Lec 5.1 Linear Regression

Recall: Supervised learning $\begin{cases} \text{discrete } y_n: & \text{classification} \\ \text{continuous } y_n: & \text{regression} \end{cases}$

1. Linear Regression setup:

↳ Training set:

$$\mathcal{D} = \{ (\underline{x}_1, y_1), (\underline{x}_2, y_2) \cdots (x_d, y_d) \}$$

$$\underline{x}_n \in \mathbb{R}^d, \quad y_n \in \mathbb{R}$$

↳ Decision Rule: (aka "Hypothesis set")

$$\hat{y} = h(\underline{x}) = w_0 + w_1 x_1 + \cdots + w_d x_d$$

Redefined augumented form: $\underline{w} = (w_0, w_1, \cdots w_d)$

$\underline{x} = (x_0 = 1, x_1, \cdots x_d)$

$$\therefore \hat{y} = h(\underline{x}) = w^T \underline{x}$$

h for hypothesis

↳ Criterion for learning:

$$E_{in}(\underline{w}) = \frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y})^2 = \frac{1}{N} \sum_{n=1}^{N} \underbrace{(y_n - w^T \underline{x})^2}_{e_n(\underline{w})}$$

"averaged squared error"

// $e_n(\underline{w})$: squared error on the n^{th} example

↳ Goal:

Given $\mathcal{D}$, find $\underline{w} \in \mathbb{R}^{d+1}$ to minimize $E_{in}(\underline{w})$

2. Example:

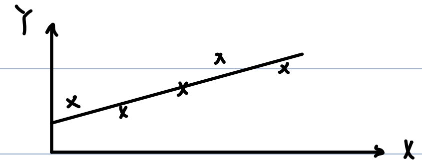
↳ Q: aim is to find the impact of advertisement on sales.

Let X = adv. cost in one week $(d = 1)$

y = sales in one week

$\mathcal{D} = \{(x_1, y_1), (x_2, y_2) \cdots (x_N, y_N)\}$

Find a linear model $y = w_0 + w_1 x$



↳ Refined Model:

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \text{adv on na of TV} \\ \text{adv on na of Radio} \\ \text{adv on na of Newspaper} \end{bmatrix} \quad (d = 3)$$

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$

larger $w_i \Rightarrow$ More profitable x_i

3. Algebra on how to solve:

↳ Data Matrix:

$$\underline{X} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix} \in \mathbb{R}^{N \times (d+1)} \quad \text{i.e.} \begin{bmatrix} x_{1,1} & \cdots & x_{1,d} \\ x_{2,1} & & x_{2,d} \\ \vdots & & \\ x_{N,1} & & x_{N,d} \end{bmatrix}$$

↳ Target vector:

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N$$

↳ Weight vector:

$$\underline{w} \in \mathbb{R}^{d+1}$$

↳ Linear Regression Model:

$$\underline{\hat{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} \underline{w}^T \underline{x}_1^T \\ \underline{w}^T \underline{x}_2^T \\ \vdots \\ \underline{w}^T \underline{x}_N^T \end{bmatrix} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix} \underline{w}$$

$$\therefore \boxed{\underline{\hat{y}} = \underline{\bar{X}}\, \underline{w}}$$

↳ Error:

$$\boxed{E_{in}(\underline{w}) = \frac{1}{N} \| \underline{y} - \underline{\hat{y}} \|^2}$$

- When is $E_{in}(\underline{w}) = 0$, i.e. $\underline{y} = \underline{\hat{y}}$?

$\because$ $y_n = \underline{w}^T \underline{x}_n$, for $n = 1, 2, \cdots N$

#. of linear equations: N

#. of variables : $d+1$

$\therefore$ In practice, $N >> d$ $\Rightarrow$ No exact solution

$\therefore$ Instead, we try to minimize $E_{in}(\underline{w})$

$\Rightarrow$ least square solution.

Lec 5.2 Least squares solution

↳ To minimize $E_{in}(\underline{w}) = \frac{1}{N} \|\underline{y} - \hat{\underline{y}}\|^2$

1. Least square sol:

↳ def: $f(\underline{w}) = \|\underline{y} - \hat{\underline{y}}\|^2$

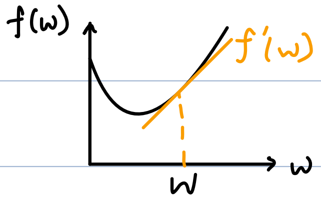
$$= \|\underline{X}\,\underline{w} - \underline{y}\|^2$$

$$= \sum_{n=1}^{N} (\underbrace{\underline{w}^T \underline{x}_n - y_n}_{e_n(\underline{w})})^2$$

↳ def: gradient of $f(\underline{w})$:

- $\nabla f(\underline{w}) = \begin{bmatrix} \partial f / \partial w_1 \\ \partial f / \partial w_2 \\ \vdots \\ \partial f / \partial w_d \end{bmatrix}$, It points in the direction of the steepest increase.

- If $\underline{w}$ is 1-dimension, only two direction: left / right



↳ Claims:

① $\boxed{\nabla f(\underline{w}) = 2\underline{X}^T(\underline{X}\,\underline{w} - \underline{y})}$

② Least squares solution $\underline{w}_{LS}$ is such that $\nabla f(\underline{w}_{LS}) = 0$

$\therefore\ 2\underline{X}^T(\underline{X}\,\underline{w} - \underline{y}) = 0$

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

Assume: The d+1 col of $\underline{X}$ are lin. indept.

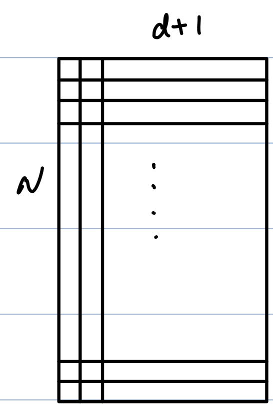
data samples

$\equiv \exists$ at least d+1 rows of $\underline{X}$ that are lin. indept.

$\therefore$ rank$(\underline{X}) = d+1 \iff \underline{X}^T\underline{X}$ is invertible

$\therefore\ \boxed{\underline{w}_{LS} = (\underline{X}^T\underline{X})^{-1}\underline{X}^T\underline{y}}$ // "w_{lin} in textbook"

$\underline{X}$:



↳ Pseudo - inverse of $\underline{\bar{X}}$:

$$\underline{\bar{X}}^{+} \triangleq (\underline{\bar{X}}^{T} \underline{\bar{X}})^{-1} \underline{\bar{X}}^{T}$$

Why we called "pseudo" inverse ?

- $\underline{\bar{X}}^{+} \underline{\bar{X}} = I$
- But $\underline{\bar{X}} \underline{\bar{X}}^{+} = \underline{\bar{X}} (\underline{\bar{X}}^{T} \underline{\bar{X}})^{-1} \underline{\bar{X}}^{T} \neq I$

Note:

① We have $\underline{y} = \underline{\bar{X}} \underline{w}$, and want to find $\underline{w}$

$\underline{w} = \boxed{?} \underline{y}$, where $\boxed{?}$ is kind of $\underline{\bar{X}}^{-1}$

But $\underline{\bar{X}}$ is not a square matrix $\Rightarrow$ Not invertible

$\therefore$ We use pseudo inverse to do instead

② $\underline{\hat{y}}_{LS} = \underline{\bar{X}} \underline{w}_{LS} = \underbrace{\underline{\bar{X}} \underline{\bar{X}}^{+}}_{\text{projection matrix from } \underline{y} \text{ to } \underline{\hat{y}}} \underline{y}$