

# 인공지능을 위한 머신러닝 알고리즘

## 3. 로지스틱 회귀 모델

# CONTENTS

1

로짓(logit) 함수

2

로지스틱 회귀 모델

3

로지스틱 회귀 모델의 파라미터 추정

# 학습 목표

- 로짓(logit) 함수를 이해할 수 있다.
- 로지스틱 회귀의 분류 원리에 대해 이해할 수 있다.
- 로지스틱 회귀 모델의 파라미터를 구할 수 있다.

A person's hands are shown holding a smartphone, with the screen glowing. The background is dark with out-of-focus, warm-toned lights (bokeh) in shades of yellow and orange. A semi-transparent dark banner is at the bottom, containing a yellow decorative mark and the title text.

## 1. 로짓(logit) 함수

## 오즈 (odds)

어떠한 사건의 확률이  $p$ 일 때, 그 사건의 오즈는 다음과 같이 계산

$$\text{odds} = p / (1-p)$$

예시

		비만		
		Yes	No	Total
혈중 콜레스테롤	Normal	402	3614	4016
	High	101	345	446
		503	3959	4462

혈중 콜레스테롤이 정상인 그룹에서 비만인 경우의 오즈

$$\text{비만일 확률} / (1 - \text{비만일 확률}) = (402/4016) / (1 - (402/4016)) = 0.1001 / 0.8889 = 0.111$$

## ■ 오즈 (odds)

◉ 혈중 콜레스테롤이 정상인 그룹에서 비만이 아닌 경우의 오즈

▪  $0.8999/0.1001 = 8.99$

◉ 혈중 콜레스테롤이 높은 그룹

▪  $\text{odds}(\text{비만}) = 101/345 = 0.293$

▪  $\text{odds}(\text{비만이 아님}) = 345/101 = 3.416$

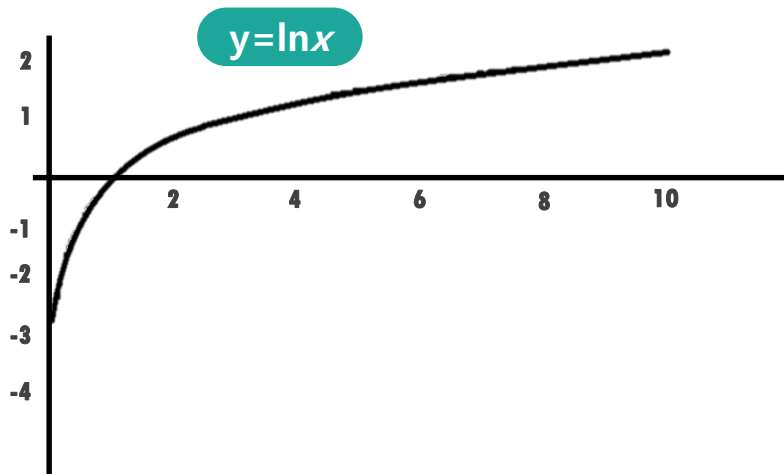
◉ 혈중 콜레스테롤이 정상에서 높은 수치로 갈 때,  
비만인 경우의 오즈는 약 세 배 증가

▪  $\text{odds 비율} = 0.293/0.111 = 2.64$

▪ 혈중 콜레스테롤이 높을 때, **2.64** 배 더 비만이 되기 쉬움

## ■ 로짓 변형

❖ 로짓은 오즈의 자연로그



$$\text{logit}(p) = \ln(\text{odds}) = \ln(p/(1-p))$$



## 2. 로지스틱 회귀 모델



### ■ 로지스틱 회귀 분석이 사용되는 예

- ◉ 종속 변수의 값을 **0 또는 1**로 (이진 변수로) 표현할 수 있는 경우

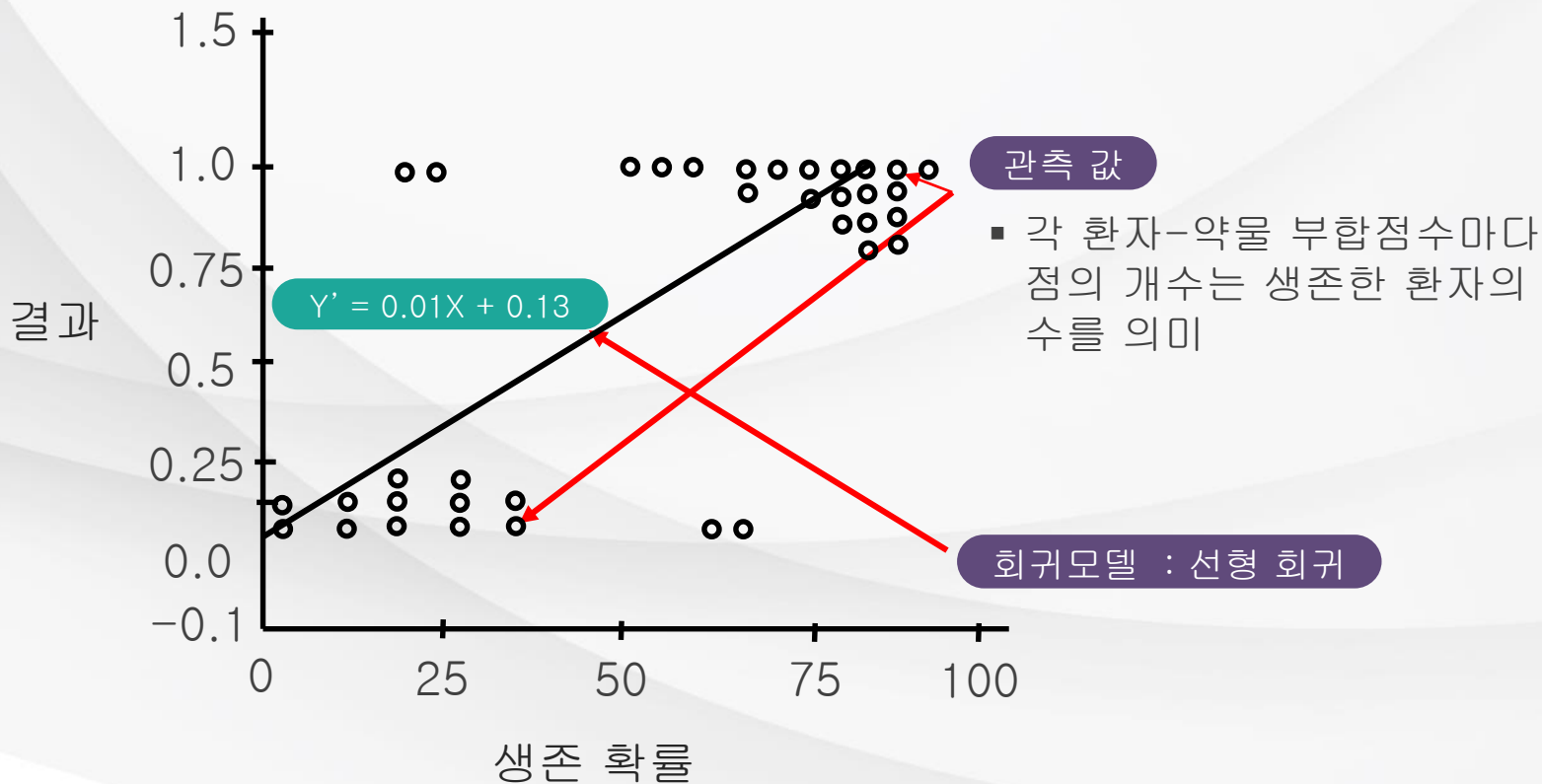
#### 예) 약물 치료 후 환자의 반응 예측

약물 치료에 대한 환자의 반응(종속 변수)을 예측하고자 할 때,

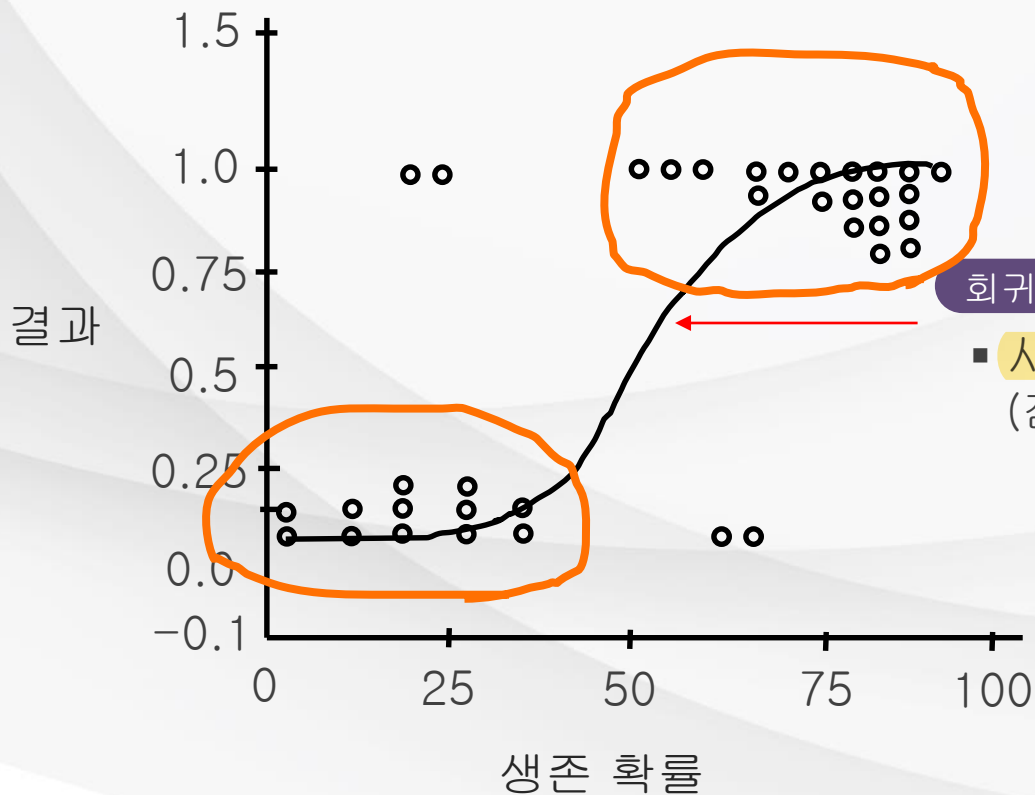
약물 치료 적용 후 환자가 살아남은 경우 **1**로,  
살아남지 못한 경우를 **0**으로 표현할 수 있음

이진변수

### ■ 선형 회귀 모델을 사용할 경우



### ■ 더 나은 솔루션



### ■ 로지스틱 회귀 모델

- 로지스틱 회귀 모델의 방정식

$$\text{logit}(p) = \beta_0 + \beta_1 X$$

- 오즈의 로그(로짓)은 설명변수  $x$ 와 선형적인 관계
- 일반적 선형 회귀 문제처럼 접근 가능

### ■ 종속 변수 $p$ 값 구하기

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

$$\Leftrightarrow \frac{p}{1-p} = e^{\beta_0 + \beta_1 X}$$

$$\rightarrow \frac{1-p}{p} = e^{\beta_0 + \beta_1 X}$$

$$\Leftrightarrow p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

↳ 선형회귀

$p$ 는 시그모이드 함수

### ■ $\beta_1$ 값 해석

#### ◎ Let

- $\text{odds1} = X \mid \text{odds} (p/(1-p))$
- $\text{odds2} = X + 1 \mid \text{odds}$

#### ◎ Then

$$\begin{aligned}\frac{\text{odds2}}{\text{odds1}} &= \frac{e^{b_0 + b_1(X+1)}}{e^{b_0 + b_1X}} \\ &= \frac{e^{(b_0 + b_1X) + b_1}}{e^{b_0 + b_1X}} = \frac{e^{(b_0 + b_1X)} e^{b_1}}{e^{b_0 + b_1X}} = e^{b_1}\end{aligned}$$

- ◎  $X$ 가 각 단위 값마다 증가할 때,  
예측된 **odds**의 비율이 **e**의 기율기( $\beta_1$ ) 제곱만큼 증가함을 의미

### Logit 변환의 의미

❖ 선형회귀에 더 적절한 함수를 도출

$$\text{Logit}(P) = \log \text{ odds} = \ln\left(\frac{P}{1-P}\right)$$

확률

0 -----  $\frac{1}{2}$  ----- 1

Odds

0 ----- 1 -----  $+\infty$

Logit

$-\infty$  ----- 0 -----  $+\infty$



### 3. 로지스틱 회귀 모델의 파라미터 추정



### 3. 로지스틱 회귀 모델의 파라미터 추정

#### ■ 최대 우도 추정법이란?



#### ❖ 동전 던지기 문제

- ◉ 앞/뒤가 나올 확률이 공정하지 않은 (**biased**) 동전이 있을 경우, 동전의 앞면이 나올 확률 **head(p)**를 계산하고자 함
- ◉ 이 때, **p**는 **unknown** 파라미터
- ◉ 동전을 **10**번 던져서 앞면이 **7**번 나왔다고 하자.  
이 때, **p**의 값으로 추정할 수 있는 값 중 가장 최선은 무엇일까?
- ◉ 데이터에 기반하여 **0.7**로 예측

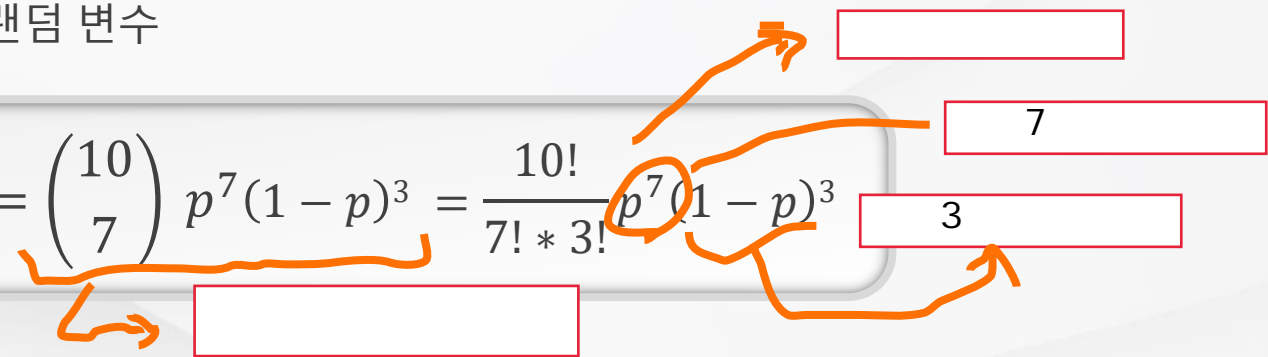
head(p) :

### 3. 로지스틱 회귀 모델의 파라미터 추정

#### ■ 최대 우도 추정법이란?

##### ❖ 동전 던지기 문제

- 동전을 10번 던졌을 때 앞면이 10번 나온 횟수는  $N=10$ 이고  $p=\text{unknown}$ 인 binomial 랜덤 변수

$$\therefore P(7\text{heads}) = \binom{10}{7} p^7 (1-p)^3 = \frac{10!}{7! * 3!} p^7 (1-p)^3$$


- 알지 못하는 파라미터  $p$ 에 대해서 데이터를 관측할 확률을 제공

### 3. 로지스틱 회귀 모델의 파라미터 추정

#### ■ 최대 우도 추정법이란?

- ◉ 이 때, 데이터의 확률을 가장 높이는 **p**의 값을 찾고자 함 (또는 우도함수를 가장 높이는 **p**)
- ◉ 즉, 데이터를 가장 잘 설명할 수 있는 파라미터 **p**를 찾고자 함
- ◉ 어떻게 찾을 수 있을까?
  - 함수에 **log**를 씌움 : 곱셈을 덧셈으로 바꿈으로써 미분을 쉽게 함
  - **p**에 대해서 미분 계산 : **p**의 변화량에 대한 함수의 변화량 계산
  - 미분 값을 **0**으로 설정하고 **p**를 계산 : 함수의 최대 값이 되는 곳은 미분 값이 **0**
- ◉ 파라미터 대입해서 확률 계산해보기

$$\text{Likelihood} = \binom{10}{7} (.7)^7 (.3)^3 = 120 (.7)^7 (.3)^3 = .267$$

$$\log \text{Likelihood} = \log \frac{10}{7! * 3!} + 7\log p + 3\log(1 - p)$$

$$\frac{d}{dp} \log \text{Likelihood} = 0 + \frac{7}{p} - \frac{3}{1 - p}$$

$$\frac{7}{p} - \frac{3}{1 - p} = 0$$

$$\frac{7(1 - p) - 3p}{p(1 - p)} = 0$$

$$7(1 - p) = 3p$$

$$7 - 7p = 3p$$

$$7 = 10p$$

$$p = \frac{7}{10}$$

### 3. 로지스틱 회귀 모델의 파라미터 추정

#### ■ 모수( $\beta_i$ ) 추정 방법

- ◎ 최대 우도(**maximum likelihood**) 추정법 이용

- 우도함수(**likelihood function**)

$$L(x_i; \beta_i) = \prod_{i=1}^n \left( \frac{1}{1 + \exp(-\eta)} \right)^{y_i} \left( 1 - \frac{1}{1 + \exp(-\eta)} \right)^{1-y_i}$$

$$\eta = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)$$

- 우도함수를 최대화하는 최대 우도 추정법(**MLE**)을 이용하여  $\hat{\beta}_i$ 을 수치적(**numerical**) 방법으로 산출

$$\text{Max}[L(x_i; \beta_i)] \Rightarrow \beta_i ??$$

y

.

### 3. 로지스틱 회귀 모델의 파라미터 추정

#### 로지스틱 회귀 모델 예제

❖ 나이에 따른 동맥 심장 질환 확률 예측하기

데이터

	55세 이상	55세 미만
동맥 심장 질환 유	21	22
동맥 심장 질환 무	6	51

로지스틱 모델

$$\log\left(\frac{P(D)}{1-P(D)}\right) = \alpha + \beta_1 X_1$$

$$X_1 = \begin{cases} 1 & \text{if age} \geq 55 \\ 0 & \text{if age} < 55 \end{cases}$$

우도함수 식

$$55 \quad 55 \quad = y$$

$$\mathcal{L}(\alpha, \beta_1) = \left(\frac{e^{-\alpha-\beta_1}}{1+e^{-\alpha-\beta_1}}\right)^6 \left(\frac{1}{1+e^{-\alpha-\beta_1}}\right)^{21} \left(\frac{e^{-\alpha}}{1+e^{-\alpha}}\right)^{51} \left(\frac{1}{1+e^{-\alpha}}\right)^{22}$$

$$1-y$$

### 3. 로지스틱 회귀 모델의 파라미터 추정

#### 로지스틱 회귀 모델 예제

$$L(\alpha, \beta_1) = \left( \frac{e^{-\alpha - \beta_1}}{1 + e^{-\alpha - \beta_1}} \right)^6 \left( \frac{1}{1 + e^{-\alpha - \beta_1}} \right)^{21} \left( \frac{e^{-\alpha}}{1 + e^{-\alpha}} \right)^{51} x \left( \frac{1}{1 + e^{-\alpha}} \right)^{22}$$

로그 우도함수

$$\therefore \log L(\alpha, \beta_1) =$$

$$6(-\alpha - \beta_1) - 6\log(1 + e^{-\alpha - \beta_1}) + 0 - 21\log(1 + e^{-\alpha - \beta_1}) - 51\alpha + 51\log(1 + e^{-\alpha}) + 0 - 22\log(1 + e^{-\alpha})$$

### 3. 로지스틱 회귀 모델의 파라미터 추정

#### 로지스틱 회귀 모델 예제

##### 로그 우도함수

$$\begin{aligned} \therefore \log L(\alpha, \beta_1) = \\ 6(-\alpha - \beta_1) - 6\log(1 + e^{-\alpha - \beta_1}) + 0 - 21\log(1 + e^{-\alpha - \beta_1}) - \\ 51\alpha + 51\log(1 + e^{-\alpha}) + 0 - 22\log(1 + e^{-\alpha}) \end{aligned}$$

##### 미분식

$$\begin{aligned} \frac{d[\log L(\beta_1)]}{d\beta_1} = \\ -6 + \frac{6e^{-\alpha - \beta_1}}{1 + e^{-\alpha - \beta_1}} + \frac{21e^{-\alpha - \beta_1}}{1 + e^{-\alpha - \beta_1}} = 0 \end{aligned}$$

$$\begin{aligned} \frac{d[\log L(\alpha)]}{d\alpha} = \\ 51 - \frac{51e^{-\alpha}}{1 + e^{-\alpha}} - \frac{22e^{\alpha}}{1 + e^{\alpha}} = 0 \end{aligned}$$



학습정리



지금까지 [로지스틱 회귀 모델]에 대해서 살펴보았습니다

1

로짓(logit) 함수

$$\text{odds} = p / (1-p)$$

로짓은 오즈의 자연로그

$$\text{logit}(p) = \ln(\text{odds}) = \ln(p/(1-p))$$

로지스틱 회귀 모델

$$\text{logit}(p) = b_0 + b_1 X$$

로짓과 설명변수 X를 선형적인 관계로 모델링  
종속변수 p와 X는 비선형 관계

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

로지스틱 회귀 모델의 파라미터 추

정

우도함수 사용

$$L(x_i; \beta_i) = \prod_{i=1}^n \left( \frac{1}{1 + \exp(-\eta)} \right)^{y_i} \left( 1 - \frac{1}{1 + \exp(-\eta)} \right)^{1-y_i}$$