Spring 2023

SWCON253: Machine Learning

Lecture 05
Logistic Regression 인데 분류 문제

Jinwoo Choi
Assistant Professor
CSE, Kyung Hee University

Disclaimer: Many slides have been borrowed from Hui Yong Kim, who may have borrowed some of them from others. Any time a slide does not show any explicit credit, I credit the slide to Hui Yong Kim.

Contents

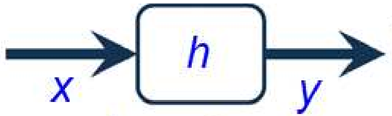
References

- *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 이진 분류

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



입력벡터

↓Binary

{0, 1}

or → 출력값이 2가지(종류)

{-1, 1}

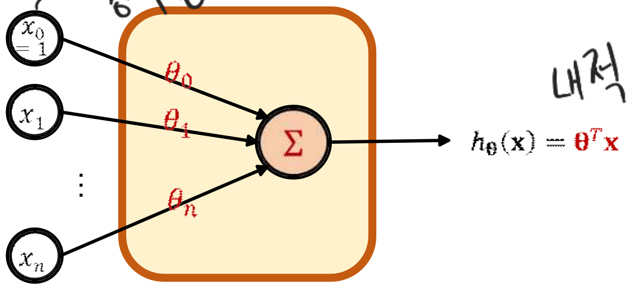
Extending Linear Regression for Binary Classification (cont'd)

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

활성함수가 logistic sigmoid

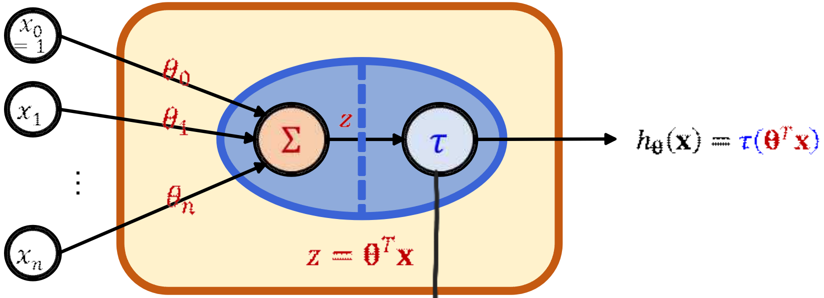
Linear Regression

상수항



내적

Linear Regression + Threshold function

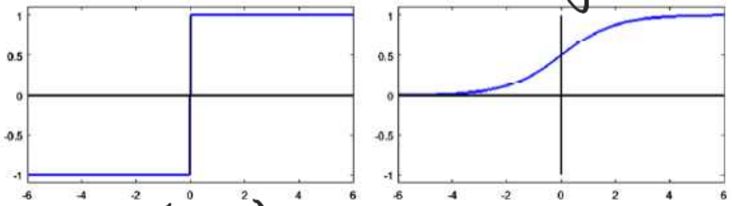


→ Logistic Sigmoid

$\mathbf{x}$ $\boldsymbol{\theta}$

특징벡터 가중치

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



y의 range [0, 1]

(a) 계단 함수 (Sign) (b) 로지스틱 시그모이드

→ Perceptron → logistic Regression

2. Logistic Regression

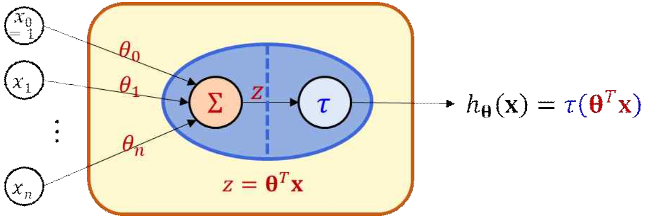
1. Model Representation
2. Cross-Entropy Loss
3. Gradient

Logistic Regression – *Model Representation*

◆ Logistic Regression Model

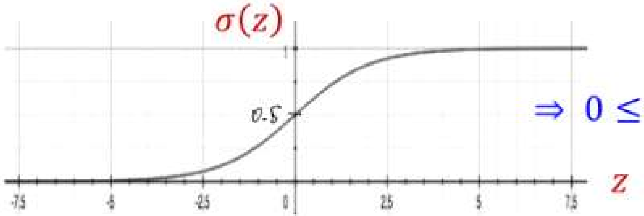
$$z = \theta^T x$$ ⟵ linear model

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



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

$\sigma(z)$



$\Rightarrow 0 \le h_\theta(x) \le 1$ → 확률로 볼 수 있음

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

★ E.g., for binary classification:

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

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

ex) 0.5 이상 → 사람
미만 → 사람X

$p(y=1|x;\theta)$ x를 넣었을 때 y가 1일 확률

Logistic Regression – *Cost Function: MSE?*

θ를 그래서 어떻게 구할거냐?

- **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.

분류 문제지 회귀 문제가 아님

- **Instead, we will use *Cross-Entropy* (CE) Loss**

Cross-Entropy Loss

- Cross-Entropy (CE) Loss: used widely in binary classification $\underline{h_\theta(x)}$ 확률

트레이닝 샘플의 개수

y=1 → 사람
y=0 → 사람X

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

⇒ m개의 평균

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

⇒ m개의 샘플을 한꺼번에 처리

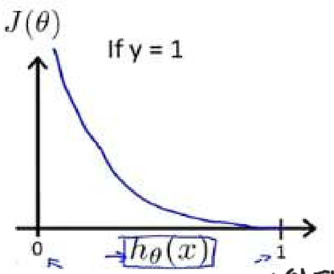
$J = -\log h$ (샘플 1개일때) $J = -\log(1-h)$

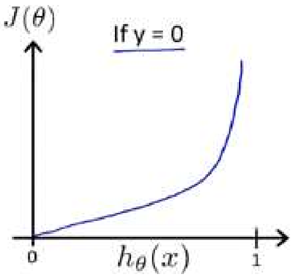
- The CE loss is convex for $0 \le h_\theta(\mathbf{x}) \le 1$.

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

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

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





틀렸으니까 ∞

y=1 일때 ŷ=1인 Loss → 정답이니까 0

Logistic Regression – *Gradient of CE Loss*

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

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

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

  - [illegible] $$\boxed{h_\theta(\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^T\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}}$$

➡ ***Gradient***:

$$\frac{\partial J(\boldsymbol{\theta})}{\partial\theta_j} = \frac{1}{m}\sum_{i=1}^{m}\left(\sigma(\boldsymbol{\theta}^T\mathbf{x}^{(i)}) - y^{(i)}\right)x_j^{(i)} \quad \text{for } j=0,\dots,n$$

★ **Classic form:**

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

★ **Vector form:**

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

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \qquad \sigma'(z) = \sigma(z)(1-\sigma(z))$$

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

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

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

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

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

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

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

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

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

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

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

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

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

$$= \underline{\sigma(z)\cdot(1-\sigma(z))}\cdot x_j$$

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