

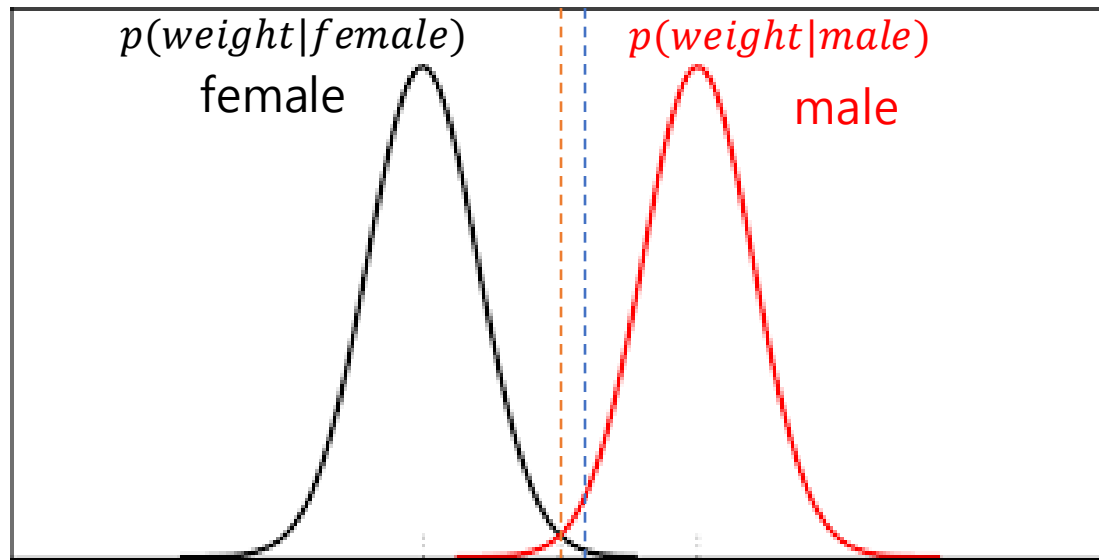
# Logistic Regression

AI/ML Teaching

# Goals

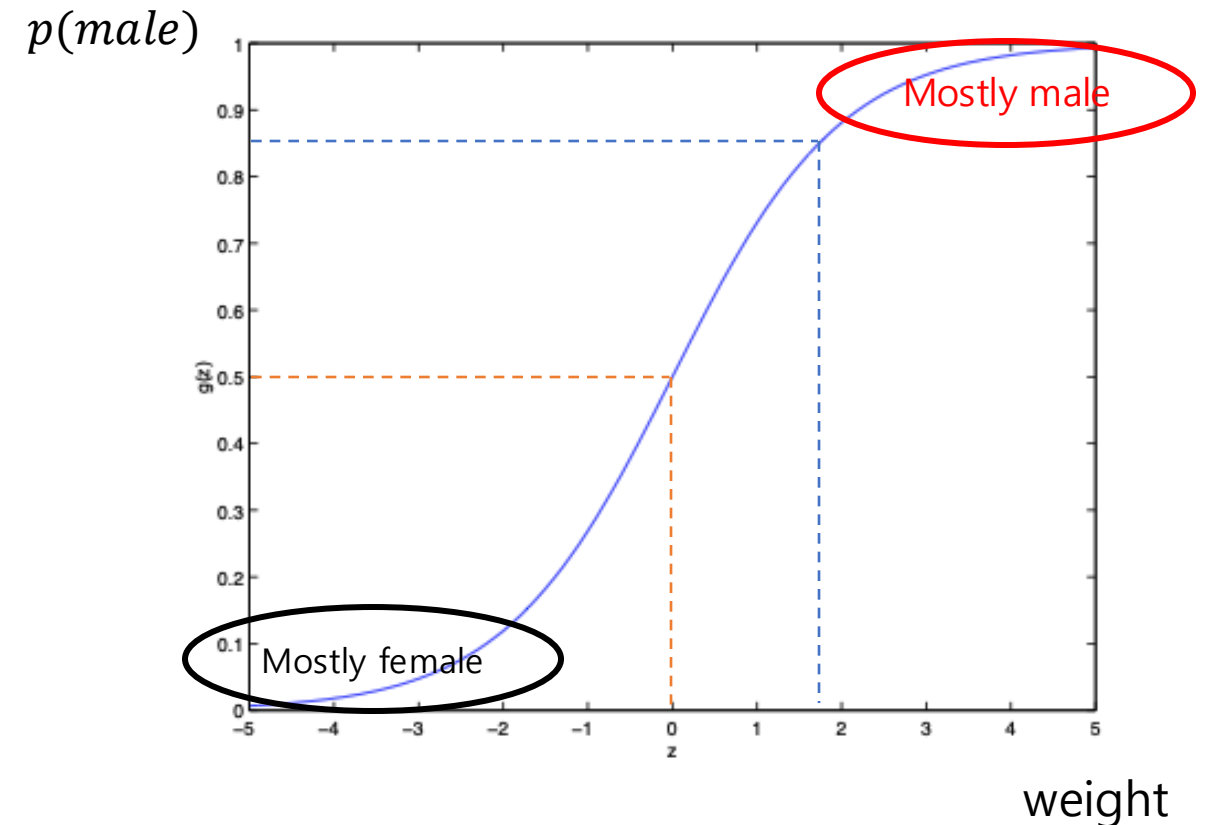
- Basic concept of logistic regression: linear vs logistic
- Binary classification & binary cross entropy loss
- Multi-class classification & cross-entropy loss

# Binary classification example



\*assume prior as 1/2

$$p(\text{male}|\text{weight}) = \frac{p(\text{weight}|\text{male})}{p(\text{weight}|\text{male}) + p(\text{weight}|\text{female})}$$



# Modeling conditional probabilities

- Linear model

- $0 \leq h_{\theta}(x) \leq 1$

- $0 \leq \frac{h_{\theta}(x)}{1-h_{\theta}(x)} < \infty$

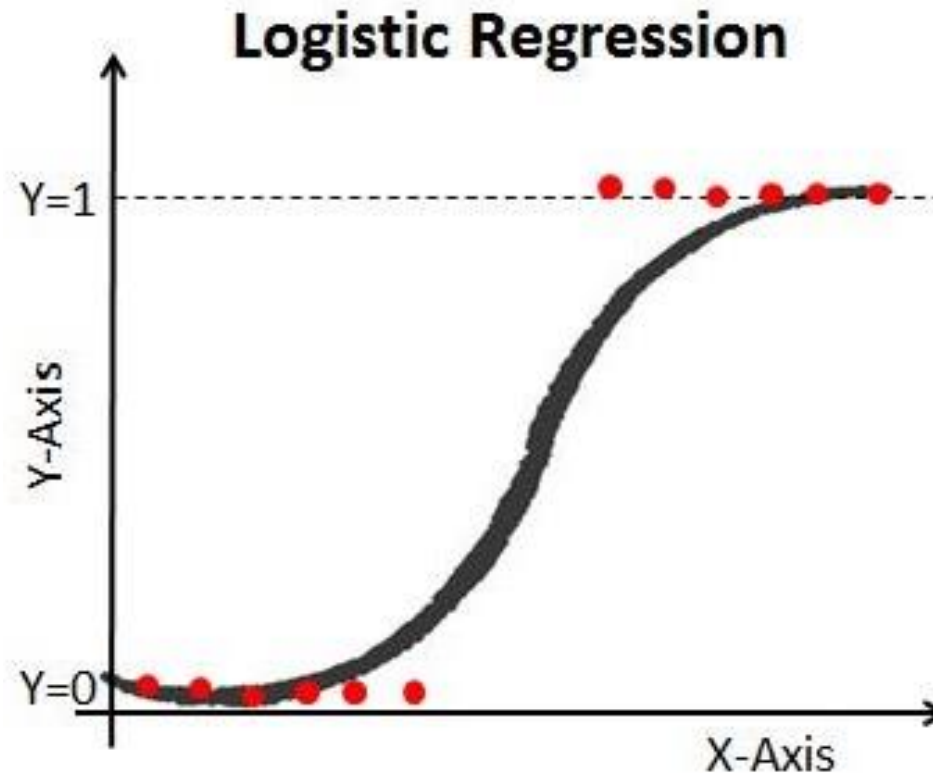
- $-\infty \leq \log \frac{h_{\theta}(x)}{1-h_{\theta}(x)} < \infty$

- $\log \frac{h_{\theta}(x)}{1-h_{\theta}(x)} = \theta^T x$

- $h_{\theta}(x) = \frac{1}{1+\exp -\theta^T x}$

- $g(z) \rightarrow 1$  as  $z \rightarrow \infty$

- $g(z) \rightarrow 0$  as  $z \rightarrow -\infty$



# Maximum likelihood & BCE loss

$$P(y = 1 \mid x; \theta) = h_{\theta}(x)$$

$$P(y = 0 \mid x; \theta) = 1 - h_{\theta}(x)$$

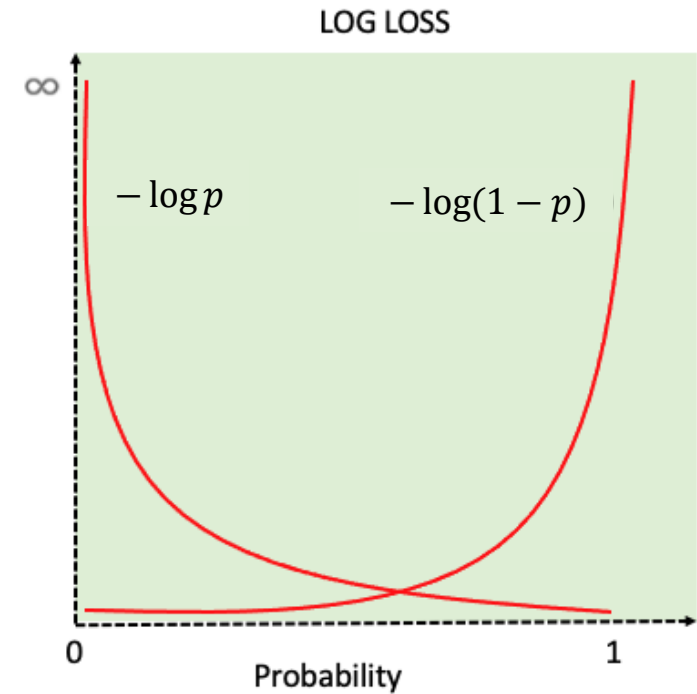
$$p(y \mid x; \theta) = (h_{\theta}(x))^y (1 - h_{\theta}(x))^{1-y}$$

$$L(\theta) = p(\vec{y} \mid X; \theta)$$

$$= \prod_{i=1}^n p(y^{(i)} \mid x^{(i)}; \theta)$$

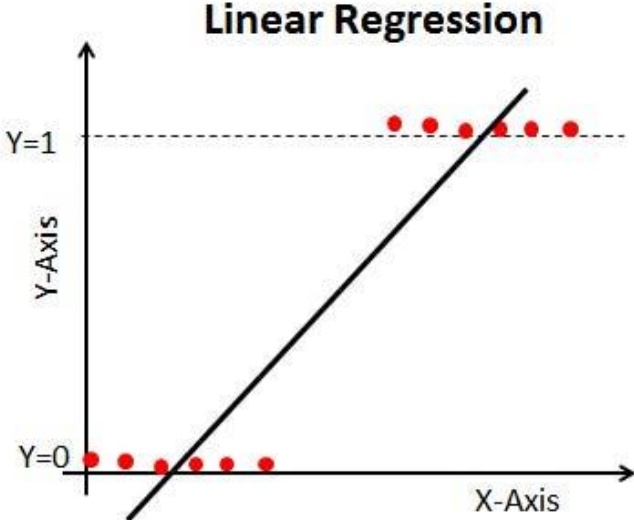
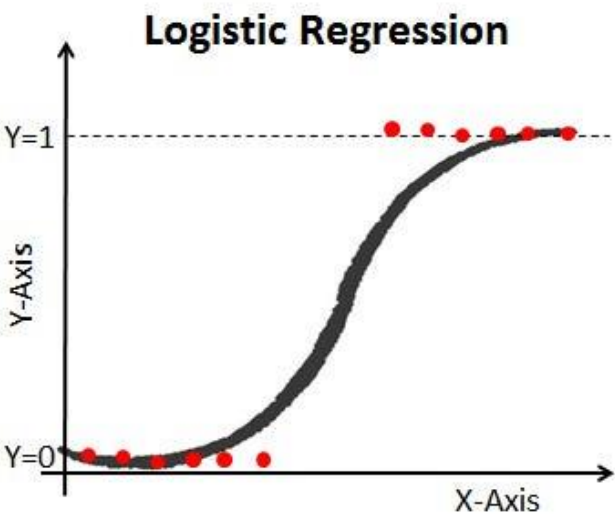
$$= \prod_{i=1}^n (h_{\theta}(x^{(i)}))^{y^{(i)}} (1 - h_{\theta}(x^{(i)}))^{1-y^{(i)}}$$

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^n y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))$$



# Linear regression & logistic regression

- Doesn't make sense for  $h_{\theta}(x)$  to take values larger than 1 or smaller than 0

Linear regression	Logistic regression
$h_{\theta}(x) = \theta^T x$	$h_{\theta}(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$
 <p>The plot titled "Linear Regression" shows a 2D coordinate system with a vertical Y-Axis and a horizontal X-Axis. A solid black line with a positive slope represents the linear hypothesis. Red dots representing data points are clustered at two horizontal levels: one near the X-axis (labeled Y=0) and another higher up (labeled Y=1). Dashed horizontal lines extend from these labels on the Y-axis. The linear line passes through the middle of the data points, illustrating that it can predict values outside the [0, 1] range.</p>	 <p>The plot titled "Logistic Regression" shows a 2D coordinate system with a vertical Y-Axis and a horizontal X-Axis. A solid black S-shaped curve (sigmoid function) represents the logistic hypothesis. Red dots representing data points are clustered at two horizontal levels: one near the X-axis (labeled Y=0) and another higher up (labeled Y=1). Dashed horizontal lines extend from these labels on the Y-axis. The sigmoid curve stays within the [0, 1] range, fitting the data points better than the linear model.</p>

# Multi-class classification: softmax

$$\begin{bmatrix} P(y = 1 | x; \theta) \\ \vdots \\ P(y = k | x; \theta) \end{bmatrix} = \text{softmax}(t_1, \dots, t_k) = \begin{bmatrix} \frac{\exp(\theta_1^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)} \\ \vdots \\ \frac{\exp(\theta_k^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)} \end{bmatrix}$$

## Sigmoid

2 classes

$$\text{out} = P(Y=\text{class1}|X)$$

## SoftMax

k>2 classes

$$\text{out} = \begin{bmatrix} P(Y=\text{class1}|X) \\ P(Y=\text{class2}|X) \\ P(Y=\text{class3}|X) \\ \vdots \\ P(Y=\text{classk}|X) \end{bmatrix}$$

X

$$\begin{bmatrix} 3 \\ 1.75 \\ -2 \\ 0.5 \end{bmatrix}$$

Input vector

(logit)

SoftMax

$$\frac{e^{x_i}}{\sum_{j=1}^K e^{x_j}}$$

Out

$$\begin{bmatrix} 0.725 \\ 0.21 \\ 0.005 \\ 0.06 \end{bmatrix}$$

Output vector

Probability  
distribution

# Cross entropy loss

$$\begin{bmatrix} P(y = 1 \mid x; \theta) \\ \vdots \\ P(y = k \mid x; \theta) \end{bmatrix} = \text{softmax}(t_1, \dots, t_k) = \begin{bmatrix} \frac{\exp(\theta_1^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)} \\ \vdots \\ \frac{\exp(\theta_k^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)} \end{bmatrix}$$

- $H(p, q) = -\sum_i p_i \log q_i$
- $p_i \in \{0,1\}$ : label (e.g.,  $\mathbf{p} = [0,1,0, \dots, 0]^T$ )
- $q_i = \frac{\exp(\theta_i^T x)}{\sum_j \exp(\theta_j^T x)}$



# BCE & cross-entropy loss

Binary classification	Multi-class classification
sigmoid	softmax
BCE	Cross-entropy loss
Maximum likelihood Estimation (MLE)	

# Reference

- Andrew Ng, CS229 Stanford Lecture Notes
- Cosma Shalizi, 36-402 CMU Lecture Notes