Assignment 4

1. $$E_D(\underline{w}) = \frac{1}{2}\sum_{n=1}^{N} g_n\left(t_n - \underline{w}^T\phi(\underline{x}_n)\right)^2$$

Consider $G = \begin{bmatrix} \sqrt{g_1} \\ \sqrt{g_2} \\ \vdots \\ \sqrt{g_n} \end{bmatrix}$ (possible $\because g_i > 0, \forall i$)

$$Y = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}, \quad X = \begin{bmatrix} (\phi(\underline{x}_1))^T \\ (\phi(\underline{x}_2))^T \\ \vdots \\ (\phi(\underline{x}_n))^T \end{bmatrix}$$

a) We can rewrite $E_D(\underline{w})$ as

$$E_D(\underline{w}) = \frac{1}{2}\| G^T(Y - X\underline{w})\|^2$$

as convex, & so, we can differentiate wrt $\underline{w}$, to get minima

$$\frac{\partial E_D(\underline{w})}{\partial \underline{w}} = \frac{\partial}{\partial \underline{w}}\left[\frac{1}{2}\| G^T(Y - X\underline{w})\|^2\right]$$

$$\|G^T(Y - X\underline{w})\|^2 = \left[G^T(Y - X\underline{w})\right]^T \left[G^T(Y - X\underline{w})\right]$$

$$= (Y - X\underline{w})^T GG^T (Y - X\underline{w})$$

$$= Y^T GG^T Y - 2\underline{w}^T X^T GG^T Y + \underline{w}^T X^T GG^T X\underline{w}$$

$$\therefore \frac{\partial E_D(\underline{w})}{\partial \underline{w}} = \frac{\partial}{\partial \underline{w}}\left[\frac{1}{2}\left(Y^T GG^T Y - 2\underline{w}^T X^T GG^T Y + \underline{w}^T X^T GG^T X\underline{w}\right)\right]$$

$$= \frac{\partial}{\partial \underline{w}}\left\{\frac{1}{2}\left(-2X\right)\right.$$

$$\frac{\partial E_D(\underline{w})}{\partial \underline{w}} = 0 - X^T GG^T Y + X^T GG^T X\underline{w}$$

So, @ minima, $\frac{\partial E_D(\underline{w})}{\partial \underline{w}} = 0$

$$\therefore X^T GG^T X\underline{w}^* = X^T GG^T Y$$

$$\therefore \underline{w}^* = \left(X^T GG^T X\right)^{\dagger} X^T GG^T Y$$

b) (i) data-dependent noise variance:

$$t_i \sim w \cdot x_i + N(0, \sigma_i^2)$$

$$\sim N(w \cdot x_i, \sigma_i^2)$$

Maximising log-likelihood (= Minimising negative log-likelihood)

$$= \underset{\theta}{\text{argmax}} \log\left(L(\theta/x)\right)$$

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

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

Maximising $\mu \Rightarrow \mu = \frac{1}{n}\sum x_i$

$\sigma_i^2 \Rightarrow \sigma_i^2 = \frac{1}{n}\sum (x_i - \mu)^2$

$$\therefore \text{Log } L(\underline{w}/s) = \sum \log P(t_i/w)$$

$$= \sum_{i=1}^{n} \frac{(\underline{w}x_i - t_i)^2}{2\sigma_i^2}$$

By setting $\frac{1}{\sigma_i^2} = g_i$, we get the given expression.

(ii) repeated data points

If our data is repeated, we can simply add a term 'qi', to the cost, & x_i's will be unique now.
qi corresponds to frequency of occurrence of every x_i.

2) Bayes Optimal Classification:

$$P(F) = \sum P(F/h_i) P(h_i/D)$$

$$= 1(0.4) + 0(0.2) + 0(0.1) + 0(0.1) + 0(0.2)$$
$$= 0.4$$

$$P(L) = \sum P(L/h_i) P(h_i/D)$$

$$= 0(0.4) + 1(0.2) + 0(0.1) + 1(0.1) + 1(0.2)$$
$$= 0.2 + 0.1 + 0.2 = 0.5$$

$$P(R) = \sum P(R/h_i) P(h_i/D)$$

$$= 0(0.4) + 0(0.2) + 1(0.1) + 0(0.1) + 0(0.2)$$
$$= 0.1$$

∴ Bayes optimal classifier recommends to turn Left. (as P(L) is maximum).

MAP estimate:

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}}\ P(h/D) \quad \text{(definition)}$$

Clearly, from definition, & data given MAP recommends to go forward.

(∵ $P(h_i/D)$ is max for h_1 & $P(F/h_1) = 1$)

Hence, they are NOT SAME

3) VC-dimension: Model class F has VC dimension d if d is the largest set of points $x_1, \cdots, x_d \subset X$, $\exists$ for all labelings of $x_1, \cdots, x_d$, $\exists$ some $f \in F$ that achieves that labelling.

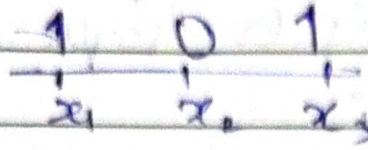
Given, 1- dimensional data.
Model class $\vdash\vdash$, characterised by $\{p, q\}$

Consider 
case of 3 points.

label 1: if $p < x < q$

label 0: else.

Clearly, for the following configuration



this class of models H, can't achieve the labellings.

$\because x_1$ labelled as $1 \Rightarrow p < x_1 < q$

$\because x_3$ " " $1 \Rightarrow p < x_3 < q$

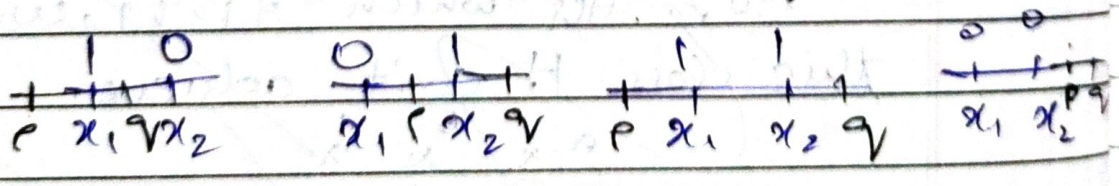
Also $x_1 < x_2 < x_3$

$\Rightarrow p < x_2 < q$.

Contradiction, as x_2 is labelled 0. i.e; $x_2 > q$ (or) $x_2 < p$.

$\therefore$ VC dimension is not 3 or more.

for 2 points.



for all combinations, H can achieve the labellings,

$\therefore$ VC dimension = 2

4) $y(\underline{x}, \underline{w}) = w_0 + \sum_{k=1}^{D} w_k x_k$

$$E(\underline{w}) = \frac{1}{2} \sum_{i=1}^{N} \left(y(\underline{x}_i, \underline{w}) - t_i \right)^2$$

If Gaussian noise $\epsilon_k \sim N(0, \sigma^2)$ is added to each of i/p variable x_k, then

$$y'(\underline{x}, \underline{w}) = w_0 + \sum_{k=1}^{D} w_k (x_k + \epsilon_k)$$

i.e, $y'(\underline{x}_i, \underline{w}) = w_0 + \sum_{k=1}^{D} w_k (x_{ik} + \epsilon_{ik})$,

$$= y(\underline{x}_i, \underline{w}) + \sum_{k=1}^{D} w_k \epsilon_{ik}$$

Also,

$$E_D'(\underline{w}) = \frac{1}{2} \sum_{i=1}^{N} \left\{ \left(y(\underline{x}_i, \underline{w}) - t_i \right)^2 + 2 \left(y(\underline{x}_i, \underline{w}) - t_i \right) \left(\sum_{k=1}^{D} w_k \epsilon_{ik} \right) + \left(\sum_{k=1}^{D} w_k \epsilon_{ik} \right)^2 \right\}$$

Applying expectation

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

Assuming

$$E[\varepsilon_i] = 0 \quad \& \quad E[\varepsilon_i \varepsilon_{i'}] = \begin{cases} \sigma^2, & i \neq i' \\ 0, & i = i' \end{cases}$$

we get,

$$E\left[\left(\sum \omega_k \varepsilon_{ik}\right)^2\right] = E\left[\sum_{k=1}^{D} \sum_{k'=1}^{D} \omega_k \omega_{k'} \varepsilon_{ik} \varepsilon_{ik'}\right]$$

$$= \sum_{k=1}^{D} \sum_{k'=1}^{D} \omega_k \omega_{k'} \, \mathbf{E}\left[\varepsilon_{ik} \varepsilon_{ik'}\right]$$

$$= \sum_{k=1}^{D} \omega_k^2 \sigma^2$$

$$= \sigma^2 \sum_{k=1}^{D} \omega_k^2$$

Hence,

$$E\left[\tilde{E}(\underline{w})\right] = E(\underline{w}) + \frac{N}{2}\sum_{k=1}^{D} w_k^2$$

∴ we could get L_2 regularization term without bias parameter w_0.