FoML

Assignment - 4

Hemanth K
ES19BTECH11003

1) Given,

dataset where datapoints are denoted by $(x_n, t_n), n = 1, \dots, N$

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

Let us consider the below Matrices

$$T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}_{n \times 1} \qquad W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}_{d \times 1}$$

$d \to$ dimensions of x_i

$$\phi = \begin{bmatrix} \phi(x_1) \\ \phi(x_2) \\ \vdots \\ \phi(x_n) \end{bmatrix}_{n \times d}, \quad G = \begin{bmatrix} g_1 & \cdots & & \\ \cdot & g_2 & \cdots & \\ \cdots & \cdot & g_3 & \cdots \\ & & \ddots & \\ \cdots & & & g_n \end{bmatrix}_{n \times n}$$

$$E_d(w) = \frac{1}{2} (T^T - w^T \phi^T) G (T - \phi w)$$

$$= \frac{1}{2} (T^T G T - T^T G \phi w - G w^T \phi^T T + w^T \phi^T G \phi w$$

So as to find minimum, we have to differentiate w.r.t to $w = 0$

$$\therefore \frac{\partial (E_d(w))}{\partial w} = \frac{1}{2} (- T^T G \phi^T - \phi^T G T + \phi^T G \phi w + \phi^T G \phi w) \quad (G^T = G)$$

(diagonal Matrix)

$$\therefore \frac{\partial (E_d(w))}{\partial w} = \frac{1}{2} (-2 \times \phi^T G T + \phi^T G \phi w) = -\phi^T G T + \phi^T G \phi w = 0$$

$$\phi^T G T = \phi^T G \phi w$$

$$\phi^T G (T - \phi w) = 0, \quad \boxed{w = (\phi^T G \phi)^{-1} \phi^T G T}$$

b) (i) for data dependent noise variances

Let us consider $t_i = \varepsilon_i + w^T x_i$

$$\varepsilon_i = \mathcal{N}(0, \sigma_i^2)$$

$$t_i = \mathcal{N}(0, \sigma_i^2) + w^T x_i$$

$$= \mathcal{N}(w^T x_i, \sigma_i^2)$$

from max Likelihood,

$$\text{max Likelihood}(W)_{Data} = \arg\max_{w} \prod_{i=1}^{N} P(E_i/w)$$

$$= \arg\max_{w} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\left(-\frac{(t_i - w^T x_i)^2}{2\sigma_i^2}\right)}$$

& for log Likelihood maximum for

$$w = \arg\max_{w} \underbrace{N \log \frac{1}{\sqrt{2\pi\sigma_i^2}}}_{\text{Independent of } w} + \left(-\sum_{i=1}^{N} \frac{(t_i - w^T x_i)^2}{2\sigma_i^2}\right)$$

$$= \arg\max_{w} \sum_{i=1}^{N} \frac{(t_i - w^T x_i)^2}{2\sigma_i^2}$$

Now compare above equation with $E_d(w) = \arg\min_{w} \frac{1}{2}\sum g_i (t_i - w^T x_i)^2$

$$\boxed{\therefore g_i = \frac{1}{\sigma_i^2}}$$

for data dependent noise variances $E_d(w) \cdot \frac{1}{2}\sum \frac{1}{\sigma_i^2}(t_i - w^T \phi(x_i))^2$

(ii) for replicated data points

Let us assume each i^{th} data points gets repeated for g_i times then again interms out as

$$E_d(w) = \sum_{i=1}^{N} g_i \frac{(t_i - w^T x_i)^2}{2} = \frac{1}{2}\sum_{i=1}^{N} g_i (t_i - w^T x_i)^2$$

$$= \frac{1}{2}\sum_{i=1}^{N} g_i (t_i - w^T \phi(x_i))^2$$

$\therefore$ Same error function as $E_d(w) = \frac{1}{2}\sum_{i=1}^{N} g_i (t_i - w^T \phi(x_i))^2$

2) As given, there are 5 hypotheses

Bayes optimal Estimate makes most probable predictions using given space of hypothesis (h)

$$\underset{y}{\text{Arg max}} \text{ of } \sum P(y/h_i) P(h/D)$$

$$\sum P(F/h_i) P(h/D) = 0.4\times1 + 0.2\times0 + 0.1\times0 + 0.1\times0 + 0.2\times0 = 0.4$$

$$\sum P(L/h_i) P(h_i/D) = 0.4\times0 + 0.2\times1 + 0.1\times0 + 0.1\times0 + 0.2\times1 = 0.5$$

$$\sum P(R/h) P(h/D) = 0.4\times0 + 0.2\times0 + 0.1\times1 + 0.1\times0 + 0.2\times0 = 0.1$$

from above 3, maximum is for Left (L) so according to Bayer's optimal Estimate, so robot goes to Left

Map estimate chooses most probable hypothesis and then it starts evaluating data

$$\underset{y}{\text{Arg max}} \text{ of } (P(y/h) P(h/D))$$

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

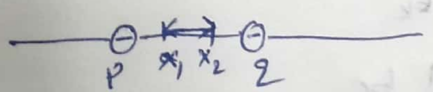
So according to map estimate, robot goes forward, So Left & forward aren't same,

3) Given, to consider, data setup of one-dimensional data $\in \mathbb{R}^1$

Hypothesis space is parameterized by p & q where x is 1 if $p < x < q$

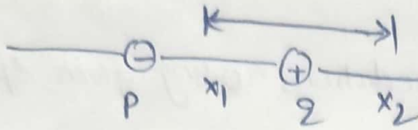
Case - 1

choose x_1, x_2 such that $p < x_1 < x_2 < q$



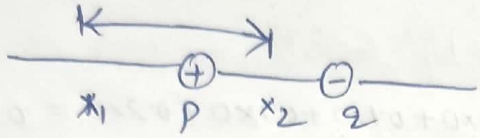
So p, q are -

Case-2 Choose x_1, x_2 such that $p < x_1 < q < x_2$



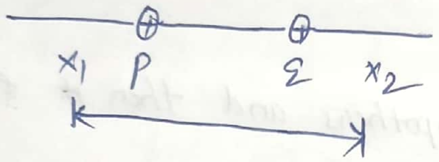
$p \to$ negative
$q \to$ positive

Case-3 Choose x_1, x_2 such that $x_1 < p < x_2 < q$



$p \to$ positive
$q \to$ negative

Case-4 Choose x_1, x_2 such that $x_1 < p < q < x_2$



$p \to$ positive
$q \to$ positive

Here we have clearly shown that using 2 dimension, But when we use 1 dimension it is not possible, with configuration, $-+-/++-/.-+-$

VC dimension of hypothesis = 2

4) Given, D-dimensional data $x = [x_1, x_2, \ldots x_D]$

Linear Model = $y(x,w) = w_0 + \sum_{k=1}^{D} w_k x_k$ & also sum of squared errors $E(w) = \frac{1}{2} \sum_{i=1}^{N} (y(x_i, w) - t_i)^2$

Now Gaussian Noise ϵ_k is being added independently to each of the Input variable x_k $\therefore$ $y'(x,w) = w_0 + \sum_{k=1}^{D} w_k(x_k + \epsilon_k)$

$$= w_0 + \sum_{k=1}^{D} w_k x_k + \sum_{k=1}^{D} w_k \epsilon_k$$

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

Now, error function will be

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

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

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

Lets take Expectation on both sides

$$E(E'(w)) = \frac{1}{2}\sum_{i=1}^{N}\left\{(y(x_i, w) - t_i)^2 + E\left[\left(\sum_{k=1}^{D} w_k \epsilon_k\right)^2\right] + 2(y(x_i, w) - t_i)\underbrace{\left(\sum_{k=1}^{D} w_k E(\epsilon_k)\right)}_{0}\right\}$$

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

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

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

$$E(E'(w)) = E_D(w) + N/2 \sum_{k=1}^{D} w_k^2$$

The relation between sum of squares avg over noise data & the standard sum of square error (nonnoisy data) is $E'_d(w) = E_d(w) + N/2 \sum_{k=1}^{D} w_k^2$