HW 2 EE5600

Bhavsar Jeel Alpesh
ES16BTECH11005

1) Region R_k is defined as

$$R_k = \{ \bar{x} \mid y_k(\bar{x}) > y_j(\bar{x}) \ \forall j \neq k \}$$

$$y_k(\bar{x}) = w_k^T \bar{x} + w_{k0}$$

~~If R_k is convex,~~

let $\bar{x}_1, \bar{x}_2 \in R_k$

If R_k is convex, $\bar{x}_0 = \lambda \bar{x}_1 + (1-\lambda)\bar{x}_2$ must also belong to R_k $\forall \lambda \in [0,1]$

So $y_k(\bar{x}_0) = y_k(\lambda \bar{x}_1 + (1-\lambda)\bar{x}_2)$

Given $y_k(\bar{x})$ is linear

So $y_k(\bar{x}_0) = \lambda y_k(\bar{x}_1) + (1-\lambda) y_k(\bar{x}_2)$

Now $\lambda y_k(\bar{x}_1) > \lambda y_j(\bar{x}_1)$ — ①

$(1-\lambda) y_k(\bar{x}_2) > \lambda y_j(\bar{x}_2)$ — ②

So adding ① & ②,

$y_k(\bar{x}_0) > y_j(\bar{x}_0)$

Hence $\bar{x}_0$ belongs to R_k

R_k is convex.

2) two class SVM

let the labels $y \in \{1, -1\}$

~~Also, for~~ $\hat{y}(\bar{x}) = \omega^T \bar{x} + \omega_0$

If $\hat{y}(\bar{x}) > 0$, $\bar{x}$ belongs to label 1

$\hat{y}(\bar{x}) < 0$, $\bar{x}$ belongs to label -1

So If $(\hat{y}(\bar{x})) * y < 0$, the its a wrong prediction

Loss function $L(y^{(i)}, \hat{y}^{(i)}(\bar{x})) = y^{(i)}(\bar{\omega}^T \bar{x}^{(i)} + \omega_0) < 0$

For M errors or M miss classifications,

$$L(y, \hat{y}(x)) = \sum_{i \in M} y^{(i)}(\omega^T \bar{x}^{(i)} + \omega_0)$$

We need to find $\bar{\omega}$ that maximizes $L(y, \hat{y}(x))$

$$\max_{\bar{\omega}, \omega_0 \|\bar{\omega}\|=1} u \quad \cancel{\text{subject to}} \; ; u > 0$$

subject to $y^{(i)}(\bar{\omega}^T \bar{x}^{(i)} + \omega_0) \geq u$

$1 \leq i \leq N$

We remove the constraint $\|\bar{\omega}\| = 1$ by

$$\max_{\bar{\omega}, \omega_0} u \quad \text{subject to} \quad \frac{y^{(i)} \cdot (\bar{\omega}^T \bar{x}^{(i)} + \omega_0)}{\|\bar{\omega}\|} \geq u$$

$$= y^{(i)}(\bar{\omega}^T \bar{x}^{(i)} + \omega_0) \geq u \|\bar{\omega}\|$$

$u > 0$

Arbitrarily assume $\|\bar{\omega}\| = \frac{1}{\mu}$

The optimization is equivalent to

$$\min_{\bar{\omega}, \omega_0} \frac{1}{2}\|\omega\|^2 \text{ such that } y^{(i)}(\omega^T x^{(i)} + \omega_0) \geq 1 \quad 1 \leq i \leq N$$

Converting it into unconstrained problem,

$$L_p = \min_{\omega_0, \bar{\omega}} \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{N} \alpha_i \left[y^{(i)}(\bar{\omega}\, \bar{x}^{(i)} + \omega_0) - 1 \right] \quad (1)$$

$\nabla_\omega L_p = 0$ ~~gives~~

~~i.e.~~ $\cancel{\nabla_\omega L_p}$ ~~to~~ $\frac{\partial L_p}{\partial \omega_0} = 0$

$$\cancel{\omega} - \sum_{i=1}^{N} \alpha_i y^{(i)} = 0 \quad (2)$$

$$\frac{\partial L_p}{\partial \bar{\omega}} = 0$$

$$\bar{\omega} = \sum_{i=1}^{N} \alpha_i x^{(i)} y^{(i)} \quad (3)$$

We cannot solve (2) & (3) without values of α

Plug (3) & (2) in (1),

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

Scanned by CamScanner

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

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

s.t. $\alpha_i \geq 0$; $\sum_{i=1}^{N} \alpha_i y^{(i)} = 0$ (4)

In addition to constraints in (4), optimal α_i must satisfy $\alpha_i \left[y^{(i)} (\bar{\omega}^T x^{(i)} + \omega_0) - 1 \right] = 0$

If $\alpha_i = 0$,

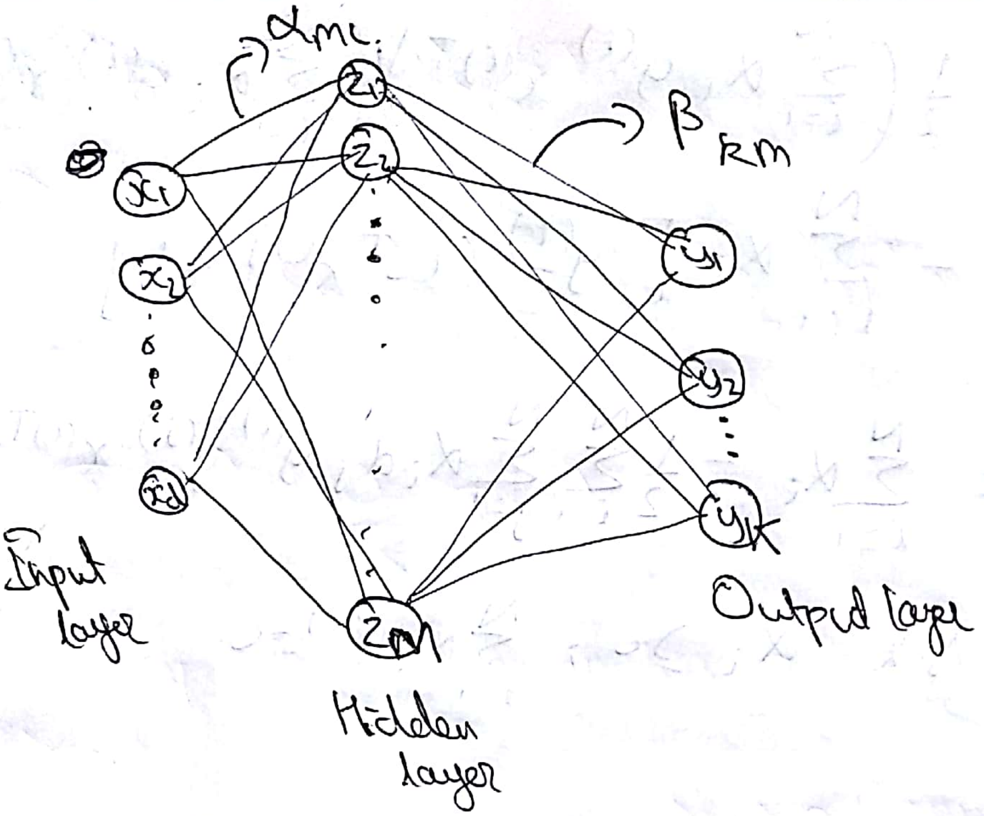
$x^{(i)}$ cannot help us find ω

If $\alpha_i > 0$,

$y^{(i)} (\omega^T x^{(i)} + \omega_0) - 1 = 0$

$x^{(i)}$ will determine the separating hyperplane

3)



$$z_m = \sigma(\alpha_{m0} + \bar{\alpha}_m^T \bar{x}) \qquad 1 \le m \le M$$

Sigmoid func^n: $\sigma(x) = \dfrac{1}{1+e^{-x}}$

$\alpha_m = [\alpha_{m1} \; . \; . \; . \; \alpha_{md}]^T$

α_{m0} : bias associated with m^{th} hidden node

$$\hat{y}_k(\bar{x}) = g_k(\beta_{k0} + \bar{\beta}_k^T \bar{z})$$

where $g_k(\bar{x}) = \dfrac{e^{x_k}}{\sum_{j=1}^{K} e^{x_j}}$ (softmax)

Generally,
$g(x)$ is softmax for classification
and Cross entropy for regression.

We are assuming $g_k(x)$ to be sigmoid function in this problem

Parameters $[\theta : \alpha_{m0}, \bar{\alpha}_m, \beta_{k0}, \bar{\beta}_k]$

$$1 \le m \le M$$
$$1 \le k \le K$$
$$\bar{\alpha}_m \in \mathbb{R}^d, \quad \beta_k \in \mathbb{R}^M$$

Cost func^n: $$R(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} \left(y_k^{(i)} - g_{(k)}(x^{(i)}) \right)^2$$

$$= \sum_{i=1}^{N} R^{(i)}(\theta)$$

$$R^{(i)}(\theta) = \sum_{k=1}^{K} \left(y_k^{(i)} - g_k(x^{(i)}) \right)^2$$

~~For finding optimal~~

For finding locally optimal parameters θ,

$$\frac{\partial R^{(i)}(\theta)}{\partial \beta_{km}} = \frac{\partial}{\partial \beta_{km}} \sum_{k'=1}^{K} \left(y_{k'}^{(i)} - g_{k'}^{(i)} \right)^2$$

$$= \frac{\partial}{\partial \beta_{km}} \sum_{k'=1}^{K} \left(y_{k'}^{(i)} - g_{k'}(\beta_{k'0} + \bar{\beta}_{k'}^T z^{(i)}) \right)^2$$

$$\frac{\partial R^{(i)}(\theta)}{\partial \beta_{km}} = \underbrace{2\left(y_k^{(i)} - g_k(\beta_{k0} + \bar{\beta}_k^T z^{(i)}) \right)\left(-g_k'(\beta_{k0} + \beta_k^T z^{(i)}) \right)}_{\delta_k^{(i)}} z_m^{(i)}$$

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

$$\frac{\partial R^{(i)}(\theta)}{\partial \alpha_{ml}} = S_m^{(i)} x_l^{(i)}$$

where $S_m^{(i)} = \left(\sum_{k=1}^{K} \delta_k^{(i)} \beta_{km}\right) \sigma'\left(\alpha_{m0} + \alpha_m^T \bar{x}^{(i)}\right)$

$$\sigma'(x) = \frac{\partial}{\partial x} \sigma(x)$$

Now update the parameters with gradient descent.

$$\beta_{km}^{(r+1)} = \beta_{km}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R^{(i)}(\theta)}{\partial \beta_{km}^{(r)}}$$

$$\alpha_{ml}^{(r+1)} = \alpha_{ml}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R^{(i)}(\theta)}{\partial \alpha_{ml}^{(r)}}$$

$\gamma_r \rightarrow$ learning rate

The convergence of α & β largely depends on the learning rate γ_r.

Subsequently we can find $\bar{z}$ & then $\bar{y}$

4) For cross entropy loss function,

$$R(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{N} y_j^{(i)} \log\left(\hat{y}_j(x^{(i)})\right)$$

Follow the same procedure as in Question 3 Solution until

$$R(\theta) = \sum_{i=1}^{N} R^{(i)}(\theta)$$

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

$$\frac{\partial R^{(i)}(\theta)}{\partial \beta_{km}} = -\sum_{k'=1}^{K} y_{k'}^{(i)} \frac{\partial}{\partial \beta_{km}}\left(\log \hat{y}_{k'}(x^{(i)})\right)$$

$$= -\sum_{k'=1}^{K} y_{k'}^{(i)} \frac{1}{\hat{y}_{k'}(x^{(i)})} g'(\beta_{k'0} + \bar{\beta}_{k'}^T z) z_m^{(i)}$$

$$= \underbrace{-y_k^{(i)} \frac{1}{\hat{y}_k(x^{(i)})} g'(\beta_{k0} + \bar{\beta}_k^T z)}_{\delta_k^{(i)}} z_m^{(i)}$$

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

Similarly we find

$$\frac{\partial R^{(i)}(\theta)}{\partial \alpha_{ml}} = \delta_m^{(i)} x_l^{(i)} = \left(\sum_{k=1}^{K} \delta_k^{(i)} \beta_{km}\right) \sigma'(\alpha_{m0} + \alpha_m^T x^{(i)})$$

Now update β & α with gradient descent And subsequently find z & $\hat{y}$ as shown in Ans 3.