Assignment-2
[EE 5600]

P. Bhargav Ram
EE16BTECH11024

1) Let y^* be the optimal estimator.

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

(where, $p(x,y)$ is the joint distribution of input x and label y.)

$$y^* = \arg\min_{\hat{y}} E(|y-\hat{y}|^2)$$

→ On partial differentiation w.r.t $\hat{y}(x)$,

$$\frac{\delta E[(y-\hat{y})^2]}{\delta \hat{y}(x)} = 2\int (y-\hat{y}(x))\, p(x,y)\, \cancel{dx}\, dy$$

$$\Rightarrow 2\int y\, p(x,y)\, \cancel{dx}\, dy dx - 2\int \hat{y}(x)\, p(x,y)\, dy dx = 0$$

(as, $p(x,y) = p(y|x)\, p(x)$)

$$\Rightarrow \int y^* p(x)\, dx = \int y p(y|x)\, p(x)\, dy\, dx$$

$$\Rightarrow \int y^* p(x)\, dx = \int E(y|x)\, p(x)\, dx$$

$$\Rightarrow y^* = E[y|x]$$

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

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

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

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

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

as, $\int y^* p(y|x)\, dy \, \cancel{\frac{p(x)}{y}} \cancel{dx}$

$$= \int y^*(x)\, p(x)\, dx - \int y\, p(y|x)\, p(x)\, dy\, dx$$

$$= 0$$

$$\Rightarrow E\left[(y-\hat{y}(x))^2\right] = \underbrace{E\left[(y-y^*)^2\right]}_{\text{square of bias}} + \underbrace{E\left[(\hat{y}-\hat{y})^2\right]}_{\text{variance}}.$$

3) Let x be d-dimensional input and y be labels.

where, $y \in C^k$

$$\Rightarrow \tilde{X} = (1, x^T)^T, \quad \tilde{X}^T = \begin{bmatrix} 1 & \cdots & 1 \\ x_{11} & & x_{N1} \\ x_{12} & & \vdots \\ \vdots & & \vdots \\ x_{1d} & & x_{Nd} \end{bmatrix}_{(d+1)\times N}$$

$$\Rightarrow \hat{y}(x) = \tilde{X}\tilde{W}$$

where, $$\tilde{W} = \begin{bmatrix} W_{10} & W_{K0} \\ W_{11} & \vdots \\ \vdots & \vdots \\ W_{1d} & W_{Kd} \end{bmatrix}_{(d+1)\times k}$$

Cost function, $E(W) = \left((\tilde{X}\tilde{W} - Y)^T(\tilde{X}\tilde{W} - \tilde{Y})\right)$

$\Rightarrow$ for min, $\frac{dE}{\delta W} = 0$

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

$$\Rightarrow \tilde{W}^* = (X^TX)^{-1}X^TY$$

Each respective class has one value from y matrix. The class with higher y value will be ~~the~~ class of respective input.

4) In Fisher linear discriminant for 2-class classifier, we create a boundary such that the separation between two classes will be as high as possible with minimum error. The separation is calculated by difference between two means.

Let x be d-dimensional input and y be 2-class label.

There let there be N number of total samples. where n_1 belongs to class 1 and n_2 belongs to class-2.

Let 'v' be unit vector along the imaginary line boundary.

⇒ Distance of x from line will be $V^T x$.

⇒ projection onto given line from Class 1 will be,

$$\mu_1 = \frac{1}{n_1}\sum_{x_i \in C_1}^{n_1} V^T x_i = V^T\left(\frac{1}{n_1}\sum_{x_i \in C_1}^{n_1} x_i\right) = V^T \mu_1'$$

$$\Rightarrow \mu_2 = \frac{1}{n_2}\sum_{x_i \in C_2}^{n_2} V^T x_i = V^T \mu_2'$$

Now, Scatter function ; $S = \sum_{i=1}^{n} (z_i - \mu_z)^2$

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

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

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

In Fisher discriminant process, J(v) will be ~~minimum~~ maximum.

⇒ $S_1 + S_2$ will be maximum.

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

$$= \sum_{y_i \in C_1} \left((x_i - \mu_1')^T V\right)^T \left((x_i - \mu_1)^T V\right)$$

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

Scanned by CamScanner

$\Rightarrow S_1 = V^T S_1 V \quad , \quad S_2 = V^T S_2 V.$

$\Rightarrow S_1 + S_2 = V^T(S_1+S_2)V = V^T S_W V$

Now, $d = (\mu_1' - \mu_2')(\mu_1' - \mu_2')^T$

$$(\mu_1 - \mu_2)^2 = V^T (\mu_1' - \mu_2')(\mu_1' - \mu_2')^T V$$

$$= V^T d V$$

$$\Rightarrow J(v) = \frac{v^T S_W V}{v^T d v}$$

$\Rightarrow \frac{d}{dv}(J(v)) = 0$ (to maximize)

$$\frac{d}{dv}(J(v)) = \frac{(v^T S_W v)\left(\frac{d}{dv}(v^T d v)\right) - (v^T d v)\left(\frac{d}{dv}(v^T S_W v)\right)}{(v^T S_W v)^2}$$

$$= \frac{v^T S_W v(2dv) - v^T dv(2S_W v)}{(v^T S_W v)^2} = 0$$

$$\Rightarrow v^T S_W v\, dv = v^T d v\, S_W v$$

$$dv = \frac{v^T d v}{v^T S_W v} S_W V = \lambda S_W V$$

$$\Rightarrow \lambda V = S_W^{-1} d v$$

as, $dv = (\mu_1' - \mu_2')(\mu_1' - \mu_2)^T \equiv \text{Const.} \times (\mu_1' - \mu_2')$

$$\Rightarrow v = S_W^{-1}(\mu_1' - \mu_2')$$

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

$$y^* = \underset{\hat{y}}{\arg\min}\ E[L(y, \hat{y})]$$

$$= \underset{\hat{y}}{\arg\min}\ E_x[E_{y|x}[L(y, \hat{y})]]$$

$$= \underset{\hat{y}}{\arg\min}\ E_x\left[\sum_{k=1}^{k} L(y, \hat{y})\, p(y = \hat{y}_k | x)\right]$$

$$= \underset{\hat{y}}{\arg\min}\ E_x[1 - P(y = \hat{y}|x)]$$

$$= \underset{\hat{y}}{\arg\max}\ E_x[P(y = \hat{y}|x)]$$

$\Rightarrow\ y^* = C_k$ if $p(c_k|x) > p(c_j|x)\ \forall\ k \neq j$

for maximum $p(y = \hat{y}|x)$, ~~the class is predicted~~ ~~as the~~ the class will be predicted.