Assignment - 4

AI20BTECH11006

1)

i) $E_D(w) = \frac{1}{2}\sum_{n=1}^{N} g_n(t_n - w^T\phi(x_n))^2$

let $\Theta = \begin{bmatrix} \sqrt{g_1} \\ \sqrt{g_2} \\ \vdots \\ \sqrt{g_N} \end{bmatrix}$ $Z = \begin{bmatrix} \phi(x_1^T) \\ \phi(x_2^T) \\ \vdots \\ \phi(x_N^T) \end{bmatrix}$ $Y = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$

$E_D(w) = \frac{1}{2}(\Theta^T(Y - Zw))^2$

this is a scalar, but for ease in calculation we will write it as

$= \frac{1}{2}(\Theta^T(Y-Zw))^T(\Theta^T(Y-Zw))$

$= Y^T\Theta\Theta^TY - Y^T\Theta\Theta^TZw - w^TZ^T\Theta\Theta^TY + w^TZ^T\Theta\Theta^TZw$

$= Y^T\Theta\Theta^TY - 2Y^T\Theta\Theta^TZw + w^TZ^T\Theta\Theta^TZw$

$\nabla E_D(w) = -2Y^T\Theta\Theta^TZ + 2w^TZ^T\Theta\Theta^TZ$

$= w^T(Z^T\Theta\Theta^TZ) = Y^T\Theta\Theta^TZ$

$w^T = Y^T\Theta\Theta^TZ(Z^T\Theta\Theta^TZ)^\dagger$

where $\dagger$ represents the pseudoinverse

$w = (Z^T\Theta\Theta^TZ)^\dagger Z^T\Theta\Theta^TY$

$\because$ it is symmetric

$w^* = (Z^T\Theta\Theta^TZ)^\dagger Z^T\Theta\Theta^TY$

b)
i) data dependent noise variance

$$z_i \sim wx_i + N(0, \sigma_i^2)$$

$$z_i \sim N(wx_i, \sigma_i^2)$$

Now, maximising log likelihood
(Minimizing negative log likelihood)

$$= \underset{\theta}{\text{argmax}} \ \log(L(\theta | x))$$

$$= \underset{\mu, \sigma_i^2}{\text{argmax}} \ \log\left(\prod P_{\mu, \sigma_i^2}(x_i)\right)$$

$$= \underset{\mu, \sigma_i^2}{\text{argmax}} \sum_{i=1}^{n} \underbrace{\left[-\frac{1}{2}\log(2\pi\sigma_i^2) - \frac{(x_i - \mu)^2}{2\sigma_i^2}\right]}_{g_i(\mu, \sigma_i^2)}$$

from slides

b) i) $t_i | x_i \sim N(x_i^T w, \sigma_i^2)$

$$L = \prod_i \exp - \frac{1}{2\sigma_i^2}(x_i^T w - y_i)^2 = \exp\left(-\frac{1}{2\sigma^2}\sum(x_i^T w - y_i)^2\right.$$

$$\log L = -\sum \frac{1}{2\sigma_i^2}(x_i^T w - y_i)^2$$

we have to minimize $-\log L$

$$\text{argmin} \sum \frac{1}{2\sigma^2}(x_i^T w - y_i)^2$$

This is similar to part a

let $\sigma^2 = 1/g$

& $g_i = g \ \forall i$

it becomes $\frac{1}{2}\sum g (x_i^T w - y_i)^2$

ii) it can be thought of as a dataset where x_i appears g_i times, so it gets summed up g_i times and thus it becomes weight sum

2) Bayes Optimal Classifier

$P(h_i \mid D)$	$P(F \mid h_i)$	$P(L \mid h_i)$	$P(R \mid h_i)$
0.4	1	0	0
0.2	0	1	0
0.1	0	0	1
0.1	0	1	0
0.2	0	1	0

$$P(F \mid D) = \sum P(F \mid h_i)\, P(h_i \mid D)$$
$$= 0.4$$

$$P(L \mid D) = \sum P(L \mid h_i)\, P(h_i \mid D)$$
$$= 0.5$$

$$P(R \mid D) = \sum P(R \mid h_i)\, P(h_i \mid D)$$
$$= 0.1$$

Bayes optimal classifies returns <u>left</u>

MAP estimate

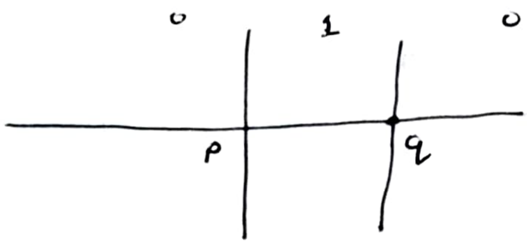
$$\underset{h \in H}{\operatorname{argmax}}\ P(h/D)$$

$P(F \mid h_1) = 0.4$ which is clearly max

$\therefore$ it recommends forward

$\therefore$ they are not the same.

3) VC dimension



for 3 dimension

let $x_1 < x_2 < x_3$

& labels 1, 0, 1 respc

the $x_1 \in (p, q)$

$x_3 \in (p, q)$

$\therefore X_2 \in (p, q)$ which is a contradiction

$\therefore$ VC dimension should be less than 3

Now for 2 dimension

$x_1 \to 0 \leftarrow x_2$

$x_1, x_2 \leq P$ or

$x_1, x_2 \geq q$

$x_1 \to 1, x_2 > x_1, x_2 \to 0$

$x_1 \in (p, q)$

$x_2 \geq q$

$x_1 \to 1$ $x_2 \to 1$

$x_1, x_2 \in (p, q)$

$\therefore$ VC dimension of H is 2

4) Given

$$y(x, \omega) = \omega_0 + \sum_{k=1}^{D} \omega_k x_k$$

$$\Rightarrow y(x_i, \omega) = \omega_0 + \sum_{k=1}^{D} \omega_k x_{ik}$$

$$E(\omega) = \frac{1}{2} \sum_{i=1}^{N} \left(y(x_i, \omega) - t_i\right)^2$$

Adding noise

$$x_{ik} \rightarrow x_{ik} + \varepsilon_{ik} \quad , \quad \varepsilon_{ik} \sim N(0, \sigma^2) \quad \forall i \ \forall k$$

$$y'(x_i, \omega) = \omega_0 + \sum_{k=1}^{D} \omega_k (x_{ik} + \varepsilon_{ik})$$

$$= y(x_i, \omega) + \underbrace{\sum_{k=1}^{D} \omega_k \varepsilon_{ik}}_{= \Sigma_i}$$

Now

$$E'(\omega) = \frac{1}{2} \sum_{i=1}^{N} \left(y'(x_i, \omega) - t_i\right)^2$$

$$= \frac{1}{2} \sum_{i=1}^{N} \left(y(x_i, \omega) + \Sigma_i - t_i\right)^2$$

$$= \frac{1}{2} \sum_{i=1}^{N} \left(y(x_i, \omega) - t_i\right)^2 + \sum_{i=1}^{N} \Sigma_i \left(y(x_i, \omega) - t\right) + \sum_{i=1}^{N} \Sigma_i^2$$

Finding expectation / average

$$E\left[E'(\omega)\right] = E\left[E_\omega + \sum_{i=1}^{N} \Sigma \left(y(x_i, \omega) - t\right) + \sum_{i=1}^{N} \Sigma_i^2\right]$$

Now $E(\Sigma_i) = \sum_{k=1}^{D} \omega_k \Sigma \varepsilon_{ik} = 0$

since mean is 0

$$E\left(\sum_{i=1}^{N} \Sigma_i^2\right), \quad \Sigma_i^2 = \left(\sum_{k=1}^{D} \omega_k \varepsilon_{ik}\right)^2 = \sum_{k=1}^{D} \sum_{j=1}^{D} \omega_k \omega_j \varepsilon_{ik} \varepsilon_{ij}$$

Now $E\left[\varepsilon_{ik} \varepsilon_{ij}\right] = 0 \quad \forall k \neq j$

& $E\left[\varepsilon_{ik} \varepsilon_{ij}\right] = \sigma^2 \quad \forall k = j$ $\quad \because \sigma^2 =$ variance

$$E[E_D'(\omega)] = E_D(\omega) + \frac{N}{2}\sigma^2(\|\omega\|^2 - \omega_0^2)$$

which is similar to l_2 normalization

$$\lambda = \frac{N}{2}\sigma^2$$

we can get L_2 regularisation term without bias parameter ω_0