MW-2

S. Sai Ashish
EE16BTECH11043

(1)

$$Loss = E[(y - \hat{y}(x))^2]$$

where $\hat{y}(x)$ is the estimator

$$L = \iint (y(x) - \hat{y}(x))^2 P(x,y)\, dy\, dx$$

$$\frac{\partial L}{\partial \hat{y}(x)} = 0$$

$$\Rightarrow \int_x \left[\int_y 2(y - \hat{y}(x)) P(x,y)\, dy \right] dx = 0$$

$$\Rightarrow \int_y 2(y - \hat{y}(x)) P(y|x) P(x)\, dy = 0$$

$$\Rightarrow \int_y y P(x,y)\, dy = \hat{y}(x) \int P(x,y)\, dy$$

$$\Rightarrow P(x) E(y/x) = \hat{y}(x)[P(x)]$$

$$\Rightarrow \boxed{\hat{y}^*(x) = E[y/x]}$$

Optimal estimator is $E[y/x]$

(2) Bias - Variance

Assuming,

A data set D and $\hat{y}_D(\vec{x})$ is the estimator associated with the data set 'D'

The error is $E\left[(y - \hat{y}_D(\vec{x}))^2\right]$

$$= E\left[\left(y - y^*(\vec{x}) + y^*(\vec{x}) - \hat{y}_D(\vec{x})\right)^2\right]$$

$$= E\left[(y - y^*(\vec{x}))^2\right] + E\left[(y^*(\vec{x}) - \hat{y}_D(\vec{x}))^2\right]$$

$$+ 2\iint_{x\,y} \underbrace{(y - y^*(\vec{x}))\left(y^*(\vec{x}) - \hat{y}_D(\vec{x})\right) P(\vec{x}, y)\, dy\, dx}_{\text{solving}}$$

$$\iint_{x\,y} (y - y^*(\vec{x}))\left(y^*(\vec{x}) - \hat{y}_D(\vec{x})\right) P(x, y)\, dy\, dx =$$

Scanned by CamScanner

$$= \int_x \left(y^*(\vec{x}) - \hat{y}_D(\vec{x})\right)\left[\int_y \left(y - y^*(x)\right) P(y|x)\, dy\right] P(x)\, dx$$

$$= \int_x \left(y^*(\vec{x}) - \hat{y}_D(\vec{x})\right)\left(\cancel{E(y|x)} - \cancel{E(y|x)}\right] P(x)\, dx$$

$$= 0.$$

Thus

$$E\left[\left(y - \hat{y}_D(\vec{x})\right)^2\right] = \underbrace{E\left[\left(y - y^*(\vec{x})\right)^2\right]}_{\text{irreducable noise}} + \underbrace{E\left[\left(y^*(\vec{x}) - \hat{y}_D(\vec{x})\right)^2\right]}_{①}$$

From ①

$$E\left[\left(y^*(\vec{x}) - \hat{y}_D(\vec{x})\right)^2\right] = E\left[\left(\left(y^* - E_D(\hat{y}_D(\vec{x}))\right) + \left(E(\hat{y}_D(x)) - \hat{y}_D(\vec{x})\right)\right)^2\right]$$

$$E\left[\left(y^*(\vec{x}) - \hat{y}_D(\vec{x})\right)^2\right] = E_D\left(\left[\left(\hat{y}_D(\vec{x}) - E(\hat{y}_D(x))\right)\right]^2\right) + E\left(\left(y^* - E_D(\hat{y}_D)\right)\right)^2$$

$$+ 2E\left[\left(y^* - E_D(\hat{y}_D)\right)\underbrace{\cancel{\left(\hat{y}_D - E_D(\hat{y}_D)\right)}}_{0}\right]$$

→ Since

$\left(y^* - E_D(\hat{y}_D)\right] \to$ constant

$$E_D\left[\hat{y}_D - E_D(\hat{y}_D(x))\right] = \underbrace{E_D(\hat{y}_D(\vec{x})) - \cancel{E_D(\hat{y}_D(\vec{x}))}}_{=0}$$

$$E\left[\left(y^*(\vec{x}) - \hat{y}_D(\vec{x})\right)^2\right] = E_D\left[\left(\hat{y}_D(\vec{x}) - E(\hat{y}_D(x))\right)^2\right] + \left[y^* - E_D(\hat{y}_D(\vec{x}))\right]^2$$

Thus

$$\mathbb{E}\left[\left(\hat{y}_D(\vec{x}) - y\right)^2\right] = \underbrace{\left(\left(y^* - E_D(\hat{y}_D)\right)\right)^2}_{\text{bias}^2} + \underbrace{E_D\left(\hat{y}_D(\vec{x}) - E_D(\hat{y}_D(\vec{x}))\right)^2}_{\text{variance}}$$

$$+ \underbrace{E\left[\left(y^* - y\right)^2\right]}_{\text{irreducible noise}}$$

There is a trade off when we reduce bias, variance shoots and vice versa; this is known as bias variance tradeoff

(3) $L = Tr\left[(Y - \tilde{X}\tilde{W})^T (Y - \tilde{X}\tilde{W})\right]$

where

$$\tilde{X} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_1^{(N)} & x_2^{(N)} & \cdots & x_d^{(N)} \end{bmatrix}$$

$$\tilde{W} = \begin{bmatrix} w_{01} & w_{11} & w_{21} & \cdots & w_{d1} \\ w_{02} & w_{12} & w_{22} & \cdots & w_{d2} \\ \vdots & \vdots & \vdots & & \vdots \\ w_{0K} & w_{1K} & w_{2K} & \cdots & w_{dK} \end{bmatrix}^T$$

Scanned by CamScanner

$$L = \sum_{i=1}^{K}\sum_{j=1}^{N} (Y - \tilde{X}\tilde{\omega})^T_{ij} \, (Y - \tilde{X}\tilde{\omega})_{ji}$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{N} (Y - \tilde{X}\tilde{\omega})^2_{ji}$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{N} (Y_{ji} - \tilde{X}\tilde{\omega}_{ji})^2$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{N} \left(-\sum_{k=0}^{d} \tilde{X}_{jk} \cdot w_{ki} + Y_{ji}\right)^2$$

$$\frac{\partial L}{\partial w_{ki}} = -2 \sum_{i=1}^{K}\sum_{j=1}^{N} \left(-\sum_{k=0}^{d} \tilde{X}_{jk} w_{ki} + Y_{ji}\right) \tilde{X}_{jk}$$

$$= -2 \sum_{i=1}^{K}\sum_{j=1}^{N} \tilde{X}^T_{kj} \left[Y_{ji} - \sum_{k=0}^{d} \tilde{X}_{jk} w_{ki} \right] \quad -(1)$$

<u>In general</u>

$$\nabla E(\vec{w}) = \begin{bmatrix} \frac{\partial E}{\partial w_{01}} & \cdots & \frac{\partial E}{\partial w_{d1}} \\ \vdots & & \vdots \\ \frac{\partial E}{\partial w_{0K}} & \cdots & \frac{\partial E}{\partial w_{dK}} \end{bmatrix} \quad -(2)$$

Thus from (1) and (2),

$$\nabla E(\vec{w}) = -2 \tilde{X}^T (Y - \tilde{X}w)$$

Setting $\nabla E(\vec{w}) = 0$

$$\boxed{W^* = \left[(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T Y \right]}$$

This similar to the output for Multidimensional Label, Linear regression Problem (Vanilla)

(4) <u>Fischer's Linear Discriminant</u> :-

$y = (\vec{w})^T \vec{x}$ is

Projection of a(D+1) dimensional vector to one dimension

→ The choice of $\vec{w}$ should be such that intra class variance is minimized and inter class variance is maximized

→ Consider a 2 classes C_1, C_2 whose means are given by

$$\vec{m}_1 = \frac{1}{N_1} \sum_{n \in C_1} \vec{x}_n$$

$$\vec{m}_2 = \frac{1}{N_2} \sum_{n \in C_2} \vec{x}_n$$

We would like to choose a vector $\vec{w}$ such that

$m_2 - m_1 = w^T(\vec{m}_2 - \vec{m}_1)$ is maximized.

This can be done by having arbitrarily large w which is not prefered because it might lead to overfitting

So we constrain w to have $\leq$ unit length

ie, $\sum_i w_i^2 \leq 1$

→ The within class variance is given by

$$s_k^2 = \sum_{y_n \in C_k} (y_n - m_k)^2$$

where

$$y_n = \vec{w}^T \vec{x}$$

$$m_k = \vec{w}^T \vec{m}$$

Fisher Criterion

$$J(\vec{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

Now,

$$(m_2 - m_1) = w^T(\vec{m}_2 - \vec{m}_1) = (\vec{m}_2 - \vec{m}_1)^T w$$

Since its a scalar $\boxed{a^T = a}$ where a is scalar.

$$(m_2 - m_1)^2 = w^T (m_2 - m_1)(m_2 - m_1)^T w$$

where

$$S_B = (m_2 - m_1)(m_2 - m_1)^T = \text{Between class covariance Matrix}$$

$$s_1^2 + s_2^2 = \sum_{n \in C_1} (y_i - m_1)^2 + \sum_{n \in C_2} (y_i - m_2)^2$$

$$= \sum_{n \in C_1} (w^T x_n - m_1)^2 + \sum_{n \in C_2} (w^T x_n - m_2)^2$$

$$= \sum_{n \in C_1} (w^T x_n - m_1)(w^T x_n - m_1)^T + \sum_{n \in C_2} (w^T x_n - m_2)(w^T x_n - m_2)^T$$

$$= w^T S_w w$$

$$S_W = \sum_{n \in C_1} (x_n - m_1)(x_n - m_1)^T + \sum_{n \in C_2} (x_n - m_2)(x_n - m_2)^T$$

Intra class variance of full data.

$$\therefore J(\vec{\omega}) = \frac{\omega^T S_B \omega}{\omega^T S_W \omega}$$

$$\frac{\partial J(\vec{\omega})}{\partial \vec{\omega}} = 0$$

$$(\omega^T S_W \omega) S_B \omega = (\omega^T S_B \omega) S_W \omega$$

we fixed $|\omega| \leq 1$

thus by,

$$S_B \vec{\omega} \propto \vec{m}_2 - \vec{m}_1$$

and $\vec{\omega}^T S_W \vec{\omega}$ & $\omega^T S_B \vec{\omega}$ are scalars

$$\boxed{\vec{\omega} = S_W^{-1} (\vec{m}_2 - \vec{m}_1)}$$

This fischer's linear discriminant.

(5) $$L(y,\hat{y}) = \begin{cases} 0 & y=\hat{y} \\ 1 & y\neq\hat{y} \end{cases}$$

$$E_{xy}(L(y,\hat{y}(\vec{x}))) = E_x[E_{y|x}(L(y,\hat{y}(\vec{x})))]$$

$$\because E_{xy}(b(x,y)) = E_x[E_{y|x}(b(x,y))]$$

Now,

$$E_{xy}(L(y,\hat{y}(\vec{x}))) = E_x\left[\sum_{y\in C_k} L(y,\hat{y}(\vec{x}))\, P(y=k|\vec{x})\right]$$

$$y^* = \underset{\hat{y}(x)\in C_k}{\arg\min}\left[E_x\left[\sum_{y\in C_k} L(y,\hat{y}(\vec{x}))\, P(y=k|\vec{x})\right]\right]$$

$$= \underset{\hat{y}(x)\in C_k}{\arg\min}\left[E_x\left[L(y_1,\hat{y})\,p(y=1|\vec{x}) + L(y=2,\hat{y})\,P(y=2|\vec{x}) + \dots L(y=k,\hat{y})\,P(y=k|\vec{x})\right]\right]$$

Let $\hat{y}(\vec{x})=k$, from lossfn def.

$$= \underset{\hat{y}(x)\in C_k}{\arg\min}\left[E_x\left[1 - P(y=k|\vec{x})\right]\right]$$

This is same as maximising for

$P(y=k|\vec{x})$ for $y\in C_k$.

ie,

$$y^* = \underset{y\in C_k}{\arg\max}\; P(y=k|\vec{x})$$