1. Non - Uniform Weights in Linear Regression

a) Given Error function,

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

To minimize the above error function, we go for equating the derivative to zero (1)

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

$$\Rightarrow \sum_{n=1}^{N} g_n \left[t_n - w^T \phi(x_n)\right] \cdot \phi(x_n)^T = 0$$

$$\Rightarrow \sum_{n=1}^{N} g_n t_n \phi(x_n)^T - w^T \left(\sum_{n=1}^{N} g_n \phi(x_n) \cdot \phi(x_n)^T\right) = 0$$

$$\Rightarrow \sum_{n=1}^{N} \sqrt{g_n} \phi(x_n)^T \cdot \sqrt{g_n} t_n - w^T \left(\sum_{n=1}^{N} \sqrt{g_n} \phi(x_n) \cdot \sqrt{g_n} \phi(x_n)^T\right) = 0$$

Let $\sqrt{g_n} \phi(x_n)^T = \phi'(x_n)^T$; $\sqrt{g_n} t_n = t_n'$

$$\Rightarrow \sum_{n=1}^{N} t_n' \cdot \phi'(x_n)^T - w^T \left(\sum_{n=1}^{N} \phi'(x_n) \cdot \phi'(x_n)^T\right) = 0$$

Solving the above,

$$\boxed{w = \left[\phi^T \cdot \phi\right]^{-1} \cdot \phi^T \cdot t}$$ where, $t = \begin{bmatrix} \sqrt{g_1} t_1 \\ \sqrt{g_2} t_2 \\ \vdots \\ \sqrt{g_n} t_n \end{bmatrix}$

& $\phi(i,j) = \sqrt{g_i} \cdot \phi_j(x_i)$

(b)

If the linear model is given as $W^T x_i + \epsilon_i$

where $\epsilon_i \sim N(0, \sigma_i^2)$, Then the error function would be

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

(i) The error function $E_D(w) = \frac{1}{2} \sum_{n=1}^{N} g_n (t_n - w^T \phi(x_n))^2$ can be obtained if $\boxed{\sigma_n^2 = 1/g_n}$

(ii) The error function in point (i) can also be obtained by ~~creating~~ treating (x_n, t_n) as repeatedly occuring g_n times. or creating g_n copies of the nth data point

2 D

2. Bayes Optimal classifier

	$P(h_i/D)$	$P(F/h_i)$	$P(L/h_i)$	$P(R/h_i)$
h_1	0.4	1	0	0
h_2	0.2	0	1	0
h_3	0.1	0	0	1
h_4	0.1	0	1	0
h_5	0.2	0	1	0

Maximum A Posteriori (MAP) Hypothesis is defined as

$$\boxed{h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \; P(h/D)}$$

From the table ~~by maxim~~, the 'h_i' that maximises $P(h/D)$ is h_1. (i.e) max of $P(h_i/D) = 0.4$ which occurs for h_1. Also, $P(F/h_1) = 1$ which is a sure event. Therefore, the MAP hypothesis suggests the robot should go forward (F).

The Bayes Optimal classification is defined as

$$\boxed{h_{BO} = \underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j/h_i) P(h_i/D)}$$

V is the set of all possible classifications

$$\sum_{h_i \in H} P(F|h_i) \cdot P(h_i|D) = P(F|h_1) \cdot P(h_1/D) + P(F/h_2) \cdot P(h_2|D) + \ldots P(F|h_5) \cdot P(h_5 D)$$

$$= (1 \times 0 \cdot 4) + (0 . 0) + \ldots + (0 . 0)$$

$$\boxed{\therefore \sum_{h_i \in H} P(F|h_i) \cdot P(h_i/D) = 0 \cdot 4} \rightarrow ①$$

$$\sum_{h_i \in H} P(L|h_i) \cdot P(h_i/D) = P(L|h_1) \cdot P(h_1/D) + P(L|h_2) \cdot P(h_2/D) + P(L|h_3) \cdot P(h_3|D) + P(L/h_4) \cdot P(h_4/D) + P(L|h_5) \cdot P(h_5/D$$

$$= (0 \times 0 \cdot 4) + (1 \times 0 . 2) + (0 \times 0 \cdot 1) + (1 \times 0 \cdot 1) + (0 . 2 \times 1)$$

$$\boxed{\sum_{h_i \in H} P(L/h_i) \cdot P(h_i/D) = 0 \cdot 5} \rightarrow ②$$

$|||^{ly}$

$$\sum_{h_i \in H} P(R/h_i) \cdot P(h_i|D) = 1 \times 0 \cdot 1 = 0 \cdot 1 \rightarrow ③$$

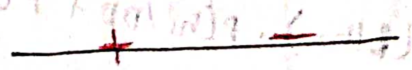
$\therefore \sum_{h_i \in H} P(L|h_i) P(h_i|D) = 0 \cdot 5$ is the Largest. Thus, the Bayes optimal recommends <u>the robot turn left</u>

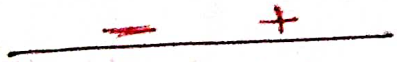
Hence MAP estimate and Bayes optimal estimate are the NOT the <u>same</u>.

3.) VC - Dimension :

Let us consider two points in the one dimensional space.

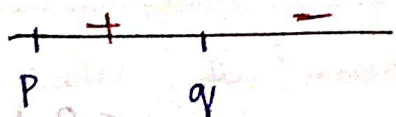
<u>Possible</u> <u>combinations</u> :

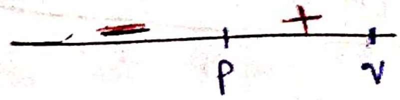




In both of the above configurations, we can classify the positive with an interval such that $p < x < q$. where x is the positive class point.

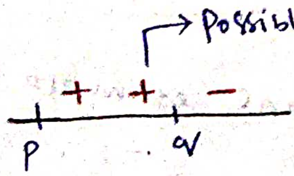
It is as shown below.

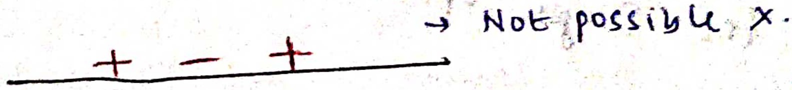




Thus $p < x < q$ <u>could classify all possible labeling</u> <u>for two points classification.</u>

Let us consider three points configuration as shown below:





For the above configurations, it is not possible for ~~atleast~~ $p < x < q$ to classify atleast one configuration.

∴ The VC dimension cannot be 3.

Hence, the VC dimension is 2 //

7) Given

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

with noise is added independently to each of the input variables x_k. Then

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

The error function is,

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

$$= \frac{1}{2} \sum_{i=1}^{N} \left(w_0 + \sum_{k=1}^{D} w_k (x_{ik} + \epsilon_k) - t_i \right)^2$$

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

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

Consider,

$$E_\epsilon \left[\left(\sum_{k=1}^{D} w_k \epsilon_k \right)^2 \right] = E_k \left[\sum_{j=1}^{D} \sum_{k=1}^{D} w_j w_k \epsilon_j \epsilon_k \right]$$

$$= \sum_{j=1}^{D} \sum_{k=1}^{D} w_j w_k E_\epsilon [\epsilon_j \epsilon_k]$$

$$= \sum_{j=1}^{D} w_j^2 E_\epsilon [\epsilon_j^2] + \sum_{j=1}^{D} \sum_{j \neq k}^{D} E_\epsilon (\epsilon_j \epsilon_k) \longrightarrow (1)$$

We know that

$E(\epsilon_j^2) = \sigma^2$ since ϵ_* is $N(0, \sigma^2)$

Since ϵ_j, ϵ_k are independent

$$E_\epsilon[\epsilon_j \epsilon_k] = E_\epsilon(\epsilon_j) E(\epsilon_k) = 0$$

$$\therefore E_\epsilon\left[\left(\sum_{k=1}^{D} w_k \epsilon_k\right)^2\right] = \sigma^2 \sum_{j=1}^{D} w_j^2$$

Consider

$$E_\epsilon\left[2\left(\sum_{k=1}^{D} w_k \epsilon_k\right)\left(y(x_i, w) - t_i\right)\right]$$

$$= 2\left(y(x_i,w) - t_i\right) E_\epsilon\left[\sum_{k=1}^{D} w_k \epsilon_k\right]$$

$$= 2\left(y(x_i,w) - t_i\right) \sum_{k=1}^{D} w_k E_\epsilon[\epsilon_k]$$

$$\Rightarrow E_\epsilon\left[2\left(\sum_{k=1}^{D} w_k \epsilon_k\right)\left(y(x_i,w) - t_i\right)\right] = 0$$

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

Consider error for noise free input variables with L2 weight decay.

$$E_D(w) = \frac{1}{2}\sum_{i=1}^{N}\left(y(x_i,w) - t_i\right)^2 + \lambda \sum_{k=1}^{D} w_k^2 \longrightarrow ③$$

Hence from ② & ③,

Minimizing the sum of squares for noise free input variables with L2 weight decay is equivalent to minimizing E_D averaged over the noise distribution