Logistic Regression

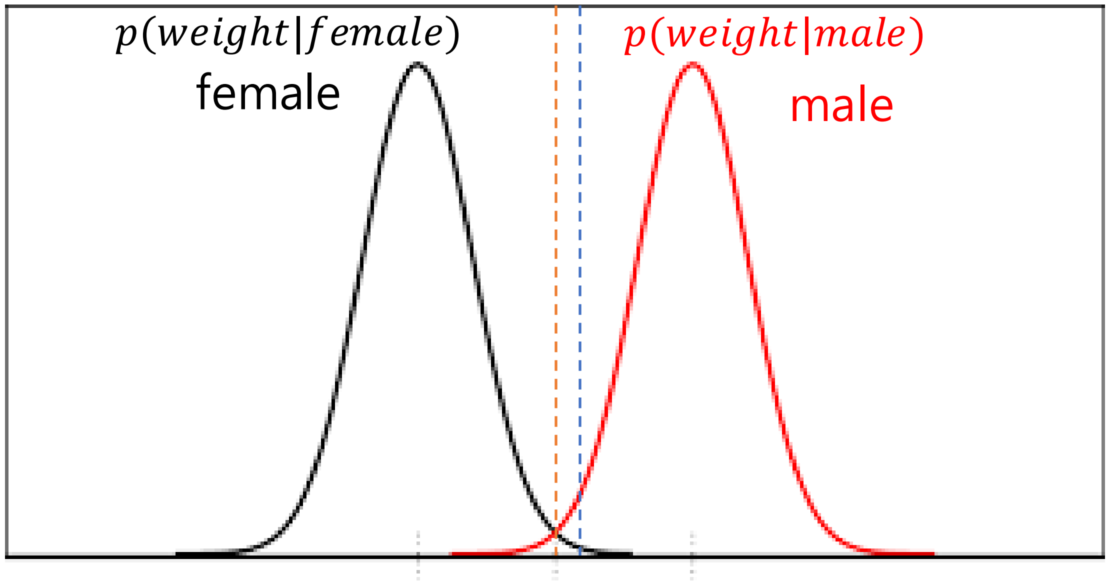
AI/ML Teaching

Goals

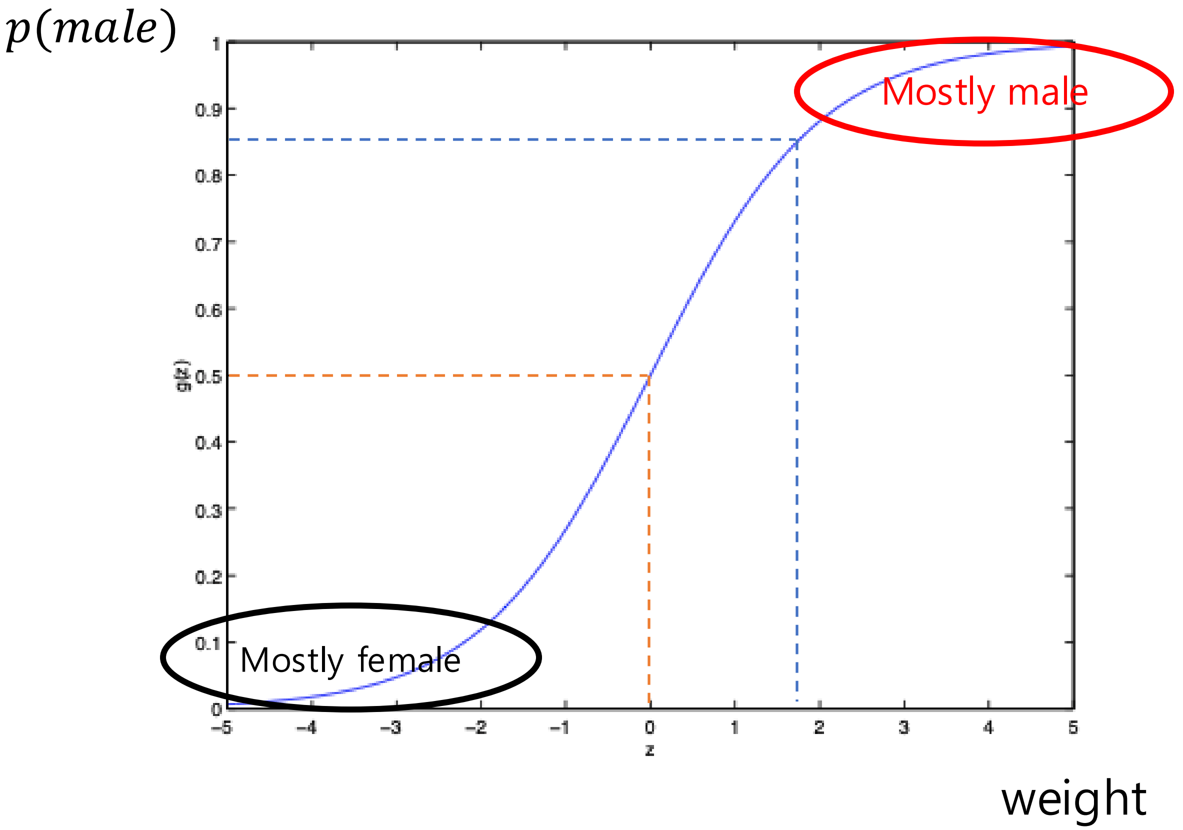
- Basic concept of logistic regression: linear vs logistic
- Binary classification & binary cross entropy loss
- Multi-class classification & cross-entropy loss

Binary classification example

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



*assume prior as 1/2

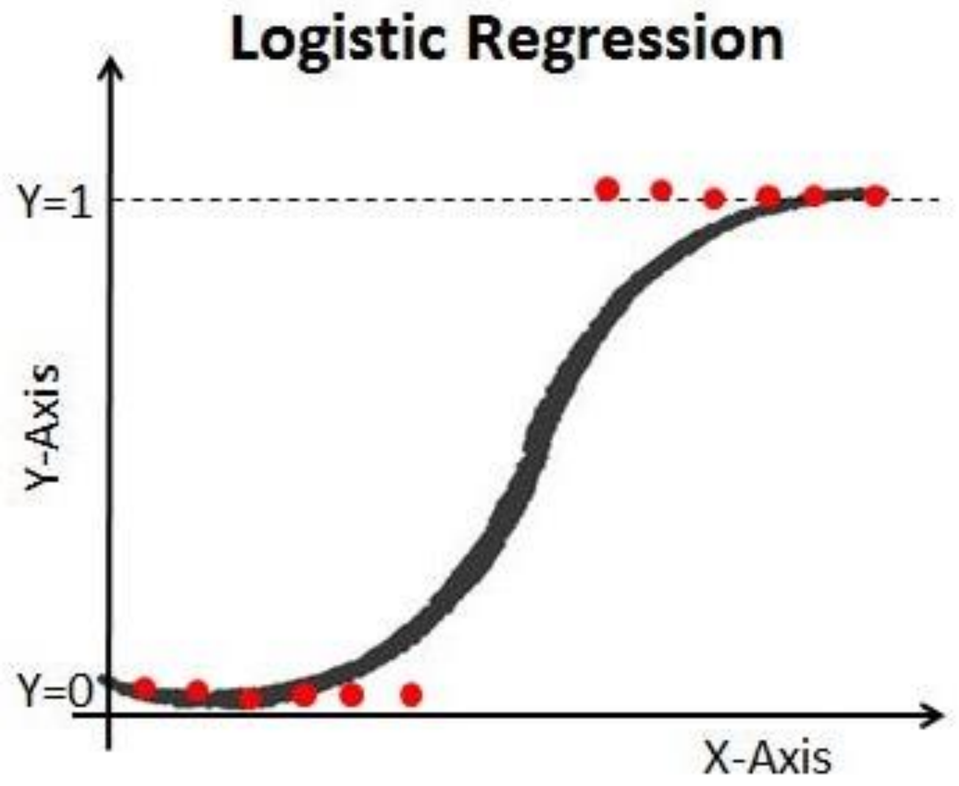


Modeling conditional probabilities

- 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 $z \rightarrow -\infty$



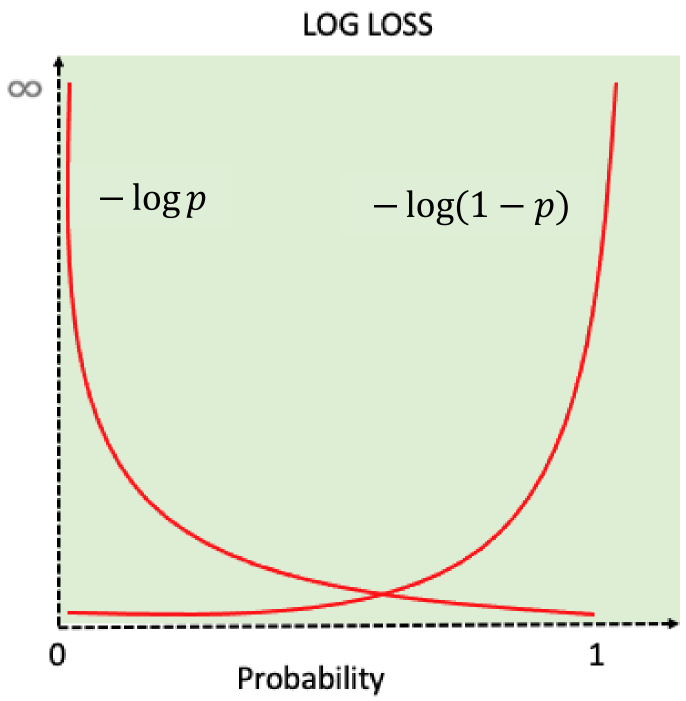
Maximum likelihood & BCE loss

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

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

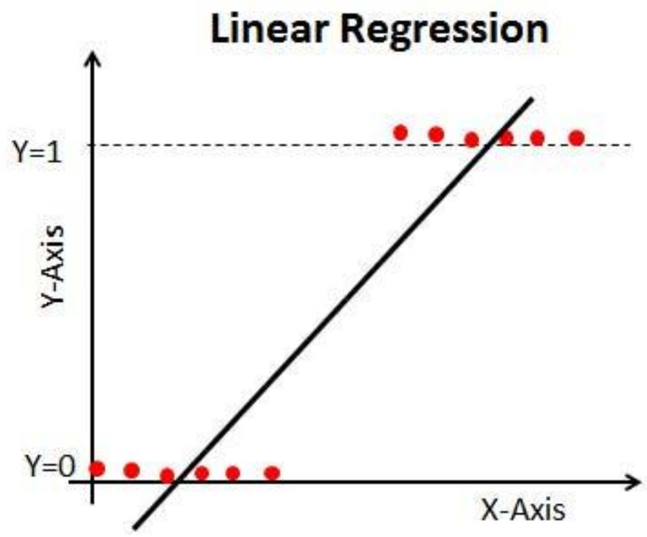
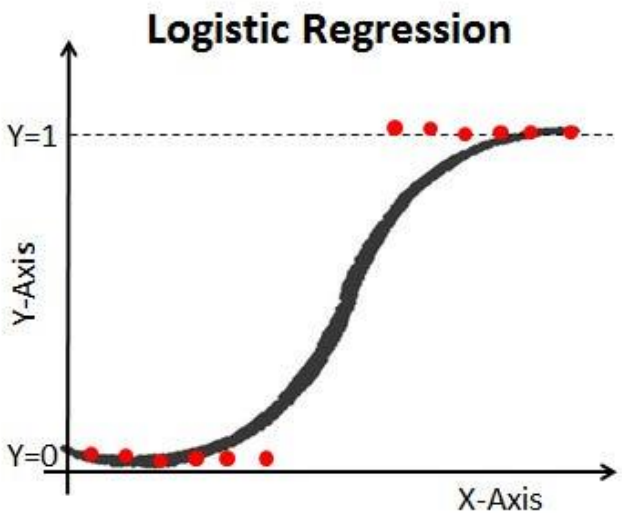
$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X;\theta) \\ &= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \prod_{i=1}^{n} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1-h_\theta(x^{(i)})\right)^{1-y^{(i)}} \end{aligned}$$

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



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

Multi-class classification: softmax

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

Sigmoid

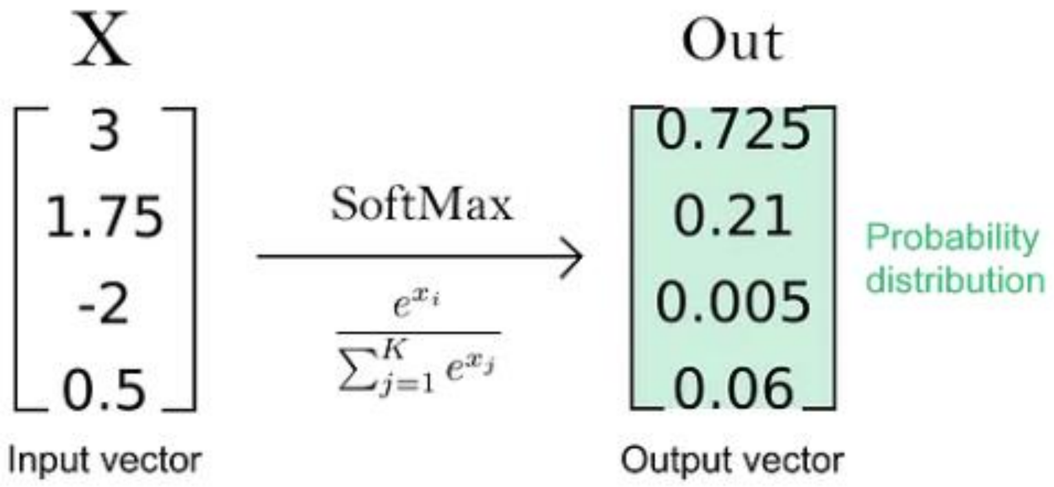
2 classes

out = P(Y=class1|X)

SoftMax

k>2 classes

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



(logit)

Cross entropy loss

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

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

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