FoML Assignment - 4

Naitik Malav
CS19BTECH11026

Ans 1. (a)

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

→ So taking gradient of $E_D(w)$ and set it to zero (For minimum)

$$\therefore \frac{\partial}{\partial w} E_D(w) = \underbrace{-\sum_{n=1}^{N} g_n \left(t_n - w^T\phi(x_n)\right)\phi(x_n) = 0}$$ — (1)

∴ From eqn (1) →

$$\sum_{n=1}^{N} g_n t_n \phi(x_n) = \left(\sum_{n=1}^{N} g_n \phi(x_n)\phi(x_n)^T\right) w$$

$$\boxed{\therefore\ w = \frac{\left(\sum_{n=1}^{N} g_n t_n \phi(x_n)\right)}{\left(\sum_{n=1}^{N} g_n \phi(x_n)\phi(x_n)^T\right)}}$$

as

Ans 1. (b) The linear model for the output eqn →

$$\hat{y}_i = w^T x_i + \hat{w}_i$$

where, $\hat{w}_i$ is $\mathcal{N}(0, \sigma_i^2)$

∴ i.e. Normalize the distribution with 0 and σ_i^2 as parameters then making $\sigma_i^2 = 1/2g_i$ we can get the original one.

~~We ea~~

It can alse be derived from making each i^{th} element g_i times so that the $J(w)$ derived.

Hence the above ~~eees~~ are the two alternative interpretations of weighted sum of sq. error function in terms of data-dependent noise variance and replicated data points.

Ans2. MAP estimate :-

$$i > \arg\max_{h_i} P\left(\frac{h_i}{D}\right) \Rightarrow h_1 \quad \left(\because P\left(\frac{h_1}{D}\right) = 0.4\right)$$

also, $P\left(\frac{F}{h_1}\right) = 1 \Rightarrow$ this implies that

MAP estimate is "Forward".

Bayes Optimal Estimate :-

$$P\left(\frac{K=F}{D}\right) = \sum_{h_i} P\left(\frac{F}{h_i}\right) P\left(\frac{h_i}{D}\right) = (1) * (0.4)$$
$$= 0.4 \quad \text{--- (1)}$$

$$P\left(\frac{K=L}{D}\right) = \sum_{h_i} P\left(\frac{L}{h_i}\right) P\left(\frac{h_i}{D}\right) = (0.2)(1) + (0.1)(1) + (0.2)(1)$$
$$= 0.5 \quad \text{--- (2)}$$

$$P\left(\frac{K=R}{D}\right) = \sum_{h_i} P\left(\frac{R}{h_i}\right) P\left(\frac{h_i}{D}\right) = 0.1 \quad \text{--- (3)}$$

$\therefore$ From eq^n (1), (2), (3) i.e. Bayes Optimal Estimate is Left.

as

$\therefore$ MAP Estimate is Forward (F)

and Bayes Optimal Estate is Left (L)

So they are not same, because MAP chooses the most probable hypothesis function then evaluate function of predicted data.

i.e. $\rightarrow$ $h_{MAP} = \arg\max (P(D|h))\, P(h)$

Forward (F) $\Rightarrow$ $\arg\max (0.4 \times 1) = 0.4$

Left (L) $\Rightarrow$ $\arg\max (0.2 \times 1, 0.1 \times 1, 0.2 \times 1)$

$\Rightarrow$ $\underline{0.2}$

Right (R) $\Rightarrow$ $\arg\max (0.1 \times 1) = 0.1$

$\longrightarrow$ Bayes Optimal Classifier makes most probable prediction using <u>entire training data</u> for <u>making predictions</u>.
It also uses <u>hypothetical Space</u> for making predictions.

Ans 3. Let two data points y_1 and y_2 and $y_1 < y_2$.

So they can always be shattered by H, no matter how they are labeled.

Explanation →

Case-a : if y_1 is positive, and y_2 negative then we'll choose $a < y_1 < b < y_2$.

Case-b : if y_1 is negative, and y_2 is positive, then choose → $y_1 < a < y_2 < b$.

Case-c : if y_1 is positive and y_2 is also positive then choose → $a < y_1 < y_2 < b$.

Case d if y_1 is negative & y_2 is also negative then choose → $a < b < y_1 < y_2$.

Now, if we have three points $y_1 < y_2 < y_3$ and if they are labeled as y_1 as positive, y_2 as negative, and y_3 as positive.

Then in above scenario they cannot be shattered by H.

$\boxed{\text{Hence } VC(H) = 2}$ as

Ans 4. Given: $y(x, w) = w_0 + \sum_{K=1}^{D} w_k x_k$

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

$\therefore$ According to question, independent noise is added to each dimension of each input variable x_i

Hence our new model becomes $\rightarrow$

$$y'(x_i, w) = w_0 + \sum_{K=1}^{D} w_k (x_{ik} + \epsilon_{ik})$$

$$y'(x_i, w) = w_0 + \underbrace{\sum_{K=1}^{D} w_k x_{ik} + \sum_{K=1}^{D} w_k \epsilon_{ik}}$$

(here ϵ_{ik} is independent of i & K)

Hence our new error function is $\rightarrow$

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

$$= \frac{1}{2} \sum_{i=1}^{N} \left(y(x_i, w) + \sum_{K=1}^{D} w_k \epsilon_{ik} - t_i \right)^2$$

$$= \frac{1}{2} \sum_{i=1}^{N} \left[(y(x_i, w) - t_i)^2 + 2 (y(x_i, w) - t_i) \left(\sum_{K=1}^{D} w_k \epsilon_{ik} \right) + \left(\sum_{k=1}^{D} w_k \epsilon_{ik} \right)^2 \right]$$

Taking the expection of above eq^n and using linearity of expectation $\rightarrow$ $\because E[\epsilon_{ik}] = 0$

$$E\left[E'_D(w)\right] = \frac{1}{2}\sum_{i=1}^{N}\left((y(x_i,w)-t_i)^2 + \cancel{2\,(y(x_i,w)-t_i) * \left(\sum_{k=1}^{D} w_k E[\epsilon_{ik}]\right)}^{\,0} + E\left[\left(\sum_{k=1}^{D} w_k \epsilon_{ik}\right)^2\right] \right)$$

Now,

$$E\left[\left(\sum_{k=1}^{D} w_k \epsilon_{ik}\right)^2\right] = E\left[\sum_{k=1}^{D}\sum_{k'=1}^{D} w_k w_{k'} \epsilon_{ik}\epsilon_{ik'}\right]$$

$$= \sum_{k=1}^{D}\sum_{k'=1}^{D} w_k w_{k'} E\left[\epsilon_{ik}\epsilon_{ik'}\right]$$

$$= \sum_{k=1}^{D}\sum_{k'=1}^{D} w_k w_{k'} \delta_{kk'}$$

$$= \sum_{k=1}^{D} w_k^2 \qquad \text{— (1)}$$

Using result of eq^n (1) we get $\rightarrow$

$$E\left[E_D'(w)\right] = \frac{1}{2}\sum_{i=1}^{N}\left[\left(y(x_i,w)-t_i\right)^2 + \sum_{k=1}^{D} w_k^2\right]$$

$$= \underbrace{E_D(w) + \frac{N}{2}\sum_{k=1}^{D} w_k^2}$$

Hence here we get a L_2 (norm) regularization term without the bias parameter w_0.

Hence Proved.