1) We know that

$$L = |y - \hat{y}|^2$$

$$E(L) = \iint \cancel{E}|(y - \hat{y})|^2 p(x,y)\, dx\, dy.$$

$$= \iint |y - \hat{y}(x)|^2 p(x,y)\, dx\, dy.$$

For the Averaged loss to be minimum, differentiale $E(L)$ w.r.t $\hat{y}(x)$.

$$\frac{\partial E(L)}{\partial \hat{y}(x)} = 2\int (y - \hat{y}(x))\, p(x,y)\, dx\, dy$$

$$\Rightarrow \int y\, p(x,y)\, dy - \int \hat{y}(x)\, p(x,y)\, dy = 0$$

$$\Rightarrow \int y\, p(x,y)\, dy - \int \hat{y}(x)\, p(x,y)\, dy = 0.$$

Scanned by CamScanner

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

$$\Rightarrow \hat{y}(x) = \int y \frac{p(x,y)}{p(x)}\, dy$$

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

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

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

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

Simplifying the third term,

$$E\left[(y-y^*)(y^*-\hat{y})\right] = \iint (\hat{y}(x) - y^*(x)).(y^*(x)-y)\, p(x,y)\, dx\, dy$$

$$= \int (\hat{y}(x) - y^*(x)) \int (y^*(x)-y)\, p(x,y)\, dy\, dx$$

But here, $\int (y^*(x) - y) p(x,y) dy\, dx$

$$= \int (y^*(x) - y)\, p(y/x)\, p(x)\, dx.$$

$$= \int y^*(x)\, p(x)\, dx - \underbrace{\int \left(y\, p(y/x) \right) p(x)\, dy\, dx}_{\int E[y/x]\, p(x)\, dx}$$

$$= 0$$

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

Expanding the 2nd term in a dataset D,

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

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

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

$$\therefore E\left[(y - \hat{y}(x))^2 \right] = \text{Varience} + (\text{bias})^2 + \text{noise}$$

3) Least Squares Solution for k-class discriminant classifier.

Let X be d-dimensional and y be labels which belong to one of the classes.

$$\tilde{X} = (\mathbb{1}\ X^T)^T, \quad \tilde{X}^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{N1} \\ \vdots & & & \vdots \\ x_{1d} & \cdots & \cdots & x_{Nd} \end{pmatrix}$$

Let $\hat{y}(x) = \tilde{X}\tilde{W}$ when $\tilde{W}^T$ represents the set of coefficients of dimensions k w.r.t to each class.

$$\tilde{W} = \begin{bmatrix} W_{10} & \cdots & W_{k0} \\ W_{11} & & \vdots \\ \vdots & & \vdots \\ W_{1d} & & W_{kd} \end{bmatrix}$$

Sum of squares $E = \text{trace}(\tilde{X}\tilde{W} - y)^T(\tilde{X}\tilde{W} - y)$

Minimising it,

$$\frac{\partial E}{\partial W} = 0 \Rightarrow \frac{\partial}{\partial \tilde{W}}(\tilde{W}^T\tilde{X}^T - y^T)(\tilde{X}\tilde{W} - y) = 0$$

$$\Rightarrow \frac{\partial}{\partial \tilde{W}}\left(\tilde{W}^T\tilde{X}^T\tilde{X}\tilde{W} - y^T\tilde{X}\tilde{W} - \tilde{W}^T\tilde{X}^T y + y^T y\right) = 0$$

$\Rightarrow \tilde{X}^T\tilde{X}\tilde{W} + \tilde{X}^T\tilde{X}\tilde{W} - \tilde{X}^T y$

$-\tilde{X}^T y = 0.$

$\Rightarrow 2\tilde{X}^T\tilde{X}\tilde{W} - 2\tilde{X}^T y = 0.$

$$\Rightarrow \boxed{\tilde{W}^* = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T y}$$

Using the $\tilde{W}^*$, calculating $\hat{y} = \tilde{X}\tilde{W}^*$ and finding the maximum $\hat{y}_K$, that is the ~~maximum~~ class of the respective input.

4) X is a d-dimensional input, Y is 2-class output Label.

N-samples of which n_1 belong to Class 1 and n_2 to class 2.

V be a unit vector along the imaginary line boundary. Distance of x from the line will be $V^T x$

Therefore mean of projection onto class '1' will be

$$\mu_1 = \frac{1}{n_1}\sum_{x_i \in C_1}^{n_1} V^T x_i = V^T\left(\frac{1}{n_i}\sum_{x_i \in C_1}^{n_1} x_i\right)$$

$$\boxed{\mu_1 = V^T \mu_1'}$$

Scanned by CamScanner

Similarly, $\mu_2 = V^T \mu_2'$

Let the varience be the measure of scatter.

$$S = \sum_{i=1}^{n} (z_i - \mu_2)^2 \quad \text{Varience}$$

$$\boxed{S_1 = \sum_{y_i \in C_1} (y_i - \mu_1)^2 = \sum_{x_i \in C_1} (V^T x_i - V^T \mu_1')^2}$$

Similarly,

$$S_2 = \sum_{y_i \in C_2} (y_i - \mu_2)^2 = \sum_{x_i \in C_2} (V^T x_i - V^T \mu_2')^2$$

In fisher's discriminant process, we will maximise the seperation between the two classes and minimize the scattering

$$J(V) = \frac{(\mu_1 - \mu_2)^2}{S_1 + S_2}$$

Let $\mathcal{S}_1 = \sum_{x_i \in C_1} (x_i - \mu_1')(x_i - \mu_1')^T$

$\mathcal{S}_2 = \sum_{x_i \in C_2} (x_i - \mu_2')(x_i - \mu_2')^T$.

$\therefore S_1 = \sum_{y \in C_1} (v^T x_i - v^T \mu_1)^2$

$= \sum_{x_i \in C_1} v^T (x_i - \mu_1')(x_i - \mu_1')^T v$.

$= v^T \mathcal{S}_1 v$

Similarly, $S_2 = v^T \mathcal{S}_2 v$.

$S_1 + S_2 = v^T (\mathcal{S}_1 + \mathcal{S}_2) v$

$= v^T \mathcal{S}_W v$.

Let $d = (\mu_1' - \mu_2')(\mu_1' - \mu_2')^T$.

Consider $(\mu_1 - \mu_2)^2 = v^T (\mu_1' - \mu_2')(\mu_1' - \mu_2')^T v$.

$= v^T d v$.

$\therefore J(v) = \dfrac{v^T d v}{v^T S_W v}$.

For maximising, $\nabla J(v) = 0$.

$\Rightarrow \dfrac{(v^T S_W v)(2dv) - (2 S_W v) v^T d v}{(v^T S_W v)^2} = 0$

Scanned by CamScanner

$\Rightarrow \quad v^T S_w v \, dv = v^T \cancel{B} \, dv \, S_w v$

$\cancel{\Leftrightarrow \quad dv = v^T dv}$

$\Rightarrow \quad dv = \boxed{\underbrace{(v^T S_w v)^{-1} v^T dv}_{\lambda}} S_w v$

$\Rightarrow \quad dv = \lambda S_w v$

$\Rightarrow \quad \lambda v = (S_w)^{-1} dv$

Consider $d \times v = (\mu_1' - \mu_2')(\mu_1' - \mu_2')^T v$

$= (\mu_1' - \mu_2') \underline{(v^T \mu_1' - v^T \mu_2')}$

$dv = \text{const} \times (\mu_1' - \mu_2')$

$\therefore \quad \lambda v = (S_w)^{-1} (\mu_1' - \mu_2') \times \text{const}$

$\Rightarrow \quad V_{op} \propto S_w^{-1} (\mu_1' - \mu_2')$

$\Rightarrow \quad V_{op} \propto S_w^{-1} (\mu_1' - \mu_2')$

5) Given loss function

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

$$E_{XY}\left[L(\hat{y},y)\right] = E_X\left[E_{Y/X}\left[L(\hat{y},y)\right]\right]$$

$$= E_X\left[\sum_{k \in C_k} L(y,\hat{y}) \cdot p(y=\hat{y}/x)\right]$$

Let $y=y^*$ be the optimal label we know that $\sum_{k=1}^{K} p(y=y^*/x) = 1$.

$$\Rightarrow E_{XY}\left[L(\hat{y},y)\right] = E_X\left[\left(1-p(y=y^*/x)\right)\right]$$

Find y^* such that $\underset{y}{\text{argmin}}.\ E_X\left[1-p(y=y^*/x)\right]$

$\Rightarrow p(y=y^*/x)$ should be maximum.

If $p(y=y^*/x)$ is maximum for a class then it will be predicted class.

$y^* = C_k$ (class k) if

$$p(C_k/x) > p(C_j/x) \quad \forall k \neq j$$