Assignment 4

AI20BTECH11011

Q.1.) a) $E_D(\bar{w}) = \frac{1}{2}\sum_{n=1}^{N} g_n (t_n - w^T \Phi(x_n))^2$

We want to find $\bar{w}$ to minimize $E_D(\bar{w})$

~~Let~~

We can also write $E_D(\bar{w})$ as

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

Let $\bar{y} = \begin{pmatrix} \sqrt{g_1}\, t_1 \\ \sqrt{g_2}\, t_2 \\ \vdots \\ \vdots \\ \sqrt{g_N}\, t_N \end{pmatrix}$

$X = \begin{pmatrix} \sqrt{g_1}\, \Phi^T(x_1) \\ \sqrt{g_2}\, \Phi^T(x_2) \\ \vdots \\ \vdots \\ \sqrt{g_N}\, \Phi^T(x_N) \end{pmatrix}$

Then we can write $E_D(\bar{w})$ as

$$E_D(\bar{w}) = \frac{1}{2}\| X\bar{w} - \bar{y} \|_2^2$$

To minimize this, we find its gradient w.r.t. $\bar{w}$ and set it to zero.

$$\frac{\partial}{\partial \bar{\omega}} E_D(\bar{\omega}) = X^T(\bar{X}\bar{\omega} - \bar{y}) = 0$$

$$\Rightarrow \bar{\omega} = (X^T X)^{\dagger} X^T \bar{y}$$

b: i.) data-dependent noise variance.

We assume that y_i's are linearly related to $\Phi(\bar{x}_i)$'s but with some added noise $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$.

Using Maximum likelihood estimation, we want to maximize,

$$\sum_{i=1}^{N} \log p(y_i | \bar{\omega})$$

$$y_i = \bar{\omega}^T \Phi(\bar{x}_i) + \epsilon_i$$

$$\Rightarrow p(y_i | \bar{\omega}) = p(\epsilon_i | \bar{\omega}).$$

$\Rightarrow$ We have to maximize

$$\sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-\epsilon_i^2}{2\sigma_i^2}\right)\right)$$

$$\propto \sum_{i=1}^{N} -\frac{(y_i - w^T\Phi(x_i))^2}{2\sigma_i^2}$$

$\Rightarrow$ We have to minimize $\displaystyle\sum_{i=1}^{N} \frac{(y_i - w^T\Phi(x_i))^2}{2\sigma_i^2}$.

If $2\sigma_i^2 = \dfrac{1}{g_i}$,

We have to minimize $\sum_{i=1}^{N} g_i (y_i - w^T \Phi(x_i))^2$

Thus the error function given in question is just for data where the noise for each sample comes from $\mathcal{N}\left(0, \frac{1}{2g_i}\right)$.
~~a Gaussi~~

Thus the variance of noise depends on the data sample.

ii.) replicated data-points.

If i^{th} data sample is replicated g_i times, then we get the same error function.

Q.2) Bayes optimal estimate:

Let w_1, w_2 & w_3 be the three partitions of the feature space.

We ~~say~~ estimate
F if $h_i \in w_1$
L if $h_i \in w_2$
R if $h_i \in w_3$.

Thus ~~P(w~~ $w_1 = \{h_1\}$
$w_2 = \{h_2, h_4, h_5\}$
$w_3 = \{h_3\}$.

$P(w_1) = 0.4$
$P(w_3) = 0.1$
$P(w_2) = 0.5$

$P(h_1|w_1) = 1$, $P(h_2 \text{ or } h_3 \text{ or } h_4 \text{ or } h_5 | w_1) = 0$
$P(h_3|w_3) = 1$, $P(h_1 \text{ or } h_2 \text{ or } h_4 \text{ or } h_5 | w_3) = 0$
$P(h_2|w_2) = 2/5$
$P(h_4|w_2) = 1/5$
$P(h_5|w_2) = 2/5$, $P(h_1 \text{ or } h_3 | w_2) = 0$.

Thus $P(h_1|w_1) \cdot P(w_1) = 0.4$
$P(h_1|w_2) \cdot P(w_2) = P(h_1|w_3) \cdot P(w_3) = 0$
$\Rightarrow$ We estimate F for h_1
$P(h_2|w_2) \cdot P(w_2) = 1/5$
$P(h_2|w_1) \cdot P(w_1) = P(h_2|w_3) \cdot P(w_3) = 0$.
$\Rightarrow$ We estimate L for h_2.

Likewise ~~two~~ we can estimate the rest.

~~his estimate~~

hypothesis	estimate
h_1	F
h_2	L
h_3	R
h_4	L
h_5	L

MAP estimate:

$$\underset{x}{\operatorname{argmax}}\ P(x \mid h_1) = F$$

Thus we estimate F for h_1.
Likewise we estimate the rest as

hypothesis	estimate
h_1	F
h_2	L
h_3	R
h_4	L
h_5	L

We observe that indeed, MAP estimate and Bayes optimal estimate are the same.

This is occuring because our system is not probabilistic but deterministic.

For any hypothesis, only one outcome is possible.

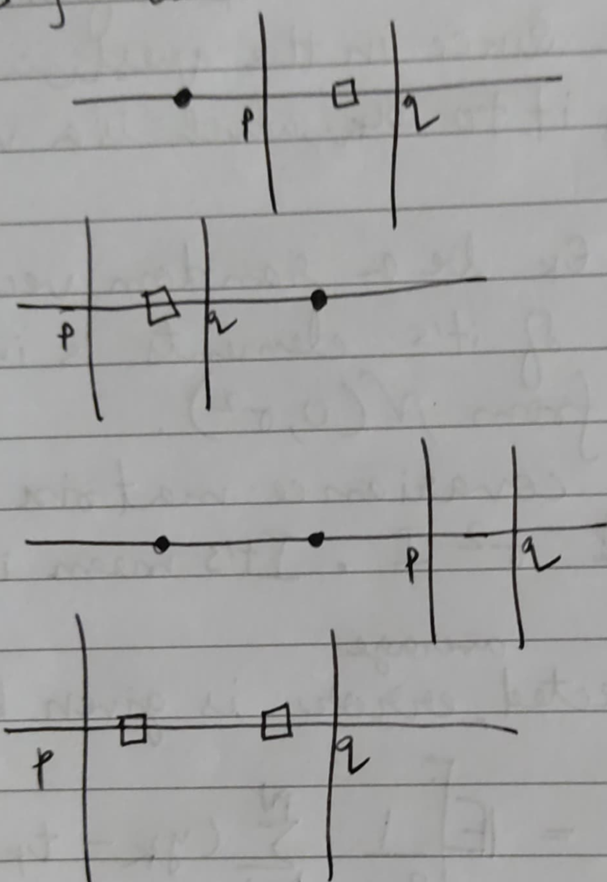
Thus ~~P~~ any hypothesis can belong to only one ~~subspace~~ outcome class.
The ~~set~~ probabilities of that hypothesis belonging to other classes is thus zero.

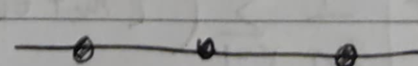
Thus Bayes optimal estimate & MAP estimate are the same.

Q.3) The VC dimension of $\mathcal{H}$ is 2.

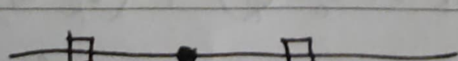
Proof: Let • represent 0 & □ represent 1.

Then A set of 2 points can be cracked as follows:



Any set of 3 points will have the configuration 

For such configuration, the labelling

 can not be cracked.

Thus VC dimension is 2.

Q.4.) If we add ϵ_k to each inputsample, then our predictions now become

$$y_k = (x_k + \epsilon_k)^T \bar{w} + w_0$$

Here, I assume that ϵ_k refers to a noise vector since in the question, we are adding it to x_k, which is a vector.

We let ϵ_k be a random vector where each of it's elements is independent and comes from $\mathcal{N}(0, \sigma^2)$.
Thus, the covariance matrix of ϵ_k is $\sigma^2 I$. It's mean is $\bar{0}$.

The expected average error is given by

$$E(w) = \mathbb{E}\left[\frac{1}{2n} \sum_{k=1}^{N} (y_k - t_k)^2\right]$$

$$\Rightarrow E(w) = \mathbb{E}\left[\frac{1}{2n} \sum_{k=1}^{N} \left((x_k + \epsilon_k)^T \bar{w} + w_0 - t_k\right)^2\right]$$

$$= \mathbb{E}\left[\frac{1}{2n} \sum_{k=1}^{N} (x_k^T \bar{w} + w_0 - t_k)^2 + (\epsilon_k^T \bar{w})^2 + 2\epsilon_k^T \bar{w}(x_k^T \bar{w} + w_0 - t_k)\right]$$

Using Linearity of expectation,

$$E(w) = \mathbb{E}\left[\frac{1}{2n}\sum_{k=1}^{N}(x_k^T\bar{w}+w_0-t_k)^2\right] + \mathbb{E}\left[\frac{\sum_{k=1}^{N} 2\epsilon_k^T\bar{w}(x_k^T\bar{w}+w_0-t_k)}{2n}\right] + \mathbb{E}\left[\frac{\sum_{k=1}^{N}(\epsilon_k^T\bar{w})^2}{2n}\right]$$

$$= \frac{1}{2n}\sum_{k=1}^{N}(x_k^T\bar{w}+w_0-t_k)^2 + \sum_{k=1}^{N}\frac{2\bar{w}^T\cancel{(x_k^T\bar{w}+w_0-t_k)}}{2n}\mathbb{E}(\epsilon_k)(x_k^T\bar{w}+w_0-t_k) + \left\{\frac{\sum_{k=1}^{N}\bar{w}^T\mathbb{E}(\epsilon_k^T\epsilon_k)\bar{w}}{2n}\right\}$$

$$= \frac{1}{2n}\sum_{k=1}^{N}(x_k^T\bar{w}+w_0-t_k)^2 + 0 + \frac{\sigma^2\bar{w}^T\bar{w}}{2}$$

$(\because \mathbb{E}(\epsilon_k^T\epsilon_k) = \sigma^2 I)$

But the error for noise-free ~~with~~ ~~regularizer~~ data is given by

$$E(\bar{w}) = \frac{1}{2n}\sum_{k=1}^{N}(x_k^T\bar{w}+w_0-t_k)^2 + \lambda\|w\|^2$$

where $\lambda\|w\|^2$ is the L_2 regularizer.

Thus if $\alpha = \frac{\sigma^2}{2}$, we are minimizing the same error function in both the cases.

In other words, adding noise to data in case of linear regressor is similar to regularizing the data.