Mini-Project #1

ECE 5268 Richard Hemphill

Problem 1

a. Show $\bar{y} = \bar{x}^T w^* + b^*$

In other words, the hyper-plane (e.g. line) used to model the data goes through the mean of the input/output (i.e. $\bar{x}$ and $\bar{y}$). ~~of the~~ When optimized, the bias term is set so that the line goes through center of data.

$$\frac{\partial}{\partial b} MSE_N(w, b) = 0$$

$$\frac{\partial}{\partial b} \frac{1}{N} \left\| \overbrace{X^T w + b^*}^{\hat{y}\text{ (predicted)}} - \underbrace{y}_{\text{actual}} \right\| = 0$$

$$\frac{1}{N} \frac{\partial}{\partial b} (X^T w + b^* - y)^T (X^T w + b^* - y) = 0$$

$$\frac{1}{N} \frac{\partial}{\partial b} \left[(w^T X + b^{*T} - y^T)(X^T w + b^* - y) \right] = 0$$

$$\frac{1}{N} \frac{\partial}{\partial b} \left[w^T X X^T w + w^T X b^* - w^T X y + b^{*T} X^T w + b^{*T} b^* - b^{*T} y - y^T X^T w - y^T b^* + y^T y \right] = 0$$

$$\frac{1}{N}\left[\cancel{(w^T X)^T} + X^T w + 2b^* - y - y\right] = 0 \quad (w^T X)^T \to X^T w$$

$$\frac{1}{N}\left[2X^T w + 2b^* - 2y\right] = 0$$

$$\frac{1}{\cancel{2}}\,\frac{\cancel{2}}{N}\left[X^T w + b^* - y\right] = 0\,\frac{1}{2}$$

$$\left(\frac{1}{N}(X^T w + b^* - y)\right)^T = (0)^T$$

$$\frac{1}{N}\left[w^T X + b^{*T} - y^T\right] = 0$$

$$\frac{1}{N} w^T X + \frac{1}{N} b^{*T} - \cancel{\frac{1}{N}(\bar{w})} - \frac{1}{N} y^T = 0$$

$$\left(\frac{1}{N} w^T X + \frac{1}{N} b^{*T} - \frac{1}{N} y^T\right) \mathbb{1}_N = 0 \cdot \mathbb{1}_N$$

$$\frac{1}{N} w^T X \mathbb{1}_N + \frac{1}{N} b^{*T} \mathbb{1}_N - \frac{1}{N} y^T \mathbb{1}_N = 0$$

(1) $\frac{1}{N} w^T X \mathbb{1}_N = \frac{1}{N} (\underbrace{X^T w}_{z})^T \mathbb{1}_N = \frac{1}{N} z^T \mathbb{1}_N = \bar{z} = \bar{X}^T w$

(2) $\frac{1}{N} b^T \mathbb{1}_N = \frac{1}{N} \begin{bmatrix} b & \cdots & b \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

$= \frac{1}{N} \begin{bmatrix} b + \dots + b \end{bmatrix} = \frac{1}{N} N b = b$

(3) $\frac{1}{N} y^T \mathbb{1}_N = \bar{y}$

$$\underset{(1)}{\bar{x}^T w} + \underset{(2)}{b} - \underset{(3)}{\bar{y}} = 0$$

$$\boxed{\bar{y} = \bar{x}^T w + b^*}$$

b. Show

$$w^* = \arg\min_{w \in R^D} MSE_{\mathcal{D}}(w, \overbrace{\bar{y} - \bar{x}^T w}^{b^*})$$

$$= \arg\min_{w \in R^D} \frac{1}{N} \| \hat{X} w - \hat{y} \|_2^2$$

In other word, the optimal weights can be determined w/o bias by centering the data through the origin. Subtracting the mean centers the data (i.e $\hat{X}$ and $\hat{y}$).

$$\arg\min_{w \in R^D} MSE_{\mathcal{D}}(w, \bar{y} - \bar{x}^T w)$$

$$\arg\min_{w \in R^D} \frac{1}{N} \| X^T w + \overbrace{b}^{\bar{y} - \bar{x}^T w} - y \|_2^2$$

$$\arg\min_{w \in R^D} \frac{1}{N} \| X^T w - \bar{x}^T w - y + \bar{y} \|_2^2$$

$$\arg\min_{w \in R^D} \frac{1}{N} \| (X^T - \bar{x}^T) w - (y + \bar{y}) \|_2^2$$

mult means by 1_N so that dimensions match

$$\arg \min_{w \in \mathbb{R}^p} \frac{1}{N} \left\| \left(X^T - \bar{x}^T 1_N\right) w - \left(y - \bar{y} 1_N\right) \right\|_2^2$$

$$\boxed{\arg \min_{w \in \mathbb{R}^p} \frac{1}{N} \left\| \tilde{X} w - \tilde{y} \right\|_2^2}$$

c. Show $C_{xx} w^* = c_{xy}$

The correlation of the ~~actual~~ centered input / target is dependent on the weights (i.e. $w^* = 0 \Rightarrow c_{xy} = 0$)

$$w^* = \arg\min_{w \in \mathbb{R}^D} \frac{1}{N} \| \check{X} w - \check{y} \|_2^2$$

$$\frac{\partial}{\partial w} \frac{1}{N} \| \check{X} w^* - \check{y} \|_2^2 = 0$$

$$\frac{1}{2} \cdot \frac{2}{N} \check{X}^T (\check{X} w^* - \check{y}) = 0 \cdot \frac{1}{2}$$

$$\frac{1}{N} \check{X}^T \check{X} w - \cancel{2} \frac{1}{N} \check{X}^T y = 0$$

$$C_{xx} w^* - c_{xy} = 0$$

$$\boxed{C_{xx} w^* = c_{xy}}$$

d. Show

$$MSE_{\gamma}(w^*, b^*) = c_{yy} - \| c_{xy} \|^2_{C_{xx}^{-1}}$$

Where

$$c_{yy} \triangleq \frac{1}{N} \| \check{y} \|_2^2 = \frac{1}{N} \sum_{n=1}^{N} (y_n - \bar{y})^2$$

$$C_{xx} w^* = c_{xy} \Rightarrow w^* = C_{xx}^{-1} c_{xy}$$

$$\frac{1}{N} \| \check{X} w^* - \check{y} \|_2^2$$

$$\frac{1}{N} \| \check{X} C_{xx}^{-1} c_{xy} - \check{y} \|_2^2$$

$$\frac{1}{N} (\check{X} C_{xx}^{-1} c_{xy} - \check{y})^T (\check{X} C_{xx}^{-1} c_{xy} - \check{y})$$

$$\frac{1}{N} (c_{xy}^T C_{xx}^{-1T} \check{X}^T - \check{y}^T)(\check{X} C_{xx}^{-1} c_{xy} - \check{y})$$

$$\frac{1}{N} \left[c_{xy}^T C_{xx}^{-1T} \check{X}^T \check{X} c_{xx}^{-1} c_{xy} - c_{xy}^T C_{xx}^{-1T} X^T y \right.$$
$$\left. - y^T X c_{xx}^{-1} c_{xy} + y^T y \right]$$

$$c_{xy}^T C_{xx}^{-1T} \overset{\to I}{\cancel{C_{xx} C_{xx}}} c_{xy} - c_{xy}^T C_{xx}^{-1T} c_{xy}$$
$$- c_{xy}^T c_{xx}^{-1} c_{xy} + c_{yy}$$

$$\cancel{C_{xy}^T C_{xx}^{-1\,T} C_{xy} - C_{xy}^T C_{xx}^{-1\,T} C_{xy}} - C_{xy}^T C_{xx}^{-1} C_{xy} + C_{yy}$$

$$-\|C_{xy}\|^2_{C_{xx}^{-1}} + C_{yy}$$

$$\boxed{C_{yy} - \|C_{xy}\|^2_{C_{xx}^{-1}} = MSE_{\sim}(w^*, b^*)}$$

e. Show:

i. $\min MSE_n(0,b) = c_{yy}$

ii. $R^2 = \frac{\|c_{xy}\|^2_{C_{XX}^{-1}}}{c_{xy}}$ for $y_n \neq$ const

iii. $\underset{\text{worst fit}}{0} \leq R^2 \leq \underset{\text{perfect fit}}{1}$

Where:

$$R^2 \triangleq \begin{cases} 1 & \text{if } y_n = \text{const for all } n \\ 1 - \dfrac{\min_{w \in \mathbb{R}^p, b \in \mathbb{R}} MSE_n(w,b)}{\min_{b \in \mathbb{R}} MSE_n(0,b)} & \text{otherwise} \end{cases}$$

e i. $\min_{b \in \mathbb{R}} MSE_n(0,b) = c_{yy}$

$$b^* = \bar{y} - \bar{X}^T w \xrightarrow{w=0} b = \bar{y}$$

$$\frac{1}{N} \| \cancel{X^T 0} + \overset{b}{\bar{y}} - y \|_2^2$$

$$\frac{1}{N} \sum_{n=1}^{N} (\bar{y} - y_n)^2 = \frac{1}{N} \sum_{n=1}^{N} \left[(-1)(y_n - \bar{y})\right]^2$$

$$\frac{1}{N} \sum_{n=1}^{N} (y_n - \bar{y})^2 = \boxed{c_{yy} = \min_{b \in \mathbb{R}} MSE_n(0,b)}$$

$$\text{eii.}\quad R^2 = 1 - \frac{\overbrace{C_{yy} - \|C_{xy}\|^2_{C_{xx}^{-1}}}^{\text{part 1d}}}{\underbrace{C_{yy}}_{\text{part 1ei}}}$$

$$R^2 = \frac{\cancel{C_{yy}} - \cancel{C_{yy}} + \|C_{xy}\|^2_{C_{xx}^{-1}}}{C_{yy}}$$

$$\boxed{R^2 = \frac{\|C_{xy}\|^2_{C_{xx}^{-1}}}{C_{yy}}}$$

eiii. $R^2 \geq 0$ due to both terms (i.e. $\|C_{xy}\|^2_{C_{xx}^{-1}}$ and C_{xy}) being positive from square

$R^2 \leq 1$ due to $1 - \frac{\min_{w \in \mathbb{R}^p b \in \mathbb{R}} MSE_\tau(w,b)}{\min_{b \in \mathbb{R}} MSE_\tau(0,b)}$

the $MSE_\tau(0,b)$ will always be larger (or eq) to $MSE_\tau(w,b)$ since optimize error ($MSE_\tau(w,b)$) should be relatively lower than non-optimized.

$\therefore\ 0 \leq R^2 \leq 1$