Home Work - 1

1) Cost function $L = (\hat{y}(\vec{x}) - y)^2$

Average cost error squared function

$$E[L] = \iint (y - \hat{y}(\vec{x}))^2 \cdot p(\vec{x}, y) \cdot d\vec{x}\, dy$$

For finding optimal estimator, keep $\frac{\partial E[L]}{\partial \hat{y}(\vec{x})} = 0$

$$\Rightarrow \frac{\partial E[L]}{\partial \hat{y}(\vec{x})} = 2 \int (y - \hat{y}(\vec{x})) \cdot p(\vec{x}, y) \cdot dy = 0$$

Since $\hat{y}(\vec{x})$ is not dependent on y

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

$$= \hat{y}(\vec{x}) \cdot p(\vec{x})$$

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

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

$$= \int y \cdot p(y \mid \vec{x})\, dy$$

$$\therefore \hat{y}(\vec{x}) = E[y \mid \vec{x}]$$

2) $(\hat{y}(\vec{x}) - y(x))^2 \rightarrow$ squared error of arbitrary model.

$\hat{y}(\vec{x})$ Defined estimator optimal $y^*(\vec{x}) = E[y|\vec{x}]$

$$(\hat{y}(\vec{x}) - y)^2 = (\hat{y}(\vec{x}) - y^*(\vec{x}) + y^*(\vec{x}) - y)^2$$

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

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

$$\Rightarrow E[(\hat{y}(\vec{x}) - y)^2] = E[(\hat{y}(\vec{x}) - y^*(\vec{x}))^2] + E[(y^*(\vec{x}) - y)^2] + 2.E[(\hat{y}(\vec{x}) - y^*(\vec{x})).(y^*(\vec{x}) - y)] \quad (1)$$

Consider third term in (1)

$$2E[(\hat{y}(\vec{x}) - y^*(\vec{x})).(y^*(\vec{x}) - y)] = 2\iint (\hat{y}(\vec{x}) - y^*(\vec{x})).(y^*(\vec{x}) - y).p(\vec{x}, y)\, d\vec{x}\, dy$$

$$= 2\int (\hat{y}(\vec{x}) - y^*(\vec{x})).\int (y^*(\vec{x}) - y).p(\vec{x}, y)\, dy\, d\vec{x}$$

$$= 2\int (\hat{y}(\vec{x}) - y^*(\vec{x})).\int (y^*(\vec{x}) - y).p(\vec{x}).p(y|\vec{x}).dy\, d\vec{x}$$

$$= 2\int (\hat{y}(\vec{x}) - y^*(\vec{x})).\int (y^*(\vec{x}).p(y|\vec{x}).dy - y.p(y|\vec{x}))\, p(\vec{x})\, d\vec{x}$$

$$= 2\int (\hat{y}(\vec{x}) - y^*(\vec{x})).[y^*(\vec{x}) - y^*(\vec{x})].p(\vec{x})\, d\vec{x}$$

$\hookrightarrow (\because \int y.p(y|\vec{x}) = E(y|\vec{x}))$

$$= 2\int (\hat{y}(\vec{x}) - y^*(\vec{x}))$$

$$= 0$$

$$E[(\hat{y}(\vec{x}) - y)^2] = E[(\hat{y}(\vec{x}) - y^*(\vec{x}))^2] + \underbrace{E[(y^*(\vec{x}) - y)^2]}_{\text{Noise}} \quad (2)$$

Let us consider, a case where model depends on dataset D.

$\hat{y}_D(\vec{x})$ be model from D.

From 2nd term in 2.

$$E_D\left[\left(\hat{y}_D(\vec{x}) - y^*(\vec{x})\right)^2\right] = ? \quad \text{--- } 3$$

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

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

$$+ 2.\left(\hat{y}_D(\vec{x}) - E_D[\hat{y}_D(\vec{x})]\right).\left(E_D[\hat{y}_D(\vec{x})] - y^*(\vec{x})\right) \quad \text{--- } 4$$

Putting 4 in 3

$$E_D\left[\left(\hat{y}_D(\vec{x}) - y^*(\vec{x})\right)^2\right] = E_D\left[\left(E_D[\hat{y}_D(\vec{x})] - y^*(\vec{x})\right)^2\right] + E_D\left[\left(\hat{y}_D(\vec{x}) - E_D[\hat{y}_D(\vec{x})]\right)^2\right] + \underbrace{2.E_D\left[\left(\hat{y}_D(\vec{x}) - E_D[\hat{y}_D(\vec{x})]\right).\left(E_D[\hat{y}_D(\vec{x})] - y^*(\vec{x})\right)\right]}_{=0}$$

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

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

3) K-class linear discriminant classifier

All the classes are distinguished by hyper plane.

$$\hat{y}_k(\vec{x}) = \vec{w}_k^T \cdot \vec{x} + w_{k0}$$

If $\hat{y}_k(\vec{x}) > \hat{y}_j(\vec{x})$ for all $j \neq k$, then output class is k

↳ decision rule.

$$X = \begin{bmatrix} x^{(1)T} \\ x^{(2)T} \\ \vdots \\ x^{(N)T} \end{bmatrix}_{N \times (d+1)} ; \quad \vec{W} = \begin{bmatrix} w_0^{(1)} & w_0^{(2)} & \cdots & w_0^{(K)} \\ \vdots & & & \\ w_d^{(1)} & \cdots & \cdots & w_d^{(K)} \end{bmatrix}^T_{K \times (d+1)}$$

$$\vec{Y} = [0 \; 0 \cdots 1 \; 0 \; 0]$$

↳ k^{th} class.

Sum of squared error

$$E(\vec{w}) = (Y - \vec{w}^T X)^T \cdot (Y - \vec{w}^T X)$$

$$E(\vec{w}) = (Y^T \cdot Y - Y^T \cdot \vec{w}^T \cdot X - \vec{w} \cdot X^T \cdot Y - w^T X^T \cdot X w)$$

$$\frac{\partial E}{\partial w} = 0 = \frac{\partial}{\partial w}(Y^T \cdot Y - \ldots$$

$$\Rightarrow 0 = 2(X^T X \vec{w} - X^T Y)$$

$$\Rightarrow \boxed{\vec{w} = (X^T X)^{-1} \cdot X^T Y}$$

↳ optimal solⁿ.

Scanned by CamScanner

4) Fisher's Linear Discriminant of 2-class classifier can.

$C_0, C_1 \rightarrow$ two classes ; $x^{(i)}, y^{(i)} \rightarrow$ training data.

N - training samples

Projecting $x^{(i)}$ onto 1-dimension axis

$\rightarrow z^{(i)} = w^T . x^{(i)}$ ($w^T \rightarrow$ projection matrix)

We would want the projected data to be as separable as possible when compared to variance.

$\Rightarrow$ same as making the difference of respective classes maximum relative to variance.

$m_0 = w^T . M_0$ $\quad M_0, M_1 \Rightarrow$ means in d dimension.

$m_1 = w^T . M_1$ $\quad m_0, m_1 \rightarrow$ means in 1 dimension.

Variance $$s_0^2 = \sum_{x_i \in C_0} (w^T . x^{(i)} - m_0)^2$$

$$s_1^2 = \sum_{x_i \in C_1} (w^T . x^{(i)} - m_1)^2$$

Maximum $$J(w) = \frac{(m_1 - m_0)^2}{s_0^2 + s_1^2}$$

$$J(w) = \frac{(w^T M_1 - w^T M_0) . (w^T M_1 - w^T M_0)^T}{\sum_{x_i \in C_0} [w^T (x^{(i)} - M_0)]^2 + \sum_{x_i \in C_1} [w^T (x^{(i)} - M_1)]^2}$$

$$\Rightarrow J(w) = \frac{w^T . (M_1 - M_0) . (M_1 - M_0)^T . w}{\sum_{x_i \in C_0} w^T . (x^{(i)} - M_0)(x^{(i)} - M_0)^T . w + \sum_{x_i \in C_1} w^T . (x^{(i)} - M_1) . (x^{(i)} - M_1)^T . w}$$

$$\therefore J(w) = \frac{w^T . S_B . w}{w^T . S_W . w}$$; $S_B, S_W \rightarrow$ symmetric.

Scanned by CamScanner

$$\frac{\partial J(\vec{w})}{\partial w} = \frac{\frac{\partial}{\partial w}(w^T \cdot S_B w)}{w^T \cdot S_w \cdot w} - \frac{w^T \cdot S_B \cdot w}{(w^T \cdot S_B \cdot w)^2} \cdot \frac{\partial (w^T S_w \cdot w)}{\partial w}$$

$$\Rightarrow \frac{\partial J(\vec{w})}{\partial w} = \frac{(S_B + S_B^T) \cdot w}{w^T \cdot S_w \cdot w} - \frac{w^T \cdot S_B \cdot w}{(w^T \cdot S_w \cdot w)^2} (S_w + S_w^T) \cdot w$$

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

$$\Rightarrow (2 S_B w) \cdot (w^T \cdot S_w \cdot w) - (w^T \cdot S_B \cdot w)(2 S_w \cdot w) = 0$$

$$\Rightarrow (w^T \cdot S_w \cdot w)(S_B w) = (w^T S_B w) \cdot (S_w w)$$

$$\Rightarrow S_B w = \left(\frac{w^T S_B w}{w^T S_w w} \right) \cdot S_w \cdot w$$

$$S_B = (M_1 - M_0) \cdot (M_1 - M_0)^T \cdot w$$

$S_B \cdot w$ is in direction of $(M_1 - M_0)$

$$S_w \cdot w = \lambda \cdot (S_B w)$$

$S_w \cdot w$ is also in direction of $(M_1 - M_0)$

$$\boxed{w = S_w^{-1} \cdot (M_1 - M_0)}$$

Optimal solution.

5) Recall $y^*(\vec{x}) = E[Y|\vec{x}]$ is the optimal estimator

Average squared error $E[(y-\hat{y}(\vec{x}))^2]$

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

↳ loss function

Zero-one loss function $\left\{ L(y,\hat{y}(\vec{x})) = \begin{cases} 0, & y = \hat{y}(\vec{x}) \\ 1, & y \neq \hat{y}(\vec{x}) \end{cases}\right.$

$$E_{XY}[L(y,\hat{y}(\vec{x}))] = E_X\left[\sum_{y \in C_k} L(y,\hat{y}(\vec{x}))\, p(y=k|\vec{x})\right]$$

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

$$= \underset{\hat{y}(\vec{x})}{\arg\min}\ E_X\left[\begin{matrix} L(y=1,\hat{y}(\vec{x})) \cdot p(y=1|\vec{x}) + \\ \vdots \\ L(y=k,\hat{y}(\vec{x})) \cdot p(y=k|\vec{x}) \end{matrix}\right]$$

Let $\hat{y}(\vec{x}) = k'$

$$y^*(\vec{x}) = \underset{\hat{y}(\vec{x})}{\arg\min}\ E_X[1 - p(y=k'|\vec{x})]$$

$$= \underset{\hat{y}(\vec{x})}{\arg\min}\ E_X[1] - E_X[p(y=k'|\vec{x})]$$

$$\therefore\ y^*(\vec{x}) = \underset{k}{\arg\max}\ p(y=k|\vec{x})$$