HW-2

1.) $\text{Loss} = E\left(\left(y(x) - \hat{y}(x)\right)^2\right)$ where $\hat{y}(x)$ is the estimator

$$\Rightarrow L = \iint \left(y(x) - \hat{y}(x)\right)^2 p(x,y)\,dy\,dx$$

$$\frac{\partial L}{\partial \hat{y}} = 0$$

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

$$\Rightarrow \int_y 2\left(y - \hat{y}\right) p(x,y)\, dy = 0$$

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

$$\Rightarrow \left(\cancel{p(x)}\right) E(y|x) = \hat{y}\left(\cancel{p(x)}\right)$$

$$\Rightarrow \boxed{\hat{y}(x) = E(y|x)}$$

2.) <u>Bias-Variance</u> :-

Say $\hat{y}_D(\vec{x})$ is estimator for the data D & $y^* = E(y|\vec{x})$

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

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

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

$$= E\left((y - y^*)^2\right) + E\left((y^* - \hat{y}_D)^2\right) + 2\iint (y - y^*)(y^* - \hat{y}_D)\, p(x,y)\, dy\, dx$$

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

$$= \int (y^* - \hat{y}_D)\left(\overbrace{\cancel{E(y|x) - y^*}}^{0}\right) dx$$

$$= 0.$$

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

Say $E_D(\hat{y}_D)$ is the mean of the estimators $\hat{y}_D$

$$\Rightarrow E_D\left((y^* - \hat{y}_D)^2\right)$$

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

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

$$+ E_D\left((\hat{y}_D - E_D(\hat{y}_D))^2\right)$$

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

$$\because E_D(y_D - E_D(\hat{y}_D)) = 0$$

$$= \underbrace{(y^* - E_D(\hat{y}_D))^2}_{\text{Bias}}$$

$$+ \text{variance}$$

$$\therefore E\left((y - \hat{y}_D)^2\right)$$

$$= (\text{Bias})^2 + \text{Variance} + \text{Noise}$$

3.) $L = tr\left[(\tilde{X}\tilde{W} - y)^T (\tilde{X}\tilde{W} - y)\right]$

where

$\tilde{X}$'s k^{th} column is $(1, x^T)^T$

& $\tilde{W}$'s " " " " $(W_{k0}, w_k^T)^T$

$$\Rightarrow L = \sum_i \sum_j (\tilde{X}\tilde{W} - y)^T_{ij} \cdot (\tilde{X}\tilde{W} - y)_{ji}$$

$$= \sum_i \sum_j (\tilde{X}\tilde{W} - Y)^2_{ji}$$

$$= \sum_i \sum_j \left((\tilde{X}\tilde{W})_{ji} - Y_{ji}\right)^2$$

$$= \sum_i \sum_j \left(\sum_k \tilde{X}_{jk} \cdot W_{ki} - Y_{ji}\right)^2$$

$$\frac{\partial L}{\partial W_{ki}} = 0 \Rightarrow \sum_i \sum_j \left(\tilde{X}_{jk}\right)\left(\sum_k \tilde{X}_{jk} \cdot W_{ki} - Y_{ji}\right) = 0$$

$$\Rightarrow \sum_i \sum_j (\tilde{X}^T)_{kj} \left((\tilde{X}W)_{ji} - Y_{ji}\right) = 0$$

$$\Rightarrow \tilde{X}^T(\tilde{X}W) - (\tilde{X})^T Y = 0$$

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

$$\boxed{\therefore\ w = (X^T X)^{-1} X^T Y}$$

The above is same as that multiple output case in Linear regression.

④ Fischer's Linear Discriminant :-

> $y = (\vec{w})^T \vec{x}$ is basically projecting a $(D+1)$ dimension vector to one dimension
$\vec{w}$ has many possibilities.

> we choose a $\vec{w}$ such that the intra-class variance is minimized & inter-class " is maximized.

> Consider 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

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

> This can be done by having arbitrarily large w (which is not preferred, cause it might lead to overfit)

So we constrain w to have ≤ unit length i.e., $\sum_i w_i^2 \leq 1 \Rightarrow \vec{w} \propto (\vec{m}_2 - \vec{m}_1)$

> the within class variance is given by

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

where

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

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

> The fisher criterion is defined as

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

But

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

$$\Rightarrow \cancel{(m_2 - m_1)^2 = (w^T(\bar{m}_2 - \bar{m}_1))^T (w^T(\bar{m}_2 - \bar{m}_1))}$$

$$\Rightarrow (m_2 - m_1)^2 = w^T \underbrace{\boxed{(\bar{m}_2 - \bar{m}_1)(\bar{m}_2 - \bar{m}_1)^T}}_{S_B} w$$

$$s_1^2 + s_2^2 = w^T S_w w$$

where

$$\underbrace{\boxed{S_W = \sum_{n \in C_1} (x_n - m_1)(x_n - m_1)^T + \sum_{n \in C_2} (x_n - m_2)(x_n - m_2)^T}}_{\text{within class variance of full data}}$$

within class variance of full data.

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

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

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

we fixed $|w| \leq 1$

so dropping the scalars

$$\Rightarrow (\bar{m}_2 - \bar{m}_1)(\bar{m}_2 - \bar{m}_1)^T w = S_w w$$

constant

$$\Rightarrow \boxed{w \propto S_w^{-1}(\bar{m}_2 - \bar{m}_1)}$$

KNN :-

- What was observed is that data is distributed almost similarly about origin.
- So if the test sample is close to origin, we need to have a bigger K value to make a concrete prediction.

5.)

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

$$E_{XY}\left(L(y, \hat{y}(\vec{x}))\right)$$

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

$$\because E_{XY}(f(x,y)) = E_X\left(E_{Y|X}(f(x,y))\right)$$

$\Rightarrow$ we need to find $\cancel{y^*}$ $\hat{y}(\vec{x}) = y^*$ such that the above expectation is minimized.

$$\sum_{y \in C_k} L(y, \hat{y}(\vec{x})) \cdot p(y=k|\vec{x})$$

If $y = \hat{y}(\vec{x})$, then $L(y, \hat{y}) = 0$

$\Rightarrow$ Summation turns out to be

$$\sum_{y \in C_k,\, y \neq \hat{y}} P(y=k|\vec{x}) = 1 - p(\hat{y}=k|\vec{x})$$

$$\Rightarrow y^* = \underset{\hat{y}(\vec{x})}{\arg\min}\, E_X\left(1 - p\left(y = \hat{y}(\vec{x}) \mid \vec{x}\right)\right)$$

say $\hat{y} = k$

This is same as maximizing

$$p(y = k \mid \vec{x}) \quad \text{for} \quad y \in C_k$$

i.e.,

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