$\boxed{\text{Mean Squared Error Regression}}$

(Also) Least-squares Regression

Use Mean-squared error (MSE)

$\therefore$ Criterion Function: $J(\underline{w}) = \frac{1}{N}\sum_{i=1}^{N}[g(\underline{x}_i) - y_i]^2 = MSE$

where $g(\underline{x}_i) = \underline{w}^T\underline{x}_i$

Algebraic Method [Ordinary Least Squares (OLS)]

$$J(\underline{w}) = \frac{1}{N}\underbrace{\sum_{i=1}^{N}[\underline{w}^T\underline{x}_i - y_i]^2}$$

$$\text{Sum}\left\{\begin{bmatrix}(\underline{w}^T\underline{x}_1 - y_1)^2 \\ (\underline{w}^T\underline{x}_2 - y_2)^2 \\ \vdots \\ (\underline{w}^T\underline{x}_N - y_N)^2\end{bmatrix}\right\} = \left\|\begin{bmatrix}\underline{w}^T\underline{x}_1 - y_1 \\ \vdots \\ \underline{w}^T\underline{x}_N - y_N\end{bmatrix}\right\|_2^2 = \left\|\begin{bmatrix}\underline{x}_1^T \\ \vdots \\ \underline{x}_N^T\end{bmatrix}\cdot\underline{w} - \begin{bmatrix}y_1 \\ \vdots \\ y_N\end{bmatrix}\right\|_2^2 = \|\underline{\underline{X}}\,\underline{w} - \underline{y}\|^2$$

$$\therefore J(\underline{w}) = \frac{1}{N}\|\underline{\underline{X}}\,\underline{w} - \underline{y}\|^2 = \frac{1}{N}(\underline{\underline{X}}\,\underline{w} - \underline{y})^T(\underline{\underline{X}}\,\underline{w} - \underline{y})$$

Minimize it: $\nabla_{\underline{w}} J(\underline{w}) = \frac{1}{N}[2\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{w} - 2\underline{\underline{X}}^T\underline{y}] = 0$

$\Rightarrow \underline{\underline{X}}^T\underline{\underline{X}}\,\hat{\underline{w}} = \underline{\underline{X}}^T\underline{y}$

$\boxed{\hat{\underline{w}} = (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T y}$ if $\underline{\underline{X}}^T\underline{\underline{X}}$ nonsingular

$\underline{\underline{X}}^{-} = (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T \leftarrow$ Moore-Penrose (left) pseudoinverse of $\underline{\underline{X}}$

Gradient Descent Method [Least Mean Squares (LMS)]

Ex: Sequential GD

$$J(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N}(\underline{w}^T\underline{x}_n - y_n)^2 = \sum_{n=1}^{N} J_n(\underline{w})$$

$$\therefore J_n(\underline{w}) = \frac{1}{N}[\underline{w}^T\underline{x}_n - y_n]^2$$

$$\nabla_{\underline{w}} J_n(\underline{w}) = \frac{2}{N}[\underline{w}^T\underline{x}_n - y_n]\cdot\underline{x}_n$$

Use sequential GD

weight update $\boxed{\underline{w}(i+1) = \underline{w}(i) - \eta(i)\frac{2}{N}[\underline{w}(i)^T\underline{x}_n - y_n]\underline{x}_n}$

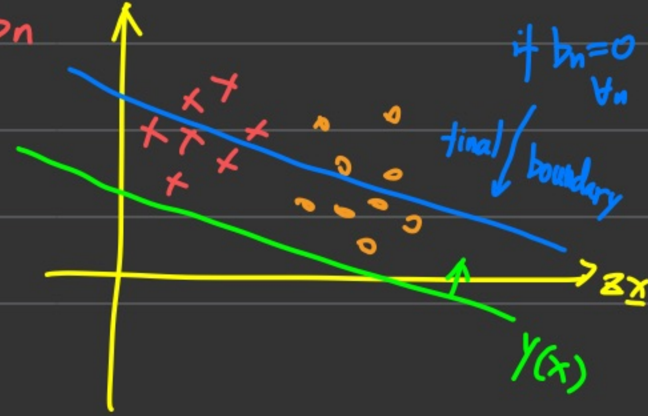
Mean Squared Error Classification

- Algebraic: Pseudoinverse learning algorithm
- Gradient Descent: Widrow-Hoff

Instead of $J(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N}(\underline{w}^T\underline{x}_n - y_n)^2$

We use $J(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N}[\underline{w}^T z_n \underline{x}_n - b_n]^2$, $b_n > 0$ $\forall n$

$\therefore$ only when all points are correctly classified ($\underline{w}^T z_n \underline{x}_n > 0$) and approaching to b_n



Then $J(\underline{w})$ approaches to zero.

Comment: b_n need to be properly chosen.

Algebraic Method

because $\frac{1}{N}\sum_{i=1}^{N}[\underline{w}^T\underline{x}_i - y_i]^2 = \frac{1}{N}\|\underbrace{\underline{\underline{X}}}_{N\times(f+1)}\underbrace{\underline{w}}_{(f+1)\times 1}-\underbrace{\underline{y}}_{N\times 1}\|^2 = \frac{1}{N}(\underline{\underline{X}}\underline{w}-\underline{y})^T(\underline{\underline{X}}\underline{w}-\underline{y})$ [in OLS]

f: features number

$$\therefore J(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N}[\underline{w}^T z_n \underline{x}_n - b_n]^2 \quad , b_n > 0 \ \forall n$$

$$= \frac{1}{N}\|\underbrace{\underline{\underline{ZX}}}_{N\times(f+1)}\underbrace{\underline{w}}_{(f+1)\times 1}-\underbrace{\underline{b}}_{N\times 1}\|^2 = \frac{1}{N}(\underline{\underline{ZX}}\,\underline{w}-\underline{b})^T(\underline{\underline{ZX}}\,\underline{w}-\underline{b})$$

Minimize it : $\boxed{\underbrace{\hat{\underline{w}}}_{(f+1)\times 1} = (\underbrace{\underline{\underline{ZX}}^T\underline{\underline{ZX}}}_{(f+1)\times(f+1)})^{-1}\underbrace{\underline{\underline{ZX}}^T}_{(f+1)\times N}\underbrace{\underline{b}}_{N\times 1}}$ if $\underline{\underline{ZX}}^T\underline{\underline{ZX}}$ nonsingular

where $\underline{\underline{ZX}} = \begin{bmatrix} z_1X_{10} & z_1X_{11} & \cdots & z_1X_{1f} \\ z_2X_{20} & z_2X_{21} & \cdots & z_2X_{2f} \\ \vdots & \vdots & & \vdots \\ z_NX_{N0} & z_NX_{N1} & \cdots & z_NX_{Nf} \end{bmatrix}$

Gradient Descent Method

Ex: Sequential GD

$$J(\underline{w}) = \sum_{n=1}^{N} J_n(\underline{w})$$

$$\therefore J_n(\underline{w}) = \frac{1}{N}[\underline{w}^T z_n \underline{x}_n - b_n]^2$$

$$\nabla_{\underline{w}} J_n(\underline{w}) = \frac{2}{N}[\underline{w}^T z_n \underline{x}_n - b_n]\cdot z_n \underline{x}_n$$

weight update $\boxed{\underline{w}(i+1) = \underline{w}(i) - \eta(i)\cdot\frac{2}{N}[\underline{w}(i)^T z_n \underline{x}_n - b_n]\cdot z_n \underline{x}_n}$