Question-1

* Linear Regression *

K. Surya Prakash
Tatipelly Vamshi

Ans: Considering linear model

$y(x,w) = w_0 + \sum w_i x_i$ for $i = 1$ to D

$\therefore$ $D \to$ features

Error: $E_D(w) = \frac{1}{2} \sum_{n=1}^{N} \{y_n(x_n, w) - t_n\}^2$

Note: This Soln is formulated in terms of matrix / vector calculations.

→ Denotions:

Data → $\underset{N \times (D+1)}{\overline{X}} = \begin{bmatrix} \overline{x}^1 \\ \vdots \\ \overline{x}^N \end{bmatrix}$

where $\overline{x}^n \to$ row vector

$\overline{x}^n = [\,x_0^n \;\; x_1^n \; \cdots \; x_{D+1}^n\,]$

Here, $\boxed{x_0^n = 1 \quad \forall \; n = 1 \text{ to } N}$

$\overline{x}^n = [\,1, x_1^n \; \cdots \; x_{D+1}^n\,]$

Here, for every data point $\bar{x}^n$, $\bar{\varepsilon}^n$ is added → which is Gaussian distributed

$$\bar{\hat{x}}^n = \bar{x}^n + \bar{\varepsilon}^n$$

Such that

$$\bar{\varepsilon}^n = [\varepsilon_0^n \; \varepsilon_1^n \dots \varepsilon_D^n]$$

Here, $\boxed{E(\varepsilon_i^n] = 0}$

$$E(\varepsilon_i \varepsilon_j) = \sigma^2 \delta_{ij}$$

⇒ <u>NOTE:</u>

Since ~~no~~, here $\bar{\varepsilon}^n_{(D+1)}$ is a row vector, where, it has (D+1) elements, Since no error is added to $\bar{x}_0^n$, ⇒ $\boxed{\varepsilon_0^n = 0}$

$$\boxed{\varepsilon_0^n = 0}$$

↪ Each element among other 'D' ⇒ $i \in [1, D]$

ε_i are generated from a guassian distribution of mean '0' and variance: σ^2.

$\therefore\ E(\varepsilon_i) = 0 \quad \forall\ i$

$\Rightarrow\ E(\varepsilon_i \varepsilon_j) = \sigma^2 \delta(i,j)\ ;\quad i, j \in \{1, D\}$

$E(\varepsilon_0 \varepsilon_j) = 0 \Rightarrow \because \boxed{\varepsilon_0 = 0}$

$$\Rightarrow\ \text{cov}(\varepsilon) = \sigma^2 \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ 0 & & 1 & & 0 \\ 0 & \cdots & & 1 & 0 \\ 0 & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & 1 \end{bmatrix}_{(D+1)\times(D+1)}$$

Note: First row of $\text{cov}(\varepsilon) = \bar{0}$

(a) Minimising E_D averaged over the noise distribution.

$$\Rightarrow\ J(w) = \underset{\varepsilon_i}{E}\left(\left(\frac{1}{2}\right) \sum_{n=1}^{N} \{y_n\{\bar{x}_n, w\} - t_n\}^2\right)$$

$$\Rightarrow\quad y_n\{\bar{x}_n, w\} = \underset{1\times(D+1)}{\bar{x}^n} \cdot \underset{(D\times 1),1)}{\bar{w}}$$

$t_n \rightarrow$ ground truth.

Matrix form

$$W = \begin{bmatrix} w_0 \\ \vdots \\ w_D \end{bmatrix}_{(D+1, 1)} \rightarrow \text{weight vector}$$

$$J(w) = E_{\varepsilon}\left(\frac{1}{2}\|\hat{x}w - t\|^2\right)$$

Here $t = \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}_{N\times 1}$ → ground truth.

Here $\hat{x}$ is the noised data

$$\underset{(N, D+1)}{\hat{x}} = \bar{x} + \bar{\varepsilon}$$

$$\Rightarrow J(w) = E_{\varepsilon_i}\left(\frac{1}{2}\|(x+\varepsilon)w - t\|^2\right)$$

We need to minimise $J(w)$.

Gradient for kth feature

$$\Rightarrow \frac{\partial J(w)}{\partial w_k} = \frac{\partial}{\partial w_k} E_{\varepsilon_i}\left(\frac{1}{2}((x+\varepsilon)w - t)^T((x+\varepsilon)w - t)\right)$$

$$\frac{\partial J(w)}{\partial w} = E_{\varepsilon}\left((x+\varepsilon)^T((x+\varepsilon)w - t)\right)$$

$$\left(\because \frac{\partial(\|A\|^2)}{\partial A} = \frac{\partial A^TA}{\partial A} = 2A\right)$$

$$\Rightarrow \frac{\partial J}{\partial W} = E_{\varepsilon}\left((X+\varepsilon)^T(X+\varepsilon)W - (X+\varepsilon)^T t\right)$$

$$= E_{\varepsilon}\left((X^TX + \varepsilon^TX + X^T\varepsilon + \varepsilon^T\varepsilon)W - X^Tt - \varepsilon^Tt\right)$$

$$\Rightarrow \quad = \left(X^TX + E_{\varepsilon}(\varepsilon^T)X + X^TE_{\varepsilon}(\varepsilon) + E_{\varepsilon}(\varepsilon^T\varepsilon)\right)W - pX^Tt - E(\varepsilon^Tt)$$

$\because$ X, t are independent of ε

$$\Rightarrow \quad E(X\varepsilon) = X E(\varepsilon)$$

$$\frac{\partial J}{\partial W} = \left(X^TX + 0 + 0 + \mathrm{Cov}(\varepsilon)\right)W - X^Tt - 0$$

$$\boxed{\frac{\partial J}{\partial W} = (X^TX + \mathrm{Cov}(\varepsilon))W - X^Tt}$$

To minimise J, $\frac{\partial J}{\partial W} = 0$

$$\Rightarrow (X^TX + \mathrm{cov}(\varepsilon))W = X^Tt$$

$$\boxed{W = (X^TX + \mathrm{cov}(\varepsilon))^{-1}(X^Tt)} \rightarrow (1)$$

$\hookrightarrow$ Result -1.

Part b :

Minimising sum of squares error for noise free variables with weight decay regularisation

ie, $$J(w) = \frac{1}{2}\sum_{n=1}^{N}\{y_n(x_n, w) - t_n\}^2 + \lambda w^T w$$

$$J(w) = \underbrace{\frac{1}{2}\| Xw - t\|^2}_{\text{noise-free data}} + \underbrace{\frac{\lambda}{2}\|\tilde{w}\|^2}_{\text{regularisation}}$$

Here, $w = \begin{bmatrix} w_0 \\ \vdots \\ w_D \end{bmatrix}_{(D+1, 1)}$

Note The regularisation has a omitted w_0

ie, $\tilde{w} = \begin{bmatrix} 0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}$

$$J(w) = \frac{1}{2}\|Xw - t\|^2 + \frac{\lambda}{2}\|\tilde{w}\|^2$$

$$\frac{\partial J(\cdot)}{\partial w} = \left(\frac{1}{2}\right)(2)\left((Xw-t)^T(X)\right) + \frac{\lambda}{2}(2(w^T)\tilde{I}) = 0$$

Note $\frac{\partial \|\tilde{w}\|^2}{\partial w} = 2w^T\tilde{I}$, $\tilde{I} = \begin{bmatrix} 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ & & \ddots & \\ & & & 1 \end{bmatrix}$

$(D+1 \times D+1)$

→ First row of $\tilde{I} = \bar{0}$

$$\Rightarrow \frac{\partial J}{\partial w} = (Xw-t)^T X + \lambda(w^T\tilde{I})$$

$\Rightarrow$ To minimise, $\frac{\partial J}{\partial w} = 0$

$$\Rightarrow (Xw-t)^T X + \lambda(w^T\tilde{I}) = 0$$

$$\Rightarrow X^T(Xw-t) + \lambda(\tilde{I}w) = 0$$

$$\Rightarrow (X^TX + \lambda\tilde{I})w = X^Tt$$

$$\Rightarrow \boxed{w = (X^TX + \lambda\tilde{I})^{-1}(X^Tt)} \rightarrow 2$$

Conclusion:

From ①:

$$W = (X^T X + \mathrm{Cov}(\varepsilon))^{-1} (X^T t)$$

From ②:

$$W = (X^T X + \lambda \tilde{I})^{-1} (X^T t).$$

We know that $\mathrm{Cov}(\varepsilon) = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ 0 & 0 & \sigma^2 \cdots & 0 \\ \cdots & & & \sigma^2 \end{bmatrix}_{(D+1 \times D+1)}$

$$\mathrm{Cov}(\varepsilon) = \sigma^2 \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & & & \\ \cdots & & & 1 \end{bmatrix} = \sigma^2 \tilde{I}$$

∴ Both results are the same

where $\boxed{\lambda = \sigma^2}$

— x —