Homework - 1

Nakka Chaitanya
EE16BTECH11022

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

by Bayes Theorem

And, in the statistical setting, we aim to get an estimate for the absolute best model $\hat{y}(\vec{x})$ for SSE cost func

$$E(\vec{x}) = \mathbb{E}\left((y-\hat{y})^2\right)$$

$$E(\vec{x}) = \iint \left(y(\vec{x}) - \hat{y}(\vec{x})\right)^2 p(x,y)\, d\vec{x}\, dy$$

for the best $\hat{y}(\vec{x})$, minimize $\vec{E}(\vec{x})$ wrt $\hat{y}$

$$\frac{d}{d\hat{y}(\vec{x})} E(\vec{x}) = 0$$

$$\Rightarrow \frac{d}{d\hat{y}(\vec{x})} \iint (y-\hat{y})^2 p(x,y)\, d\vec{x}\, dy = 0$$

$$\Rightarrow \iint -2(y-\hat{y})\, p(x,y)\, d\vec{x}\, dy = 0$$

$\rightarrow$ $\cancel{\iint y\, p(\vec{x},y)\, d\vec{x}\, dy = \iint \hat{y}\, p(\vec{x},y)\, d\vec{x}}\, dy$

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

$$\iint \cancel{y}\, p(y|\vec{x})\, p(\vec{x})\, (y-\hat{y})\, d\vec{x}\, dy = 0$$

$$\iint \left(y\, p(y|\vec{x})\, dy \right) p(x)\, dx = \iint \hat{y}\, p(x,y)\, dx\, dy$$

$$= \iint \mathbb{E}(y|x)\, p(x)\, dx = \hat{y} \iint p(x,y)\, dx\, dy$$

$\mathbb{E}(y|x)$ is indepenedent of $\vec{x}$

& $\iint p(x,y)\, dx\, dy = 1$ & $\int p(x)\, dx = 1$

$$\rightarrow \boxed{\hat{y} = \mathbb{E}(y|\vec{x})}$$

(2) Let y be the labels for the given data

i.e, $y(x)$

$\hat{y}(x)$ is the arbitrary model's label estimate

$y^*(x)$ is the absolute best model for the given data

then :-

$$E(\vec{x}) = \text{Cost function} = \mathbb{E}((y-\hat{y})^2)$$

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

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

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

Consider $\mathbb{E}((y-y^*)(y^*-\hat{y}))$

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

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

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

$\int p(y|x)\, dy = 1$ & y^* is independent of y

$$= \int_x y^* - \hat{y} \left[\int \left(y\, p(y|x)\, dy \right) - y^* \right] p(x)\, dx$$

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

$$\Rightarrow \mathbb{E}\left((y - y^*)(y^* - \hat{y}) \right) = 0$$

Consider $\mathbb{E}\left((y^* - \hat{y})^2 \right)$

We actually have access to say D datasets of the same experiment. Then

$$\mathbb{E}\left((\hat{y} - y^*)^2 \right) = \mathbb{E}_D\left((\hat{y}_D - y^*)^2 \right)$$

$$= \mathbb{E}_D\left(\hat{y}_D^2 + y^{*2} - 2\hat{y}_D y^* \right) \text{ (or)}$$

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

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

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

Again consider $\mathbb{E}_D\left[\left(\hat{y}_D - \mathbb{E}_D(\hat{y}(x)) \right) \left(\mathbb{E}_D(\hat{y}(x)) - y^* \right) \right]$

$$= \int \left(\hat{y}_D - \mathbb{E}_D(\hat{y}(x)) \right) \left(\mathbb{E}_D(\hat{y}(x)) - y^* \right) p(x, \hat{y})\, dx\, d\hat{y}$$

$$= \int \left(\mathbb{E}_D(\hat{y}(x)) - y^* \right) \int \underbrace{\left(\hat{y}_D - \mathbb{E}_D(\hat{y}(x)) \right) p(y)\, dy}_{\to 0} \cdot p(x|y)\, dx$$

$= 0$

$$\Rightarrow \mathbb{E}_D\left((y-\hat{y})^2\right) = \mathbb{E}\left((y-y^*)^2\right) + \mathbb{E}_D\left(\left(\hat{y}-\mathbb{E}_D(\hat{y})\right)^2\right)$$

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

here $\mathbb{E}\left((y-y^*)^2\right)$ = noise

$\mathbb{E}\left[\left(\hat{y}-\mathbb{E}_D(\hat{y}(x))\right)^2\right]$ = Variance

$\mathbb{E}\left[\left(\mathbb{E}_D(\hat{y})-y^*\right)^2\right]$ = Bias.

(5) $Lf = \begin{cases} 0 & \hat{y} = y \\ 1 & \text{else} \end{cases}$ $\hat{y} \to$ estimate

$y \to$ label

for the statistical setting

$\to p(y|x)$ must be maximized

$p(y|x) \propto p(x|y)\,p(y)$

We know the samples follow some distribution.

and for optimal solution

y^* for which $\underset{\hat{y}}{\operatorname{argmin}} \left(\mathbf{E}\, Lf(y, \hat{y}) \right)$

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

$$= E_X\left[\int Lf(y,\hat{y})\, p(y|x)\, dy \right]$$

now since discrete, $\int \to \Sigma$

$$= E_X\left[\sum_{y \in C_k} Lf(y,\hat{y})\, p(y|x) \right]$$

let $\hat{y} = k'$ $k' \in C_k$

$$= E_X\left[Lf(y,k')\, pr(y=1|x) + Lf(y,k')\, pr(y=2|x) + \dots \right]$$

$$= E_X\left[\sum_{y \in C_k} pr(y=k|x) - pr(y=k'|x) \right]$$

$= \mathbb{E}_x \left(1 - \Pr (y = k' | x) \right)$

for optimal solution

$\mathbb{E}_x \left[1 - \Pr (y = k' | x) \right]$ is minimized

$\Rightarrow \quad \Pr (y = k' | x)$ is maximized

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

which is intuitive and what we try to do in each and every classification problem $\rightarrow$ maximize the prob. of a label given $\vec{x}$

③ K-Class Linear Discriminant Classifier

Each class is seperated by a hyperplane

$$y_K(x) = w_K^T \boldsymbol{x} + w_{K0}$$

Note: Till now we've been using $\hat{y} = xW$

$$x \rightarrow \begin{bmatrix} x_0^1 & x_1^1 & \cdots & x_d^1 \\ \vdots & & & \end{bmatrix}$$

here $x = \begin{bmatrix} x_0^1 \\ x_1^1 \\ \vdots \\ x_d^1 \end{bmatrix}$ &

$$X = \begin{bmatrix} \mathbf{x}_0^1 & x_1^1 & \cdots & \mathbf{x}_d^1 \\ x_0^2 & x_1^2 & \cdots & x_d^2 \\ \vdots & & & \vdots \end{bmatrix}$$

$x \rightarrow$ Column vector of components

$X \rightarrow$ whole data in standard format

To merge bias term, consider

$$\tilde{W} = \begin{bmatrix} W_{10} & W_{20} & \cdots & W_{N0} \\ W_{11} & W_{21} & & W_{N1} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1d} & W_{2d} & \cdots & W_{Nd} \end{bmatrix}$$

→ Cost function $= \mathbb{E}\left((y-\hat{y})^2\right)$

$=$ Consider t_n as the estimate vector for x_n which will be of ~~to~~ a column of K elements, each element being either 0 or 1

(4) Consider the SSE cost function

$$CF = E(W) = \sum_{i=1}^{N} \left[y^{(i)} - \hat{y}^{(i)} \right]^2$$

for i^{th} row of y and k^{th} term (dimention)

$$CF_{ik} = \left(y_k^{(i)} - \sum_{j=0}^{d} x_j^{(i)} w_{jk} \right)^2$$

for optimal w^*

$$\frac{\partial CF_{ik}}{\partial w_{jk}} = 0 \Rightarrow -2x_j^{(i)} \left(y_k^{(i)} - \sum_{j=0}^{d} x_j^{(i)} w_{jk} \right) = 0$$

Ref for multidimensional labels proof

→ In matrix format

$$X^T (Y - XW^*) = 0$$

which yeilds $\boxed{W^* = (X^T X)^{-1} X^T Y}$

In case of basis functions;

$$\boxed{W^* = (\phi^T \phi)^{-1} \phi^T Y}$$

their orders being $d+1 \times K$ and

$(M+1) \times K$ respectively

Fischer's Linear Discriminant

Consider two class classification problem.

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

$$\vec{m}_2 = \frac{1}{N_1} \sum_{n \in C_2} \vec{x}_n$$

The main idea of Fischer's LDA is to prevent overlap of classes during dimensionality reduction

To get to a single dimension from multiple dimension vector, consider its component on a vector W of the same dimensionality

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

$$m_k = W^T \vec{m}_k$$

To restrict the vector itself, consider $||w|| = 1$

To reduce intra-class covariance and increase inter-class variance, Fischer's criterion is used.

$$J(w) = \frac{(m_2 - m_1)^2}{S_1^2 + S_2^2}$$

$$(m_2 - m_1)^2 = (w^T \vec{m}_k)^2$$

$$= w^T \vec{m}_k \vec{m}_k^T w.$$

$$\boxed{\vec{m}_k \vec{m}_k^T = S_B = \text{between class covariance matrix}}$$

$|||^{ly}$ $\quad S_1^2 + S_2^2$

$$= \sum_{n\in C_1} (y_1 - m_1)^2 + \sum_{n\in C_2} (y_2 - m_2)^2$$

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

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

$$= w^T S_w w$$

$$\boxed{S_w = \sum_{n\in C_1} (\vec{x}_n - \vec{m}_1)(\vec{x}_n - \vec{m}_1)^T + \sum_{n\in C_2} (\vec{x}_n - \vec{m}_2)(\vec{x}_n - \vec{m}_2)^T}$$

(within class variance)

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

$$= (\vec{m}_2 - \vec{m}_1)\, \cancel{m_k^T} \rightarrow \text{scalar}$$

$\Rightarrow$ $\boxed{S_B(\vec{w}) \text{ is in direction of } (\vec{m}_2 - \vec{m}_1)}$

for best discrimination

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

$$\frac{d}{dw}\left(\frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_W \vec{w}}\right) = 0$$

$$\Rightarrow (w^T S_W w)(S_B w) = (w^T S_B w)(S_W w)$$

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

$w^T S_W w$ & $w^T S_B w$ are scalars

$$\Rightarrow \boxed{w = S_W^{-1} (\vec{m}_2 - \vec{m}_1}$$

which is known as fischer's linear discriminant