[SWCON253] Machine Learning – Lec.08

Logistic Regression

Fall 2025

김 휘 용

hykim.v@khu.ac.kr

경희대학교
KYUNG HEE UNIVERSITY

Contents

1. Extending Linear Regression for Binary Classification
2. Logistic Regression

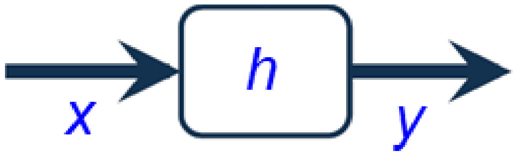
References

- *Machine Learning* by Andrew Ng, Coursera (https://www.coursera.org/learn/machine-learning)

1. Extending Linear Regression for Binary Classification

Extending Linear Regression for Binary Classification

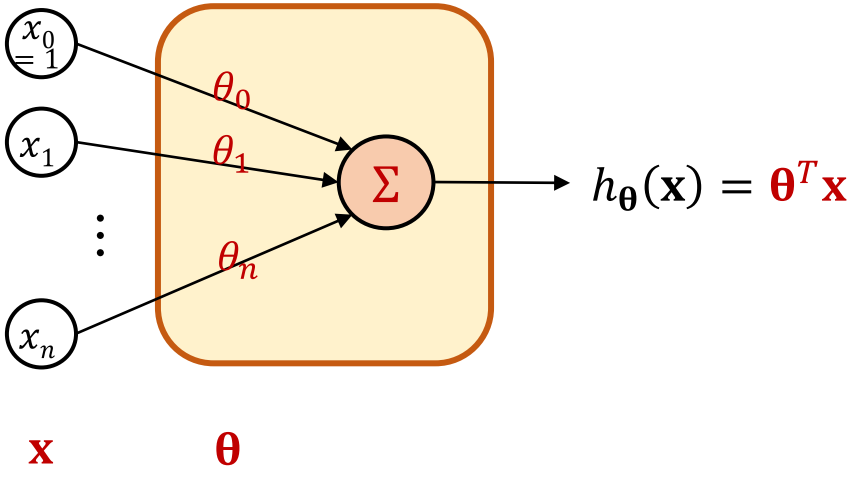
- ***Binary Classification*** **Problems:** $y \in \{0, 1\}$ or $y \in \{-1, 1\}$
  - Examples:
    - ★ *Email*: Spam / Non-Spam
    - ★ *Tumor*: Malignant / Benign
    - ★ *Housing Price*: Cheap / Expensive



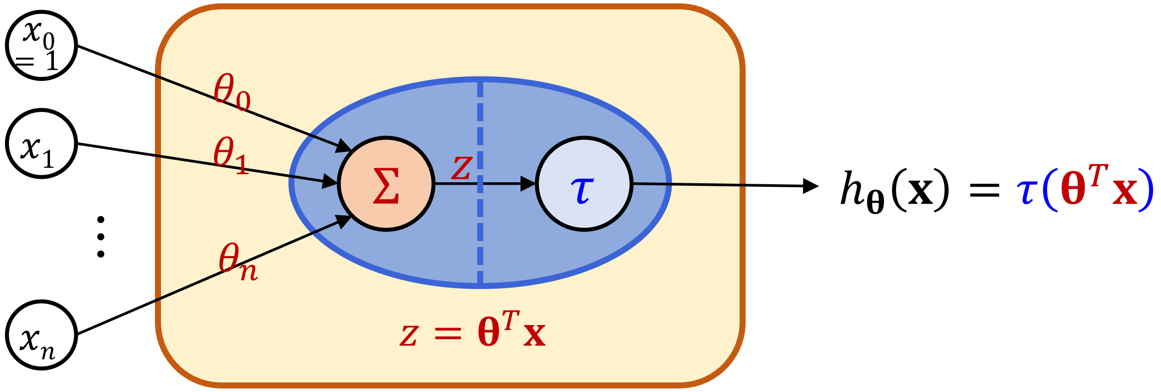
Extending Linear Regression for Binary Classification (cont'd)

◆ For *Binary classification*, we can extend *Linear regression* by adding some ***Activation function*** $\tau()$.

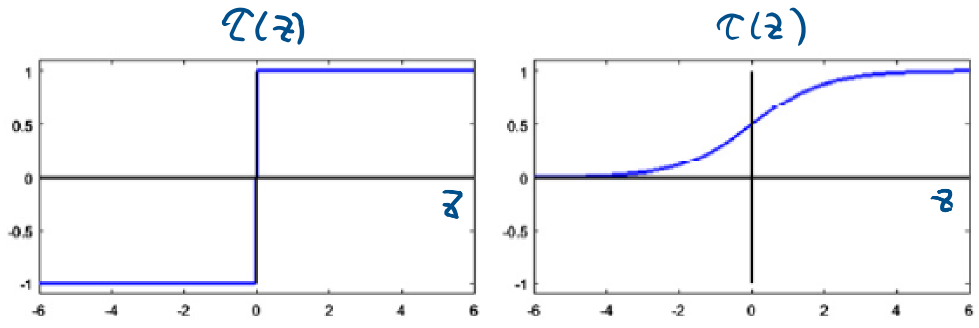
Linear Regression



Linear Regression + Threshold function



- Examples of $\boldsymbol{\tau}(z)$
  - ★ ***step*** function: ***hard*** threshold
  - ★ ***sigmoid*** function: ***soft*** threshold



(a) 계단 함수 (hard) ~> Perceptron
(b) 로지스틱 시그모이드 (soft) ~> Logistic Regression

2. Logistic Regression

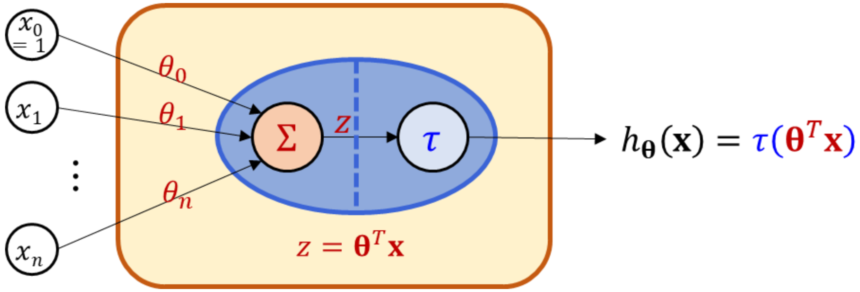
- ✓ Model Representation
- ✓ Cross-Entropy Loss
- ✓ Gradient

Logistic Regression – *Model Representation*

Logistic Regression Model

$$z = \boldsymbol{\theta}^T \boldsymbol{x}$$ ← linear model

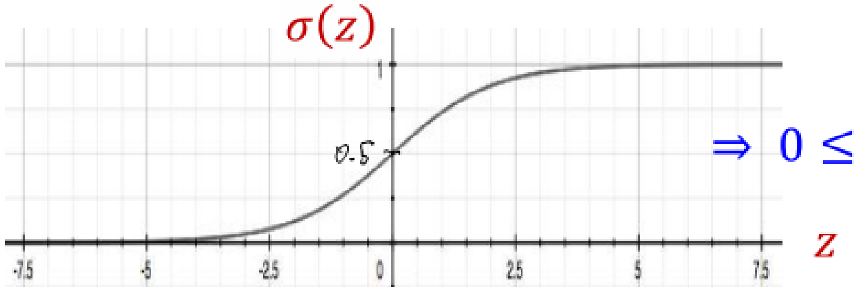
$$h_\theta(z) = \sigma(z) = \frac{1}{1+e^{-z}}$$ ← logistic function (sigmoid function)



$$z = \underline{\theta}^T \underline{x} = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n = \sum_{i=0}^{n} \theta_i x_i$$

$$y = \sigma(z)$$

- $\sigma(z)$: ***Sigmoid*** function (*soft-threshold*):



$\Rightarrow 0 \le h_\theta(\boldsymbol{x}) \le 1$

Thus, $h_\theta(\boldsymbol{x})$ will give us the probability that our output is 1.

★ E.g., for binary classification:

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

$$P(y=0|x;\theta) + P(y=1|x;\theta) = 1$$

★ $h_{\underline{\theta}}(\underline{x}) = \sigma(\underline{\theta}^T \underline{x}) = \dfrac{1}{1+e^{-\underline{\theta}^T \underline{x}}}$

$\left(\text{cf. } \dfrac{1}{1+e^{-z}} = \dfrac{e^z}{1+e^z}\right)$

Logistic Regression – *Cost Function: MSE?*

- **We will not use MSE for logistic regression**
  - The logistic sigmoid will cause the output to be *wavy*, causing many local minima.
  - In other words, the MSE cost will *not* be a *convex* function.

- **Instead, we will use *Cross-Entropy* (CE) Loss**
  - For use of CE Loss, the target values should be $y \in \{0,1\}$.

Cross-Entropy Loss

- **Cross-Entropy (CE)** Loss: used widely in binary classification

* NOTE : "log" here denotes "natural logarithm" (i.e. base = e)

- **Classic form:**

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(\underline{x}^{(i)})) + (1-y^{(i)})\log(1-h_\theta(\underline{x}^{(i)}))]$$

- **Vector form:**

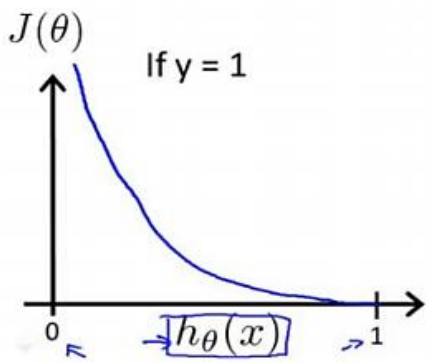
$$J(\boldsymbol{\theta}) = -\frac{1}{m}(\boldsymbol{y}^T\log(\boldsymbol{h}) + (\mathbf{1}-\boldsymbol{y})^T\log(\mathbf{1}-\boldsymbol{h}))$$

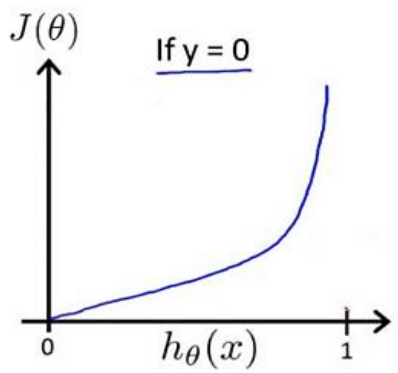
- The CE loss is *convex* for $0 \le h_{\boldsymbol{\theta}}(\mathbf{x}) \le 1$.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(h_\theta(\underline{x}^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(\underline{x}), y) = -\log(h_\theta(\underline{x})) \qquad \text{if } y = 1$$
$$\mathrm{Cost}(h_\theta(\underline{x}), y) = -\log(1-h_\theta(\underline{x})) \qquad \text{if } y = 0$$

$$\Rightarrow \mathrm{Cost}(h_\theta(\underline{x}), y) = -y\log(h_\theta(\underline{x})) - (1-y)\log(1-h_\theta(\underline{x}))$$





Logistic Regression – *Gradient of CE Loss*

- ***Gradients*** **of Cross-Entropy Loss for Logistic Regression**

  - The CE Loss: $$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(\underline{x}^{(i)})) + (1-y^{(i)})\log(1-h_\theta(\underline{x}^{(i)}))]$$

    $$J(\boldsymbol{\theta}) = -\frac{1}{m}(\boldsymbol{y}^T\log(\boldsymbol{h}) + (\boldsymbol{1}-\boldsymbol{y})^T\log(\boldsymbol{1}-\boldsymbol{h}))$$

  - The LR Model: $$\boxed{h_\theta(\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^T\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}}$$

➡ ***Gradient***:

★ **Classic form:** $$\frac{\partial J(\boldsymbol{\theta})}{\partial\theta_j} = \frac{1}{m}\sum_{i=1}^{m}\left(\sigma\left(\boldsymbol{\theta}^T\mathbf{x}^{(i)}\right) - y^{(i)}\right)x_j^{(i)}$$ for j=0,…,n

★ **Vector form:** $$\nabla J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\left(\sigma\left(\boldsymbol{\theta}^T\mathbf{x}^{(i)}\right) - y^{(i)}\right)\mathbf{x}^{(i)} = \frac{1}{m}\mathbf{X}^T(\sigma(\mathbf{X}\boldsymbol{\theta}) - \mathbf{y})$$

Logistic Regression – *Gradient Descent* (cont'd)

- **Proof**
  - Cross-Entropy Loss:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(\underset{\sim}{x}^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(\underset{\sim}{x}), y) = -y\ \log(h_\theta(\underset{\sim}{x})) - (1-y)\log(1-h_\theta(\underset{\sim}{x}))$$

  - Partial derivatives:

$$\frac{\partial \mathrm{Cost}}{\partial \theta_j} = -y\frac{1}{h_\theta(\underset{\sim}{x})}\cdot\frac{\partial h_\theta(\underset{\sim}{x})}{\partial\theta_j} + (1-y)\frac{1}{1-h_\theta(\underset{\sim}{x})}\cdot\frac{\partial h_\theta(\underset{\sim}{x})}{\partial\theta_j}$$

$$= \left(-y\frac{1}{h_\theta(\underset{\sim}{x})} + (1-y)\frac{1}{1-h_\theta(\underset{\sim}{x})}\right)\cdot\frac{\partial h_\theta(\underset{\sim}{x})}{\partial\theta_j}$$

$$= -y(1-h_\theta(\underset{\sim}{x}))\cdot x_j + (1-y)\cdot h_\theta(\underset{\sim}{x})\cdot x_j$$

$$= -y\cdot x_j + h_\theta(\underset{\sim}{x})\cdot x_j = \boxed{(h_\theta(\underset{\sim}{x}) - y)\cdot x_j}$$

$$z = \underset{\sim}{\theta}^T\underset{\sim}{x}, \quad \sigma(z) = \frac{1}{1+e^{-z}}$$

$$\frac{\partial z}{\partial\theta_j} = x_j$$

$$\frac{\partial\sigma(z)}{\partial z} = \frac{e^{-z}}{(1+e^{-z})^2}$$

$$= \frac{1}{(1+e^{-z})}\cdot\left(1-\frac{1}{(1+e^{-z})}\right)$$

$$= \sigma(z)\cdot(1-\sigma(z))$$

$$\Rightarrow \frac{\partial h_\theta(\underset{\sim}{x})}{\partial\theta_j} = \frac{\partial\sigma(\overbrace{\underset{\sim}{\theta}^T\underset{\sim}{x}}^{z})}{\partial\theta_j} = \frac{\partial\sigma(z)}{\partial z}\cdot\frac{\partial \overbrace{z}^{\theta^T x}}{\partial\theta_j}$$

$$= \sigma(z)\cdot(1-\sigma(z))\cdot x_j$$

$$= h_\theta(\underset{\sim}{x})\cdot(1-h_\theta(\underset{\sim}{x}))\cdot x_j$$

Q & A