1 Given,
data points (x_n, t_n) $n = 1 \dots N$
weighing factor $g_n > 0$
error function
$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N} g_n\left(t_n - w^T\phi(x_n)\right)^2$$
$\phi(\cdot)$ is any representation of data.

a) we have to find an expression for the solution w^* that minimizes the above error function -
we can take gradient and set to zero.
$$\frac{\partial}{\partial w} E_D(w) = -\sum_{n=1}^{N} g_n\left(t_n - w^T\phi(x_n)\right)\phi(x_n) = 0$$

Now,
let solve for w:-
$$\sum_{n=1}^{N} g_n t_n \phi(x_n) = \left(\sum_{n=1}^{N} g_n \phi(x_n)\phi(x_n)^T\right) w$$
$$w = \left(\sum_{n=1}^{N} g_n \phi(x_n)\phi(x_n)^T\right)^{-1}\left(\sum_{n=1}^{N} g_n t_n \phi(x_n)\right)$$
$$w = \frac{\sum_{n=1}^{N} g_n t_n \phi(x_n)}{\sum_{n=1}^{N} g_n \phi(x_n)\phi(x_n)^T}$$

b) let us assume a linear model of the output
$y_i = w^T x_i + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the noise.

i) For data dependent noise variance, the above objective function (E_D) can be derived by minimizing the negative log-likelihood of the output if we set $\sigma^2 = \frac{1}{g_i}$ or,
$$= \frac{1}{2g_i}$$

ii) For replicated data points, the above objective function (E_D) can be derived if we create g_i copies of the i^{th} data point.

2 Bayes Optimal Estimate $\qquad y = \{F, L, R\}$

$$= \arg\max \sum_{h_i} P(y_j | h_i)\, P(h_i | D)$$

$$= \arg\max \left(\overset{F}{\sum}(0.4 * 1),\ \overset{L}{\sum}(0.2 * 1 + 0.1 * 1 + 0.2 * 1),\ \overset{R}{\sum}(0.1 * 1) \right)$$

$$= \arg\max\ (0.4,\ 0.5,\ 0.1)$$

$$= \arg\ (0.5)$$

$$= L$$

So according to Bayes Optimal Estimator for a new data instance, the most probable prediction is Left (L) move by the robot.

MAP estimate

$$= \arg\max_{\theta} P(\theta | D)$$

$$= \arg\max\ (P(h_1|D),\ P(h_2|D),\ P(h_3|D), \ldots, P(h_5|D))$$

$$= \arg\max\ (0.4,\ 0.2,\ 0.1,\ 0.1,\ 0.2)$$

$$= \arg\ (0.4)$$

$$= h_1$$

According to MAP estimator, h_1 ~~hypotheses~~ is the most probable hypothesis that describes the training dataset.

Hence, MAP estimate $\neq$ Bayes Optimal Estimate.

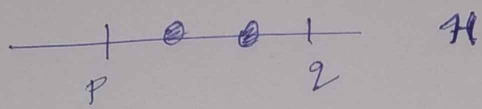
$\boxed{3}$ R^1 - 1D data

we can represent the data in a number line.

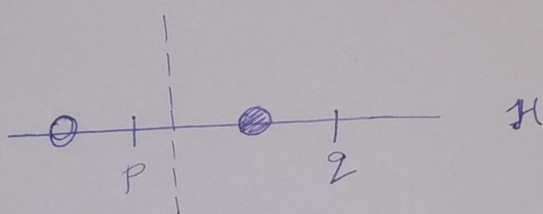
let ● denote class 1 when $p < x < q$
○ denote class 0 when $x <= p$ or, $x \geq q$.

let's take two points and place them, such that they are correctly classified.

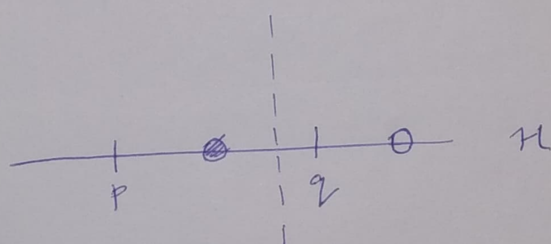
case 1: Both points are between p and q



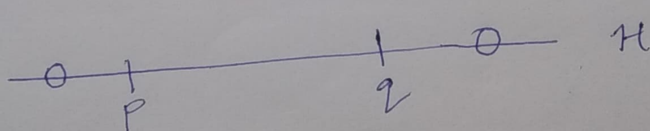
case 2: One point $<= p$ and the other betwen p & q



case 3: one point $>= q$ and other between p & q



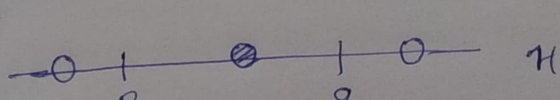
case 4: One point $<= p$ and the other $>= q$



Hence, $\boxed{VC(\mathcal{H}) = 2}$

Also, for three points we can prove that the classifier cannot separate all points. For ex -
(for atleast one labelling order)

$VC(\mathcal{H}) \neq 3$
proved.



4

Noisy data

$\hat{x}_i = x_i + \varepsilon_i$

$\varepsilon \sim \mathcal{N}(0, \sigma^2)$

$\hat{y}(x_i, W) = y(x_i, W) + \sum_{k=1}^{D} w_k \varepsilon_k$

$\hat{y}(x_i, W) = y(x_i, W) + W^T \varepsilon$

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

$\frac{\partial}{\partial w} \hat{E}(w) = \sum_{i=1}^{N} \left(\hat{y}(x_i, W) - t_i \right) (x + \varepsilon)$

setting $\frac{\partial}{\partial w} \hat{E}(w)$ to 0 for minimizing,

$\sum_{i=1}^{N} \left(\hat{y}(x_i, W) - t_i \right) (x + \varepsilon) = 0 \quad — ①$

Noise-free data

similar to above, we can directly write it in derivative form —

$\frac{\partial}{\partial w} E(w) = \sum_{i=1}^{N} \left(y(x_i, W) - t_i \right) x$

$\sum_{i=1}^{N} \left(y(x_i, W) - t_i \right) (x) = 0 \quad — ②$

let us substract ② from ① to find a relation —

$\sum_{i=1}^{N} \left[\left(\hat{y}(x_i, W) - t_i \right) (x + \varepsilon) - \left(y(x_i, W) - t_i \right) (x) \right] = 0$

$\Rightarrow \sum_{i=1}^{N} \left[\left(y(x_i, W) + w^T \varepsilon - t_i \right) (x + \varepsilon) - \left(y(x_i, W) - t_i \right) (x) \right] = 0$

$\Rightarrow \sum_{i=1}^{N} \left[x w^T \varepsilon + y(x_i, W) \varepsilon + \underset{\downarrow \; L_2 \text{ norm}}{w^T \varepsilon^2} - \varepsilon t_i \right] = 0$

$$\Rightarrow \sum_{i=1}^{N} (y(x_i, w) - t_i)\, \varepsilon = -x w^T \varepsilon - w^T \varepsilon^2$$

$$\Rightarrow \sum_{i=1}^{N} (y(x_i, w) - t_i) = -(x w^T + w^T \varepsilon)$$

$$\boxed{= -w^T(x + \varepsilon)}$$

or,

$$\boxed{= -\sum_{k=1}^{D} w_k (x_k + \varepsilon_k)} \quad — (3)$$

(3)

Putting this result into

(1), we get -

$$\sum_{i=1}^{N} (\hat{y}(x_i, w) - t_i)(x + \varepsilon) = 0$$

or,

$$\sum_{i=1}^{N} (y(x_i, w) + w^T \varepsilon - t_i)(x + \varepsilon) = 0$$

or $\cancel{\sum}\,(w^T \varepsilon - w^T(x + \varepsilon))(x + \varepsilon) = 0$

or $\cancel{\sum}\,(x + \varepsilon) w^T(-x) = 0$

$$\boxed{\text{or} \quad -\cancel{\sum}\, x(x + \varepsilon) w^T = 0}$$

Similarly in (2), we get

$$-w^T(x + \varepsilon)(x) = 0$$

$$\boxed{\text{or,} \quad -x(x + \varepsilon) w^T = 0}$$

Hence, both (1) and (2) are equal, on substituting the result we found in (3).