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

a) Make $\nabla E_D(w) = 0$ for getting w^*

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

lets denote $\sqrt{g_n}\,\phi(x_n) = \phi'(x_n)$ & $\sqrt{g_n}\, t_n = t'_n$.

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

$$\Rightarrow \sum_{n=1}^{N} \phi'(x_n)^T t'_n - \sum_{n=1}^{N} w^T * \phi'(x_n) * \phi'(x_n)^T = 0.$$

$$\Rightarrow \sum_{n=1}^{N} \phi'(x_n)^T t'_n = w^T \Big(\sum_{n=1}^{N} \phi'(x_n) * \phi'(x_n)^T\Big)$$

From this we can simply derive a result for w^* which minimizes the above error function.

$$\boxed{w^* = (\phi^T\phi)^{-1} * \phi^T . t}$$

But t is defined as:

$$t = [\sqrt{g_1}t_1, \sqrt{g_2}t_2, \ldots \sqrt{g_N}t_N]^T$$

Lets define ϕ as $N*M$ matrix, with element $\phi(i,j) = \sqrt{g_i}\,\phi_j(x_i)$

1b. Interpretation:-

(i) Data dependent noise variance:-

Considering a gaussian noise model, let us assume target variable t is given by deterministic function $y(x,w)$ with additive gaussian noise so that $t = y(x,w) + \epsilon$ where ϵ is a zero mean Gaussian RV with precision (inverse variance) β.

$$p(t/x, w, \beta) = \mathcal{N}(t/y(x,w), \beta^{-1})$$

$$E(t/x) = \int t\, p(t/x)\, dt = y(x,w)$$

$$p(t/x, w, \beta) = \prod_{i=1}^{n} \mathcal{N}(t_n / w^T\phi(x_n), \beta^{-1})$$

Now considering logarithm of likelihood.

$$\ln P(t/x, w, \beta) = \sum_{n=1}^{N} \ln \left(\mathcal{N}(t_n / (w^T \phi(x_n), \beta^{-1}) \right)$$

$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \beta E_D(w)$$

Here sum of squares error is given by

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

Consider gradient of log of likelihood

$$\nabla \ln P(t/w, \beta) = \sum_{n=1}^{N} \{ t_n - w^T \phi(x_n) \} \phi(x_n)^T$$

Setting gradient to zero,

$$0 = \sum_{n=1}^{N} t_n (\phi(x_n))^T - w^T \sum_{n=1}^{N} \phi(x_n) \phi(x_n)^T$$

$$\Rightarrow \boxed{w = (\phi^T \phi)^{-1} \phi^T t} \longrightarrow ②$$

(ii) From given question, g_n can also be viewed as. effective no. of observation; (x_n, t_n) i.e (x_n, t_n) can be treated as repeatedly occurring r_n times (replication of datapoints)

Scanned by Scanner Go

(2)

Bayes estimate.

$$\sum_{h_i \in H} P(F/h_i)\, P(h_i/D) = 0.4.$$

$$\sum_{h_i \in H} P(L/h_i)\, P(h_i/D) = 0.2 + 0.1 + 0.2 = 0.5$$

$$\sum_{h_i \in H} P(R/h_i)\, P(h_i/D) = 0.1.$$

Thus bayes optimal recommend the robot turn left

MAP Hypothesis is defined as follows.

$$h_{MAP} = \operatorname{argmax}_{h \in H} P(h_i/D).$$

$\Rightarrow$ ~~= argmax~~

$P(h_1/D) = 0.4$; $P(h_2/D) = 0.2$; $P(h_3/D) = 0.1$; $P(h_4/D) = 0.1$

$P(h_5/D) = 0.2$.

$\Rightarrow$ Max values occurs at hypothesis 1 which is h_1

$\Rightarrow$ in $P(h_1/D) \Rightarrow P(F/h_1) = 1,\ P(L/h_1) = 0,\ P(R/h_1) = 0$

$\Rightarrow$ The robot should go forward

3. 1. Given is one dimensional data $\in R^2$. The parameters are $\{p, q\}$ where x is classified as 1

Iff $p < x < q$ then the vc dimensions of H is given by

2. (a) Let us suppose that the training points are in sphere of radius R. Then let $G(x) = \text{sign}[f(x)] = \text{sign}[\alpha^T x + \beta_0]$

(b) So, the class of functions $\{G(x) \; \|B\| \leq A\}$, it has VC dimension h satisfying $\boxed{h \leq R^2 A^2}$

Hence VC dimension of $H \leq R^2 A^2$

---

4. 1. Given D dimensional data $= (x_1, x_2 \ldots x_D)$ of linear model

$$y(x, w) = w_0 + \sum_{K=1}^{D} w_K x_K$$

2. N such data samples with labels $(x_i, t_i) = i = 1, 2, \ldots N$

the sum of square function is given by $\boxed{\text{correct}}$ ★

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

3. So rearranging:

$$E_D(w) = \frac{1}{2} \sum_{i=1}^{N} \{[w_0 + \sum_{i=1}^{D} w_i (x_i + \varepsilon_i) - t_i\}^2$$

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

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

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

4. where we have used $y(x_i, w)$ to denote the output of the linear model when the input variable is x_i without noise added for the 2nd term equation we have.

$$E_\varepsilon[(\sum_{i=1}^{D} w_i \varepsilon_i)^2] = E_\varepsilon[\sum_{i=1}^{D} \sum_{i=1}^{D} w_i w_j \varepsilon_i \varepsilon_j]$$

$$\Rightarrow \sum_{i=1}^{D} \sum_{j=1}^{D} w_i w_j E_\varepsilon[\varepsilon_i \varepsilon_j] = \sigma^2 \sum_{i=1}^{D} \sum_{j=1}^{D} w_i w_j \delta_{ij}$$

which gives

$$E_\varepsilon[(\sum_{i=1}^{D} w_i \varepsilon_i)^2] = \sigma^2 \sum_{i=1}^{D} w_i^2$$

5. for the third term we can obtain:-

$$E_\varepsilon\left[2\left(\sum_{i=1}^{D} w_i\varepsilon_i\right)\left(y(x_i,w)-t_i\right)\right] = 2/y(x_i,w)-t_i)\, E_\varepsilon\left(\sum_{i=1}^{D} w_i\varepsilon_i\right)$$

$$= 2\left(y(x_i,w)-t_i\right)\sum_{i=1}^{D} E_\varepsilon\left[w_i\varepsilon_i\right] = 0$$

Therefore if we calculate the expectation of $E_D(w)$ with respect to ε we can obtain:

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