Lec 7 Logistic Regression

1. Recap:

↳ **supervised learning set up:**

Given $\mathcal{D} = \{(\underline{x}_1, y_1)\ (\underline{x}_2, y_2) \cdots (\underline{x}_N, y_N)\}$

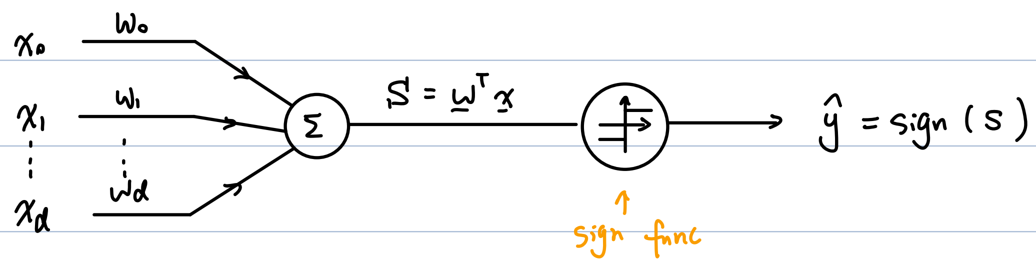
unknown target func. $y = f(\underline{x})$

Hypothesis: $\hat{y} = h(\underline{x})$ where $h \in \mathcal{H}$ // $\mathcal{H}$ is the set of hypothesis

↳ **Linear classification:**

- $x \in \mathbb{R}^{d+1}$, $y = \{+1, -1\}$

  $\hat{y} = \text{sign}(\underline{w}^T \underline{x})$

- 

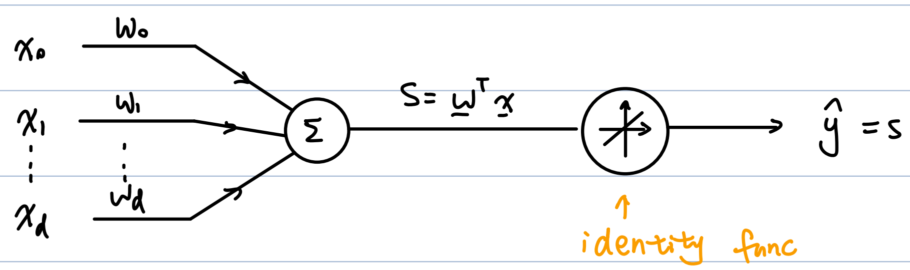
- The error for each of the sample is:

  $e(\underline{w}) = 1\{y \neq \text{sign}(s)\}$ // whether y is equal to estimation or not

↳ **Linear Regression:**

- $\underline{x} \in \mathbb{R}^{d+1}$, $y \in \mathbb{R}$

  $\underline{\hat{y}} = \underline{w}^T \underline{x}$ // compare w/ classification, this don't take the sign(·)

- 

- $e(\underline{w}) = (y - s)^2$

All above is deterministic hypothesis

2. What if we want randomness in h?

↳ example:

- $\underline{x}$ = average (fat, sugar) in diet.

  y = heart attack

- This is not determinstic. Nobody say this amount of fat & sugar will definitely have heart attack

- We want to say

  large $\underline{w}^T\underline{x} \Rightarrow$ more likely that $y=+1$

- ∴ Want new target function

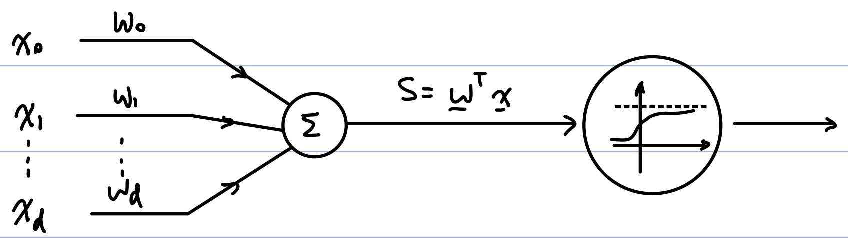
  $$f_p(\underline{x}) = \Pr\{y=+1 \mid \underline{x}\}$$

  " soft decision"

3. Logistic Regression:

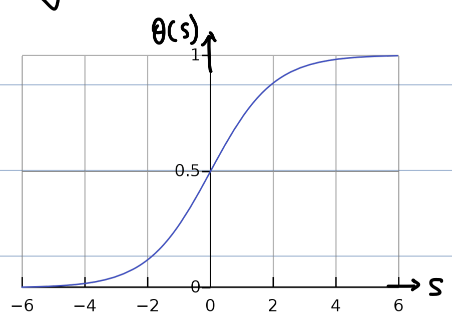
↳ Setup:

- $x \in \mathbb{R}^{d+1}$, $y \in \{+1, -1\}$

- 

  $\theta(s) = \dfrac{e^s}{1+e^s}$ ← defalt one

- $\theta(s)$: sigmoid function // "sigmoid" means S-shaped



// another $\theta(s)$: tanh

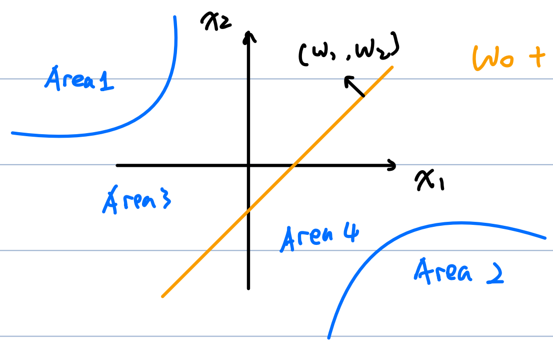
$$\tanh = \frac{e^s - e^{-s}}{e^s + e^{-s}}$$

↳ Hypothesis: (guess)

$$Pr\{y=+1 \mid \underline{x}\} = \theta(s) = \theta(\underline{w}^T \underline{x}) = \frac{e^{\underline{w}^T \underline{x}}}{1+e^{\underline{w}^T \underline{x}}}$$

↳ Note:

① Geometry: $(d=2)$



$\underline{w}^T \underline{x} = 0$

$w_0 + w_1 x_1 + w_2 x_2 = 0$

- Area 1: large $\|\underline{w}^T \underline{x}\|$ positive

$$\Rightarrow \theta(s) = \frac{e^{\underline{w}^T x}}{1+e^{\underline{w}^T x}} = \frac{\infty}{1+\infty} = 1$$

- Area 2: large $\|\underline{w}^T \underline{x}\|$, negative

$$\Rightarrow \theta(s) = \frac{e^{\underline{w}^T x}}{1+e^{\underline{w}^T x}} = \frac{0}{1+0} = 0$$

- Area 3: small $\|\underline{w}^T \underline{x}\|$, positive
- Area 4: small $\|\underline{w}^T \underline{x}\|$, negative

② $Pr\{y=-1 \mid \underline{x}\} = 1 - Pr\{y=+1 \mid \underline{x}\}$

$$= 1 - \frac{e^{\underline{w}^T \underline{x}}}{1+e^{\underline{w}^T \underline{x}}} = \frac{1}{1+e^{\underline{w}^T \underline{x}}} = \frac{e^{-\underline{w}^T \underline{x}}}{1+e^{-\underline{w}^T \underline{x}}} = \theta(-\underline{w}^T \underline{x}) = \theta(-s)$$

4. loss function of logistic regression:

↳ Notation:

- $\underbrace{\hat{P}_{\underline{w}}(1 \mid \underline{x})}_{\text{the estimate for } \Pr\{y=+1 \mid \underline{x}\}} = \theta(\underline{w}^T \underline{x})$ // $\hat{P}_{\underline{w}}$ means is a func. of $\underline{w}$

  $\hat{P}_{\underline{w}}(-1 \mid \underline{x}) = \theta(-\underline{w}^T \underline{x})$

- Combine the two above:

$$\hat{P}_{\underline{w}}(y \mid \underline{x}) = \theta(y\,\underline{w}^T \underline{x}) \quad , \quad \text{for } y \in \{+1, -1\}$$

$$= \frac{e^{y \underline{w}^T \underline{x}}}{1 + e^{y \underline{w}^T \underline{x}}}$$

$$= \frac{1}{1 + e^{-y \underline{w}^T \underline{x}}}$$

↳ Error criterion:

- For nth sample in training data, we define the error

$$\boxed{\begin{aligned} e_n(\underline{w}) &= -\log\left[\hat{P}_{\underline{w}}(y_n \mid \underline{x}_n)\right] \\ &= \log\left(1 + e^{-y \underline{w}^T \underline{x}_n}\right) \end{aligned}}$$

// log loss function

↳ Note:

- If $\underline{w}^T \underline{x} \gg 0$ (i.e. very far away from the line). and $y = +1$ (true label)

  $\Rightarrow -y\,\underline{w}^T \underline{x} = -\infty \quad \Rightarrow e^{-y \underline{w}^T \underline{x}} = 0$

  $\Rightarrow$ loss $= \log(1 + 0) = 0$

- If $\underline{w}^T \underline{x} \ll 0$, and $y = -1 \Rightarrow$ loss $= 0$

J. Example:

Q: suppose output of logistic regression is

$$\hat{P}_{\underline{w}}(-1 \mid \underline{x}_n) = 0.8 \qquad \hat{P}_{\underline{w}},(+1 \mid \underline{x}_n) = 0.2$$

if $y_n = +1$, $e_n(\underline{w}) = -\log(0.2) \approx 1.61$

if $y_n = -1$, $e_n(\underline{w}) = -\log(0.8) \approx 0.22$

↳ Note: Loss is smaller if $y_n = -1$

Because our $\hat{P}_{\underline{w}}$ assigns higher prob. for $y_n = -1$

Suppose $\hat{P}_{\underline{w}}(-1 \mid \underline{x}_n) = 0.999$, $\hat{P}_{\underline{w}}(+1 \mid \underline{x}_n) = 0.001$,

if $y_n = +1$, $e_n(\underline{w}) = -\log(0.001) = 10$

if $y_n = -1$, $e_n(\underline{w}) = -\log(0.999) = 10^{-4}$

Summary: Logic Regression

Given $\mathcal{D} = \{(\underline{x}_1, y_1)(\underline{x}_2, y_2) \cdots (\underline{x}_N, y_N)\}$, find $\underline{w} \in \mathbb{R}^{d+1}$,

to minimize

$$E_{in}(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N} e_n(\underline{w})$$
$$= \frac{1}{N}\sum_{n=1}^{N} \log(1 - e^{-y_n \underline{w}^T \underline{x}_n})$$