$x_\lambda = \lambda x_2 + (1-\lambda) x_1 \quad \boxed{\cancel{\lambda \geq 0}} \quad \lambda \in [0,1]$$

$W^T x_\lambda + W_0 < 0$ if ~~the separating hyperplane is convex~~

$$W^T(\lambda x_2 + (1-\lambda) x_1) + W_0$$

$$= \lambda W^T x_2 + W_0 + W^T (1-\lambda) x_1 .$$

Now for j-k boundary, consider.

$$x^{i} \in C_k \quad \text{if} \quad \hat{y}_k(x^{(i)}) - \hat{y}_j(x^{(i)}) > 0$$

$$(W_k - W_j)^T x^{(i)} + (W_{0k} - W_{0j}) > 0.$$

$$y_k(\hat{x}) = \lambda y_k(x_1) + (1-\lambda)\, y_k(x_2)$$

if $x_1, x_2 \in C_k$

$$y_k(x_1),\ y_k(x_2) > 0$$

$$\Rightarrow \lambda y_k(x_1) + (1-\lambda)\, y_k(x_2) > 0$$

$$\Rightarrow \underline{\underline{\hat{x} \in C_k}}$$

$$\Rightarrow \boxed{y_k \text{ is convex}}$$