



# Logistic Regression

강필성

고려대학교 산업경영공학부

pilsung\_kang@korea.ac.kr

# 로지스틱 회귀분석: Logistic Regression

- 다중 선형 회귀분석

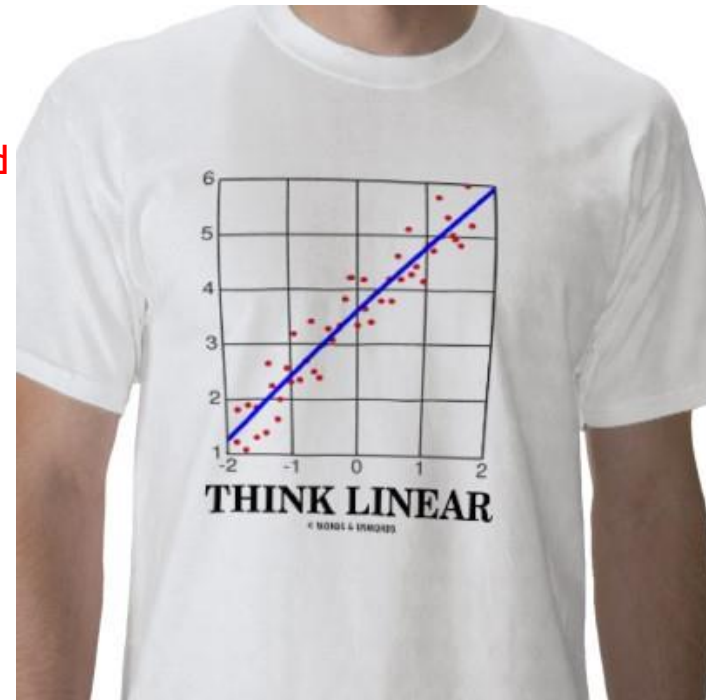
- ✓ 수치형 설명변수 X와 종속변수 Y간의 관계를 선형으로 가정하고 이를 가장 잘 표현할 수 있는 회귀 계수를 추정

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_d x_d + \epsilon$$

unexplained

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$

coefficients



# 로지스틱 회귀분석: Logistic Regression

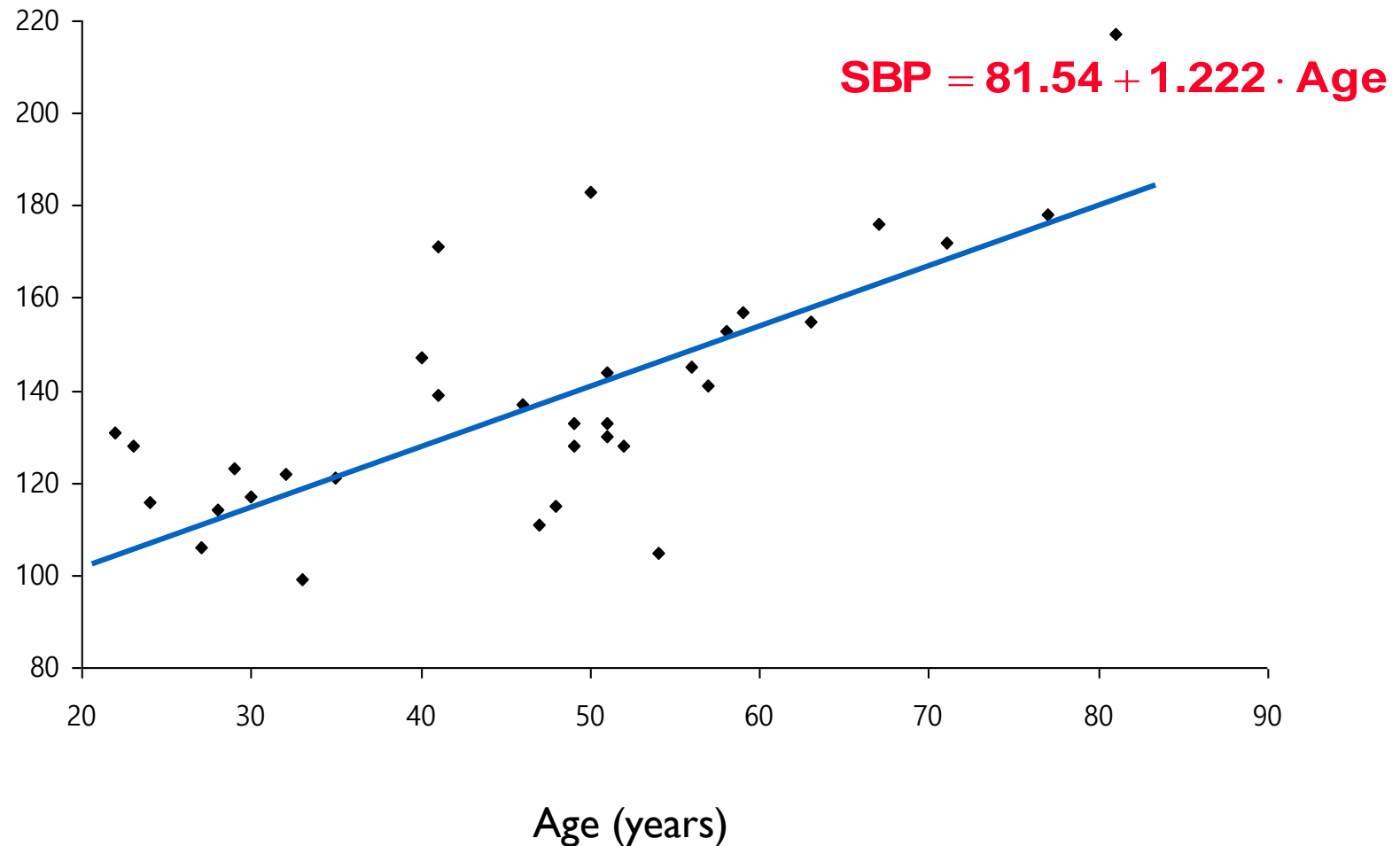
- 예시 I

✓ 33명의 성인 여성에 대한 나이와 혈압 사이의 관계

Age	SBP	Age	SBP	Age	SBP
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217

# 로지스틱 회귀분석: Logistic Regression

SBP (mm Hg)



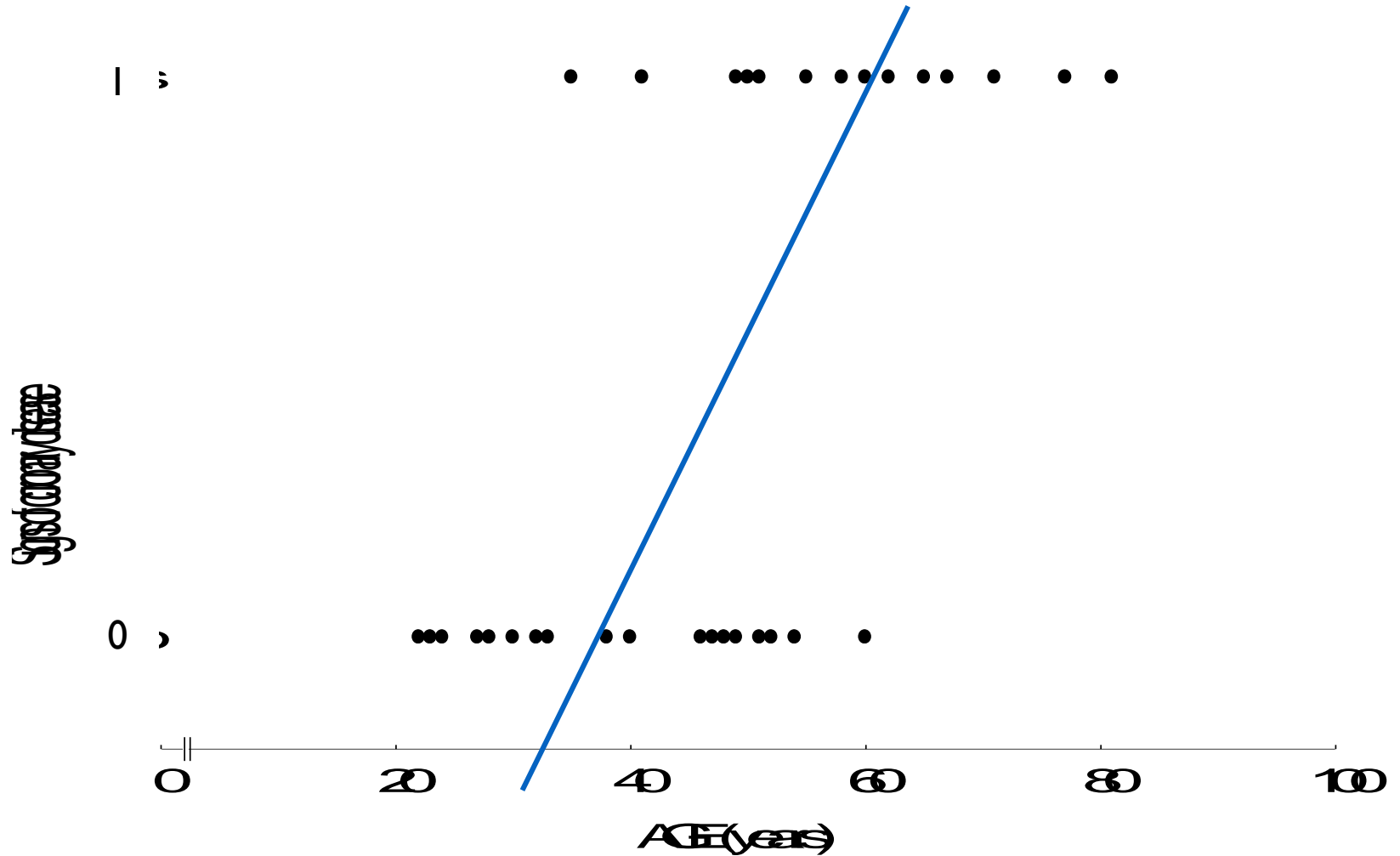
# 로지스틱 회귀분석: Logistic Regression

- 예시 2

✓ 연속형 변수가 아닌 이진형<sup>Binary</sup> 변수인 Cancer Diagnosis를 사용한다면?

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

# 로지스틱 회귀분석: Logistic Regression



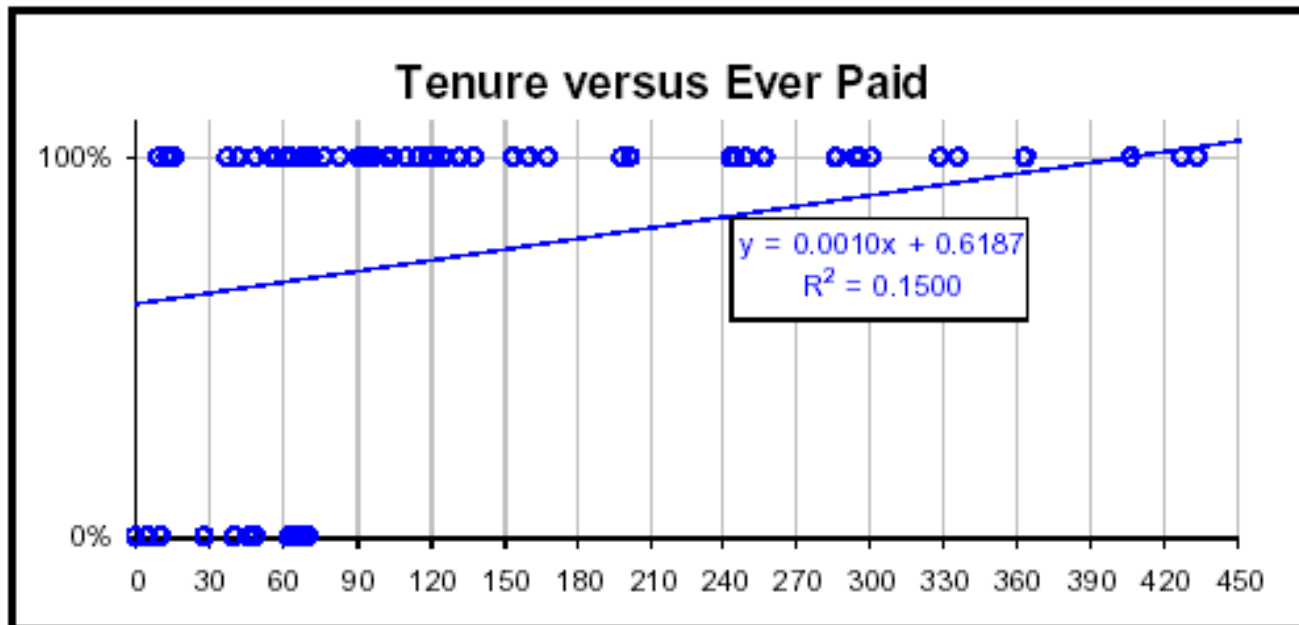
# 로지스틱 회귀분석: Logistic Regression

- 0/1의 이진 값이 아닌 확률값을 종속 변수로 사용한다면?
  - ✓ 선형회귀분석의 우변의 범위에 대한 제한이 없기 때문에 종속변수(좌변) 역시 범위의 제한을 받지 않으므로 적절하지 않음

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$



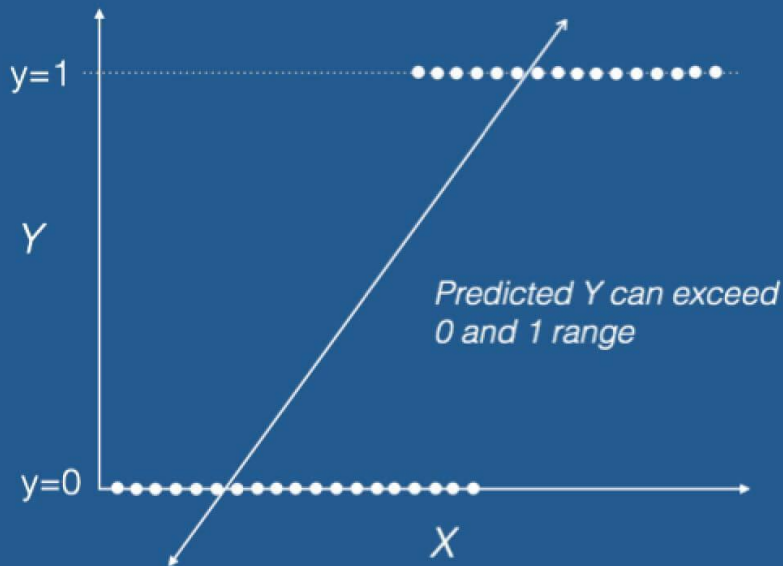
$$P(y = 1) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$



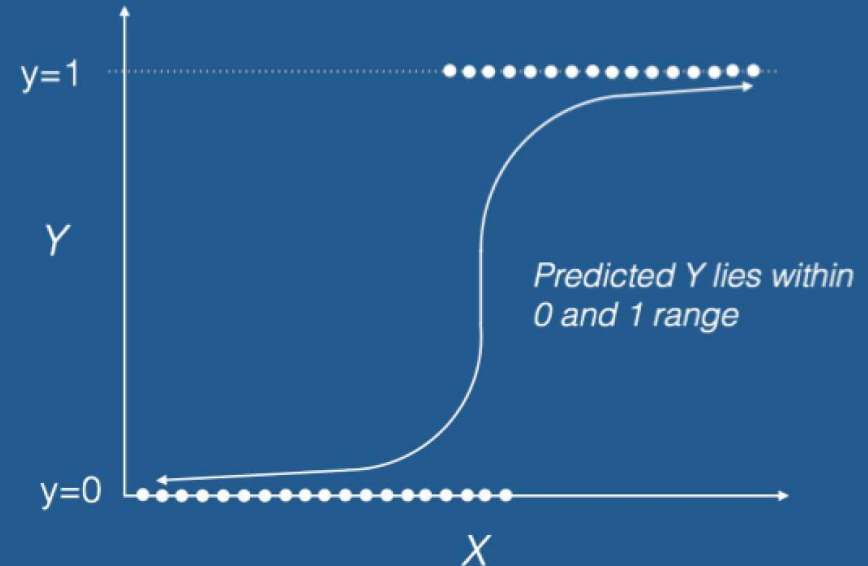
# 로지스틱 회귀분석: Logistic Regression

- 0/1의 이진 값이 아닌 확률값을 종속 변수로 사용한다면?
  - ✓ 선형회귀분석의 우변의 범위에 대한 제한이 없기 때문에 종속변수(좌변) 역시 범위의 제한을 받지 않으므로 적절하지 않음

## Linear Regression



## Logistic Regression





# 로지스틱 회귀분석: Logistic Regression

- 목적

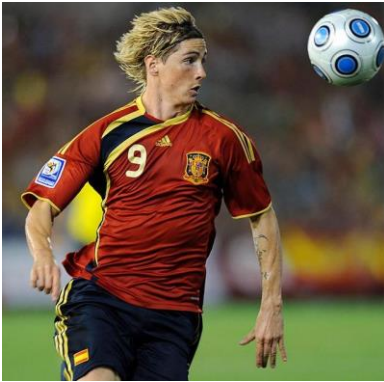
- ✓ 이진형(0/1)의 형태를 갖는 종속변수(분류문제)에 대해 회귀식의 형태로 모형을 추정하는 것

- 속성

- ✓ 종속변수  $Y$  자체를 그대로 사용하는 것이 아니라  $Y$ 에 대한 로짓 함수  $\text{logit function}$ 를 회귀식의 종속변수로 사용
- ✓ 로짓함수는 설명변수의 선형결합으로 표현될 수 있음
- ✓ 로짓함수의 값은 종속변수에 대한 성공 확률로 역산될 수 있으며, 이는 따라서 분류 문제에 적용할 수 있음

# 로지스틱 회귀분석: Logistic Regression

- 2010 World Cup Betting Odds



9 : 2



9 : 2



6 : 1



9 : 1



200 : 1



250 : 1



500 : 1



1000 : 1

# 로지스틱 회귀분석: 승산 (Odds)

- 승산 (Odds)

- ✓  $p$ : 성공 범주(class = 1)에 속할 확률

$$Odds = \frac{p}{1 - p}$$

- 이전 예시에 대해

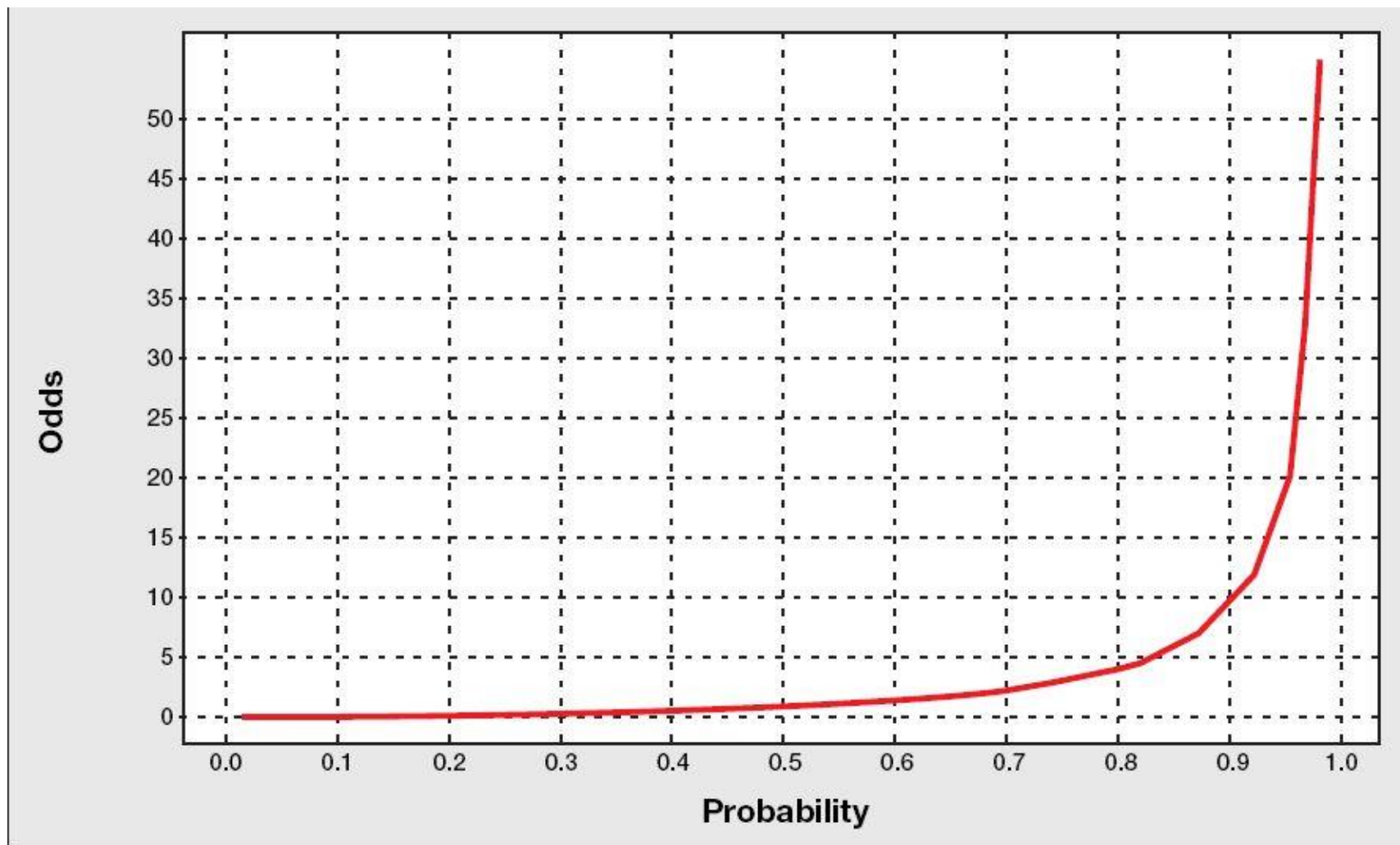
- ✓ 스페인의 우승 odds는 2/9이므로 스페인의 우승 확률은 2/11임

- ✓ 대한민국의 우승 odds는 1/250 이므로 대한민국의 우승확률은  $1/251 \approx 0.00398$  (0.398%)임

- ✓ 1,000년을 살면 대한민국이 월드컵에서 한 번 우승하는 모습을 목격할 수 있음

# 로지스틱 회귀분석: 승산 (Odds)

- 확률값이 0부터 1로 변화함에 따라 승산<sup>Odds</sup>은 0부터 무한대의 값을 가짐



# 로지스틱 회귀분석: 승산 (Odds)

- Odds의 한계

- ✓ 여전히 범위에 대한 제약이 존재함:  $0 < odds < \infty$

- ✓ 비대칭성<sup>Asymmetric</sup>

- Odds에 로그를 취하자

$$\log(Odds) = \log\left(\frac{p}{1-p}\right)$$

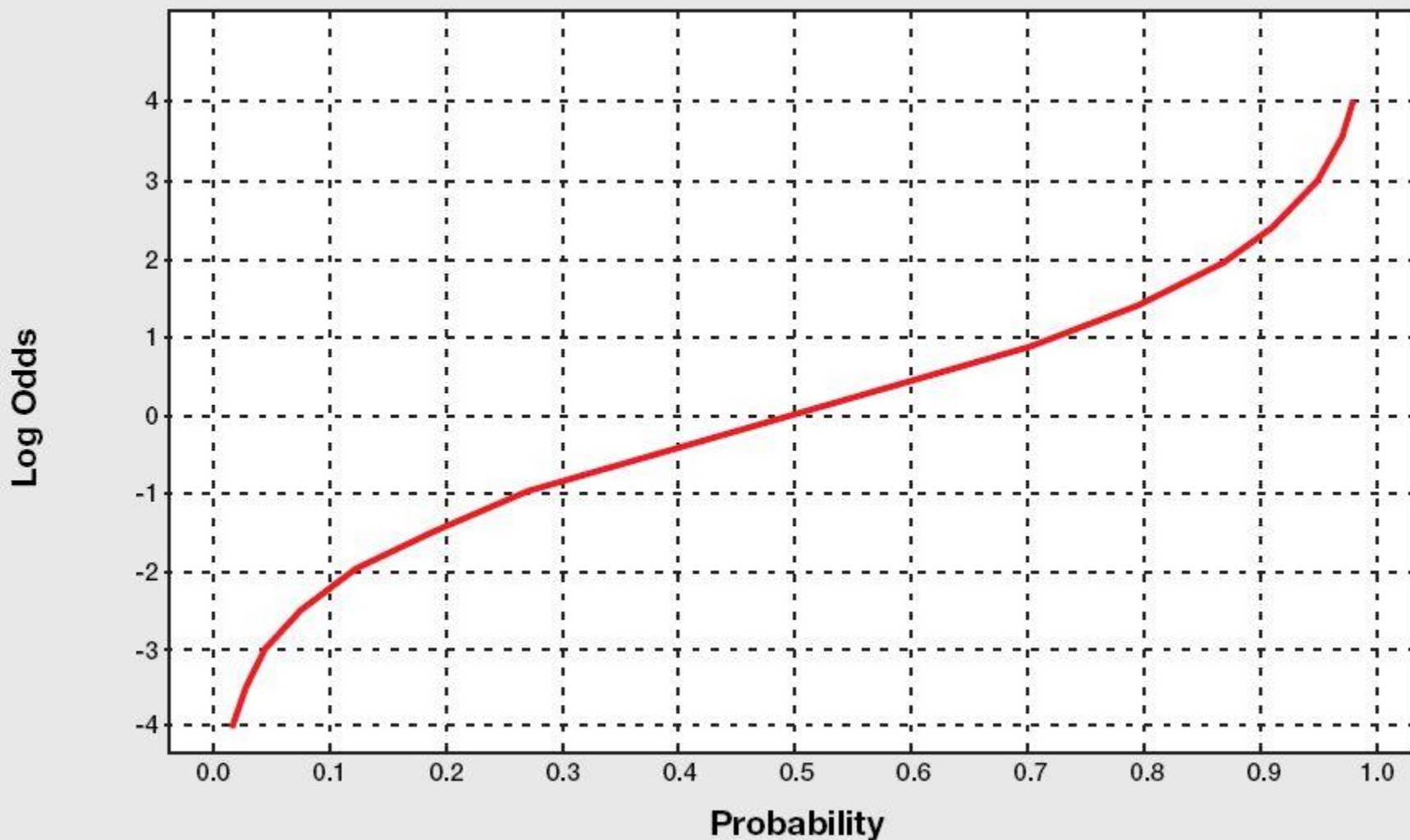
- ✓ 드디어 범위에 대한 제약이 없어짐:  $-\infty < \log(odds) < \infty$

- ✓ 대칭성 확보

- ✓ 성공확률  $p$ 가 작으면 음수값을 갖고, 성공확률  $p$ 가 크면 양수값을 가짐

# 로지스틱 회귀분석: 승산 (Odds)

- 확률값이 0부터 1까지 변화함에 따라 로그 승산은  $-\infty \sim \infty$ 의 값을 가지며 대칭임



# 로지스틱 회귀분석: Equation

- 로지스틱 회귀분석 식

- ✓ Log Odds를 이용한 회귀분석 식

$$\log(Odds) = \log\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$

- ✓ 양변에 로그를 취하면

$$\frac{p}{1-p} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d}$$

- ✓ 성공확률에 대한 식으로 표현

$$p = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d)}} = \sigma(\mathbf{x}|\beta)$$

# 로지스틱 회귀분석: Equation

- 로지스틱 회귀분석 식

Logistic  
Regression  
선형식

$$\ln\left(\frac{p}{1-p}\right) \quad : \text{logit} \quad (\text{odds에 자연로그를 취한 상태})$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i$$

• 로지스틱 회귀 모형은 종속변수가 이분형일 때 선형회귀모형의 제약을 극복하기 위해 확률에 대한 로짓 변환을 고려하여 분석

$$P = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i)}$$

• 위의 모형식에서 추정된 회귀계수로부터 사후확률에 대한 추정식을 계산



# 로지스틱 회귀분석: 학습

- 로지스틱 회귀분석에서 회귀 계수의 추정

✓ 동일한 데이터셋에 대해 다음과 같이 두 가지의 로지스틱 회귀분석 모형이 존재한다고 하면 어떤 모형이 현재 데이터를 더 잘 설명하는 모형인가?

Model A

고객	대출이용	P(Y=1)	P(Y=0)
1	1	0.908	0.092
2	0	0.201	0.799
3	1	0.708	0.292
4	0	0.214	0.786
5	1	0.955	0.045
6	0	0.017	0.983
7	1	0.807	0.193
8	0	0.126	0.874
9	1	0.937	0.063
10	0	0.068	0.932

Model B

고객	대출이용	P(Y=1)	P(Y=0)
1	1	0.557	0.443
2	0	0.425	0.575
3	1	0.604	0.396
4	0	0.387	0.613
5	1	0.615	0.385
6	0	0.356	0.644
7	1	0.406	0.594
8	0	0.508	0.492
9	1	0.704	0.296
10	0	0.325	0.675

✓ 실제 정답이 1일 때 1범주로 예측할 확률이 높고 실제 정답이 0일 때 0범주로 예측할 확률이 높으므로 Model A가 더 우수한 모형임

# 로지스틱 회귀분석: 학습

- 로지스틱 회귀분석에서 회귀 계수의 추정

✓ 우도 함수 Likelihood function

- 개별 객체의 우도 함수는 해당 학습 데이터가 정답 범주에 속할 확률 (Glass 1의 우도 함수 값은 0.908, Glass 2의 우도 함수 값은 0.799)
- 데이터의 생성 과정이 독립임을 가정할 수 있을 때, 전체 데이터셋의 우도 함수는 개별 객체의 우도 함수를 모두 곱한 값임
- 일반적으로 데이터셋의 우도 함수는 매우 작은 값을 가지므로(1보다 작은 소수가 계속 곱해지므로) 로그 우도 함수를 주로 사용

Model A

고객	대출이용	P(Y=1)	P(Y=0)
1	1	<b>0.908</b>	0.092
2	0	0.201	<b>0.799</b>
3	1	<b>0.708</b>	0.292
4	0	0.214	<b>0.786</b>
5	1	<b>0.955</b>	0.045
6	0	0.017	<b>0.983</b>
7	1	<b>0.807</b>	0.193
8	0	0.126	<b>0.874</b>
9	1	<b>0.937</b>	0.063
10	0	0.068	<b>0.932</b>

# 로지스틱 회귀분석: 학습

- 로지스틱 회귀분석에서 회귀 계수의 추정

✓ 우도 함수 Likelihood function

Model A

고객	대출이용	P(Y=1)	P(Y=0)	우도	로그 우도
1	1	0.908	0.092	0.908	-0.0965
2	0	0.201	0.799	0.799	-0.2244
3	1	0.708	0.292	0.708	-0.3453
4	0	0.214	0.786	0.786	-0.2408
5	1	0.955	0.045	0.955	-0.0460
6	0	0.017	0.983	0.983	-0.0171
7	1	0.807	0.193	0.807	-0.2144
8	0	0.126	0.874	0.874	-0.1347
9	1	0.937	0.063	0.937	-0.0651
10	0	0.068	0.932	0.932	-0.0704
				<b>0.233446</b>	<b>-0.1455</b>

Model B

고객	대출이용	P(Y=1)	P(Y=0)	우도	로그 우도
1	1	0.557	0.443	0.557	-0.5852
2	0	0.425	0.575	0.575	-0.5534
3	1	0.604	0.396	0.604	-0.5042
4	0	0.387	0.613	0.613	-0.4894
5	1	0.615	0.385	0.615	-0.4861
6	0	0.356	0.644	0.644	-0.4401
7	1	0.406	0.594	0.406	-0.9014
8	0	0.508	0.492	0.492	-0.7093
9	1	0.704	0.296	0.704	-0.3510
10	0	0.325	0.675	0.675	-0.3930
				<b>0.004458</b>	<b>-0.5413</b>

✓ Model A의 (로그) 우도 함수가 Model B의 (로그) 우도 함수보다 큼

✓ Model A가 Model B보다 데이터셋을 더 잘 설명하는 모델

# 로지스틱 회귀분석: 학습 (Optional)

- 최대 우도 추정법 Maximum likelihood estimation (MLE)

- ✓ 학습 데이터의 개별 객체들이 갖는 label에 대한 확률을 극대화 하자

- ✓ i번째 객체에 대한 우도 함수

$$P(\mathbf{x}_i, y_i | \boldsymbol{\beta}) = \begin{cases} \sigma(\mathbf{x}_i | \boldsymbol{\beta}), & \text{if } y_i = 1 \\ 1 - \sigma(\mathbf{x}_i | \boldsymbol{\beta}), & \text{if } y_i = 0 \end{cases}$$

- ✓ 출력변수가 1과 0임을 고려하여 다음과 같이 변형 가능

$$P(\mathbf{x}_i, y_i | \boldsymbol{\beta}) = \sigma(\mathbf{x}_i | \boldsymbol{\beta})^{y_i} (1 - \sigma(\mathbf{x}_i | \boldsymbol{\beta}))^{1-y_i}$$

# 로지스틱 회귀분석: 학습 (Optional)

- 최대 우도 추정법 Maximum likelihood estimation (MLE)

- ✓ 학습 데이터셋의 객체들이 독립적으로 발생됨을 가정할 경우 전체 데이터 셋에 대한 우도 함수는 다음과 같이 표현됨

$$L(\mathbf{X}, \mathbf{y} | \boldsymbol{\beta}) = \prod_{i=1}^N P(\mathbf{x}_i, y_i | \boldsymbol{\beta}) = \prod_{i=1}^N \sigma(\mathbf{x}_i | \boldsymbol{\beta})^{y_i} (1 - \sigma(\mathbf{x}_i | \boldsymbol{\beta}))^{1-y_i}$$

- ✓ 양변에 로그를 취하면

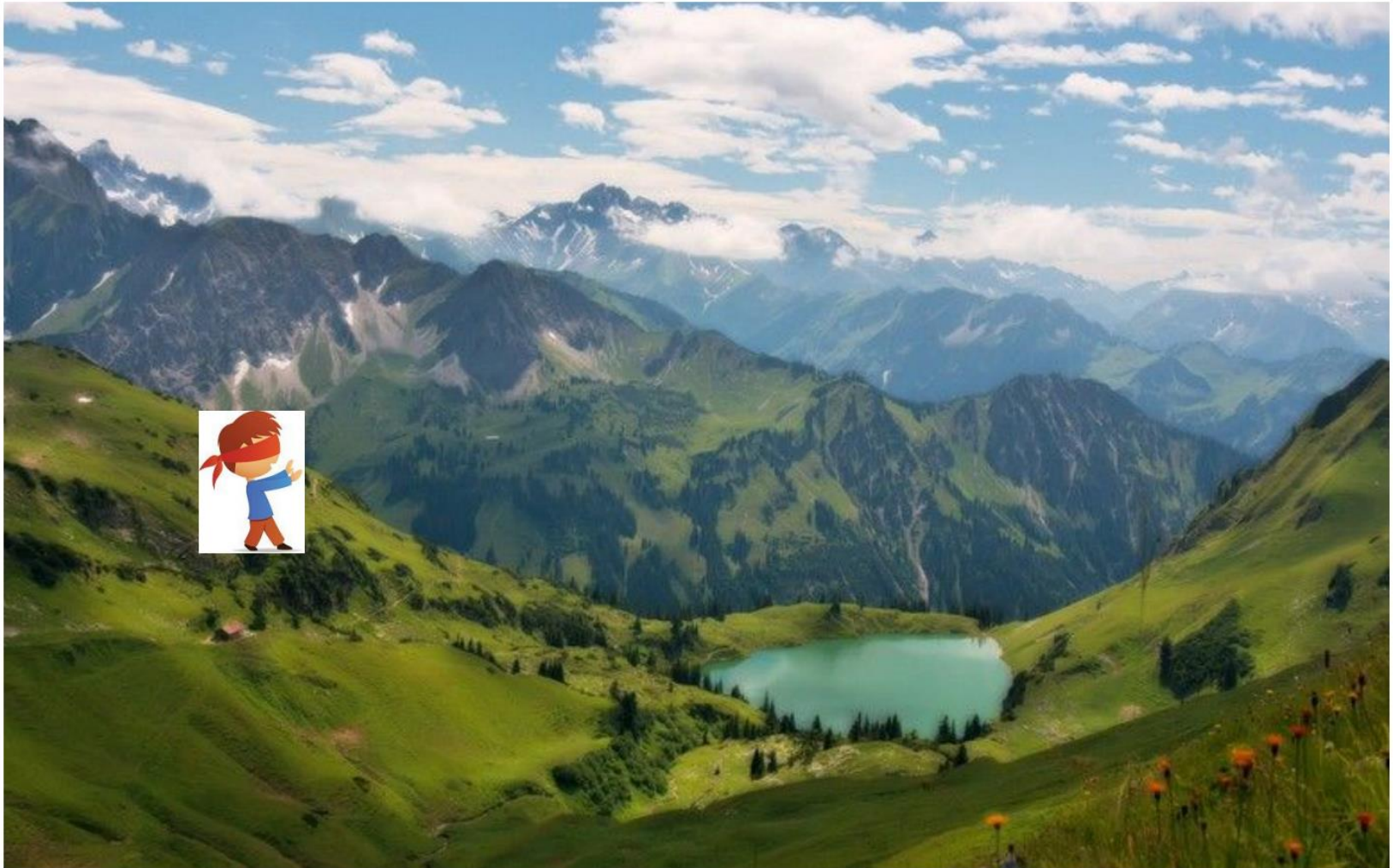
$$\log L(\mathbf{X}, \mathbf{y} | \boldsymbol{\beta}) = \sum_{i=1}^N \left( y_i \log(\sigma(\mathbf{x}_i | \boldsymbol{\beta})) + (1 - y_i) \log(1 - \sigma(\mathbf{x}_i | \boldsymbol{\beta})) \right)$$

- ✓ 우도함수와 로그-우도함수는 회귀계수  $\boldsymbol{\beta}$ 에 대해 비선형이므로 선형회귀분석과 같이 명시적인 해가 존재하지 않음

- Conjugate gradient 등의 최적화 알고리즘을 차용하여 해를 구함

# 기울기 하강: Gradient Descent

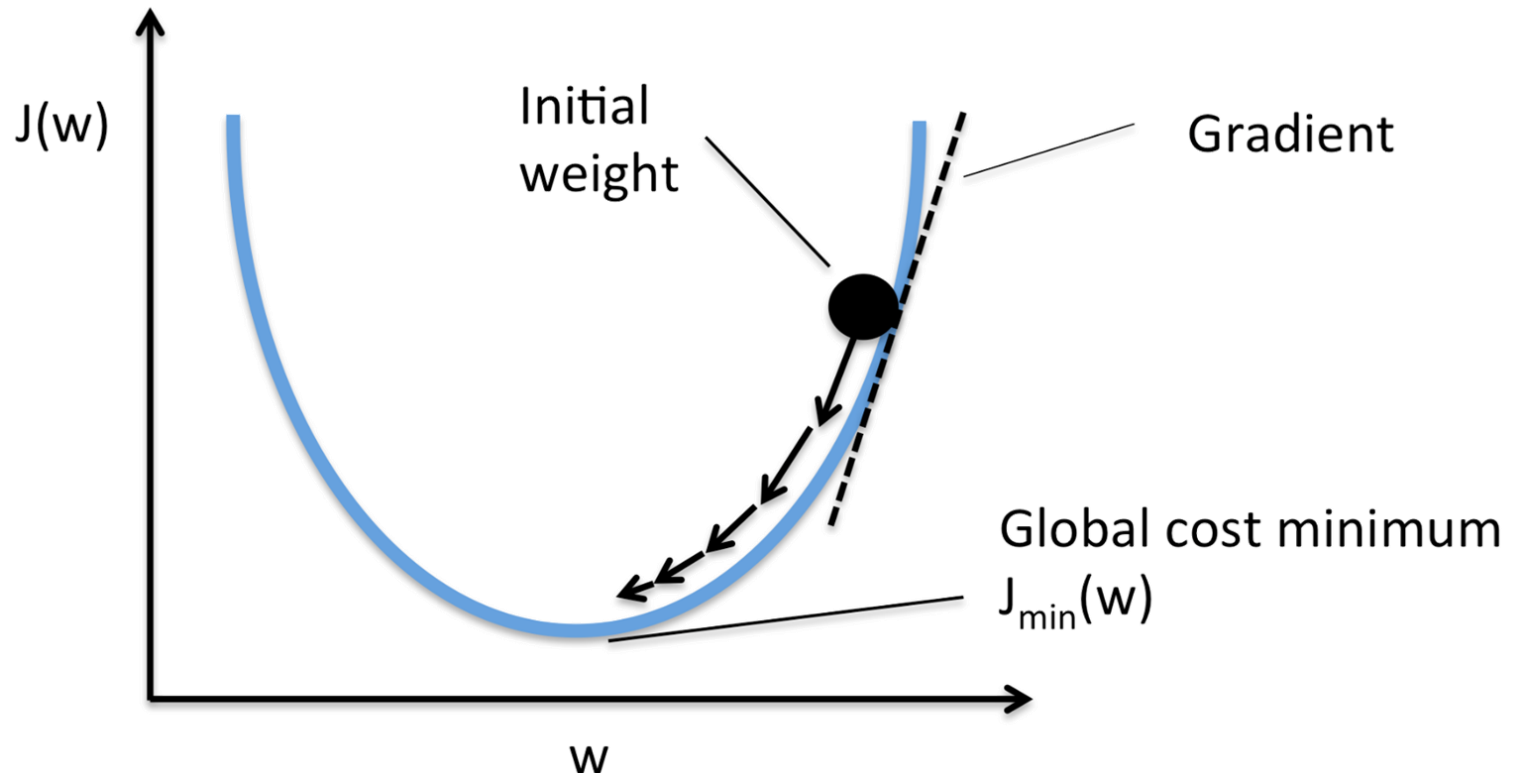
- 눈을 가린 채로 산에서 가장 낮은 곳을 찾아가기



# 기울기 하강: Gradient Descent

- 기울기 하강 Gradient descent algorithm

- ✓ 파란색 선: 미지수  $w$ 의 변화에 따른 목적함수 값의 변화
- ✓ 검은색 점: 현재 해의 위치
- ✓ 화살표: 목적함수를 최적화하기 위해 미지수  $w$ 가 이동해야 하는 방향



# 기울기 하강: Gradient Descent

- 비용함수를 현재의 가중치 값( $w$ )에 대해 1차 미분을 수행한 뒤 아래의 절차를 따름
  - ✓ 1차 미분 값( $\text{gradient}$ )이 0인가?
    - **그렇다**: 현재의 가중치 값이 최적! → 학습 종료
    - **아니다**: 현재의 가중치 값이 최적이지 않음 → 좀 더 학습해봐
  - ✓ 1차 미분 값( $\text{gradient}$ )가 0이 아닐 경우 어떻게 해야 좀 더 잘하는 퍼셉트론을 만들 수 있는가?
    - 1차 미분 값( $\text{gradient}$ )의 부호에 대한 반대 방향으로 가중치를 이동
  - ✓ 반대 방향으로 얼마나 움직여야 하는가?
    - 그건 잘 모름...
    - 조금씩 적당히(???) 움직여보고 그 다음에 다시  $\text{gradient}$ 를 구해보자
    - 하다 보면 되겠지...



# 기울기 하강: Gradient Descent (optional)

- 기울기 하강: Gradient descent algorithm

- ✓ 함수의 테일러 전개

$$f(w + \Delta w) = f(w) + \frac{f'(w)}{1!} \Delta w + \frac{f''(w)}{2!} (\Delta w)^2 + \dots$$

- ✓ 목적함수가 최소화인 경우 함수의 1차 미분 값(gradient)이 0이 아니면 Gradient의 반대 방향으로 이동해야 목적함수의 값을 감소시킬 수 있음

$$w_{new} = w_{old} - \alpha f'(w), \quad \text{where } 0 < \alpha < 1.$$

어느 방향으로 갈 것인가?

얼마만큼 갈 것인가?

- ✓ 이동 후의 새로운 함수 값은 이동 전의 함수 값보다 작음

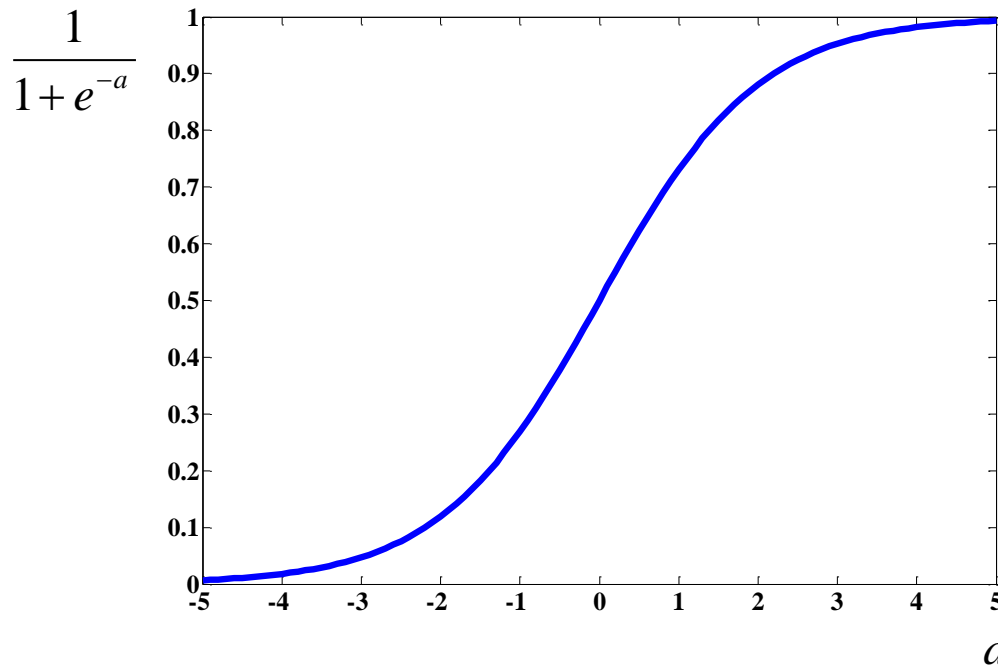
$$f(w_{new}) = f(w_{old} - \alpha f'(w_{old})) \cong f(w_{old}) - \alpha |f'(w)|^2 < f(w_{old})$$

# 로지스틱 회귀분석: 성공 확률

- 성공 확률

- ✓ 회귀계수가 추정되고 나면 주어진 설명변수집합에 대한 성공확률을 다음과 같이 계산할 수 있음

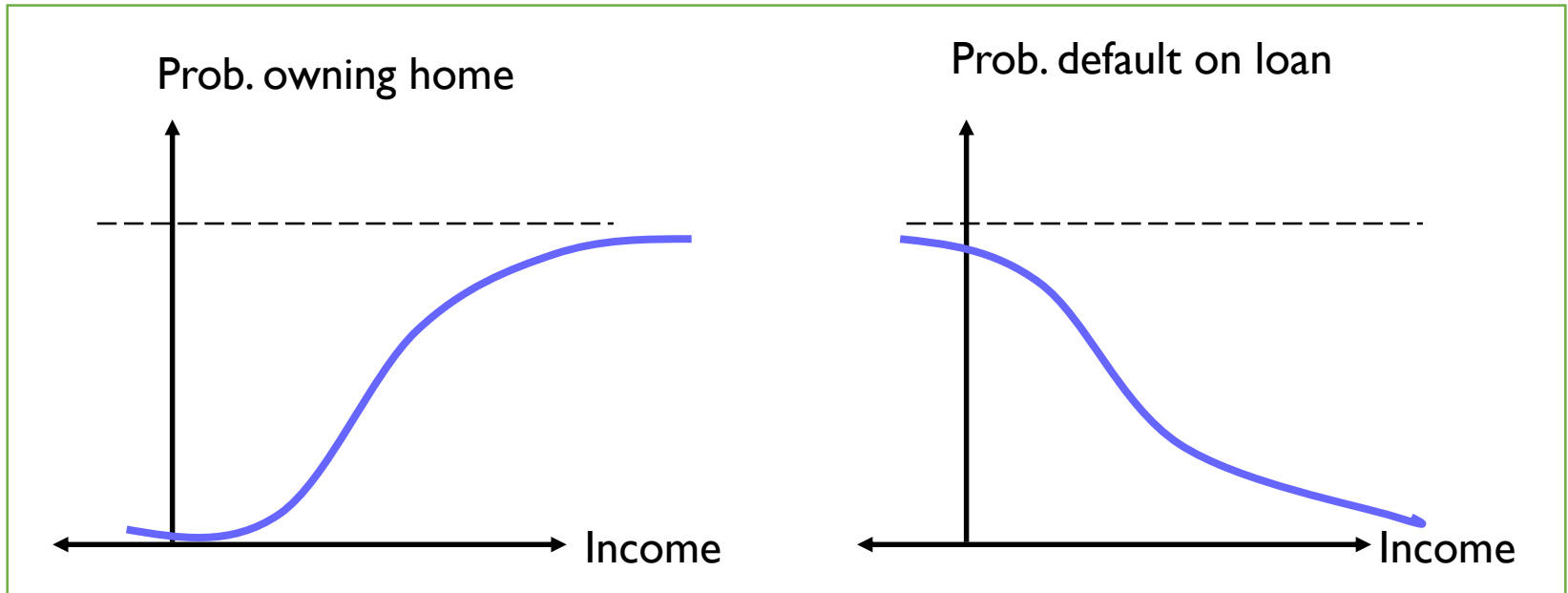
$$P(y = 1) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d)}}$$



# 로지스틱 함수의 의미

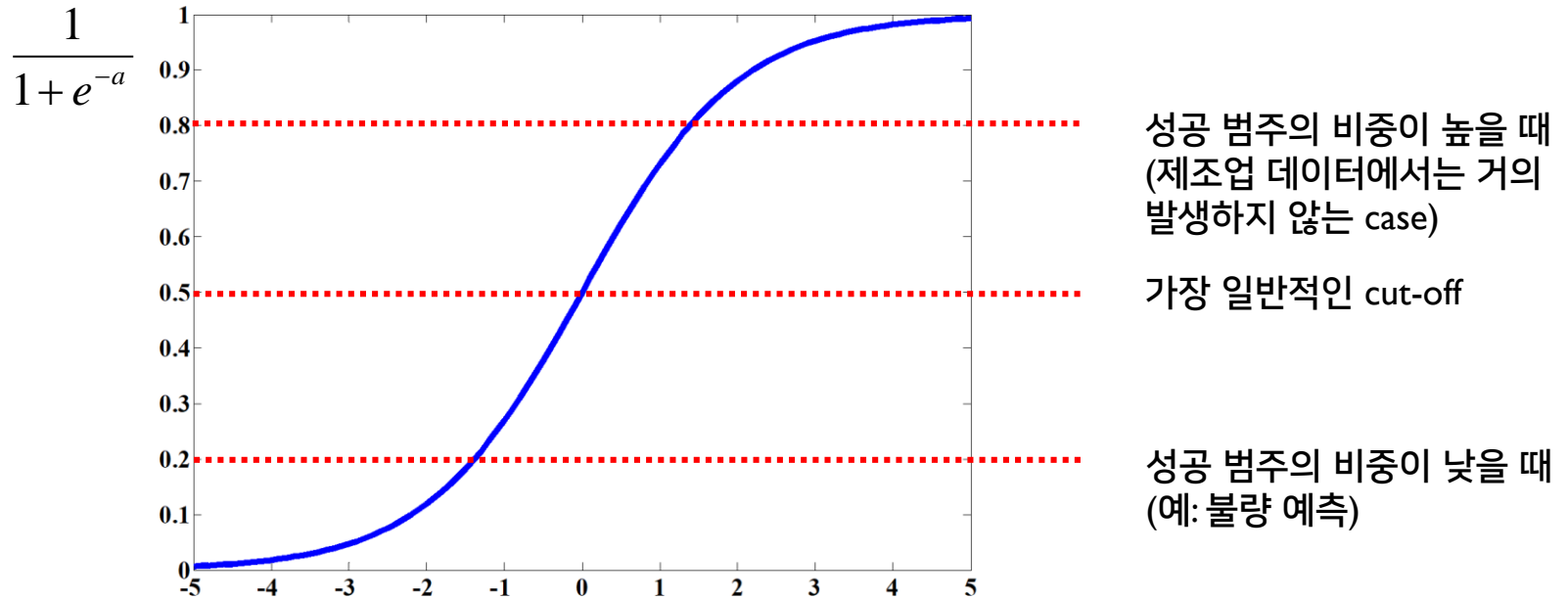
- 실제 상황에서는

- ✓ 특정 변수에 대한 확률 값은 선형이 아닌 S-커브 형태를 따르는 경우가 많음



# 로지스틱 함수의 의미

- 이진분류를 위한 cut-off 설정



- ✓ 일반적으로 0.5가 주로 사용됨
- ✓ 사전확률을 고려한 cut-off나 검증데이터의 정확도를 최대화하는 cut-off 등이 사용될 수도 있음

# 로지스틱 회귀분석: 해석

- 로지스틱 회귀분석 회귀계수의 의미

- ✓ 선형 회귀분석 회귀식

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$

- ✓ 선형 회귀분석에서의 회귀계수는 해당 변수가 1 증가함에 따른 종속변수의 변화량

- ✓ 로지스틱 회귀분석 회귀식

$$\log(Odds) = \log\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$
$$p = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d)}}$$

- ✓ 로지스틱 회귀분석에서의 회귀계수는 해당 변수가 1 증가함에 따른 **로그 승산의 변화량**

# 로지스틱 회귀분석: 해석

- 승산 비율: Odds Ratio

- ✓ 로지스틱 회귀분석에서 나머지 변수는 모두 고정시킨 상태에서 한 변수를 1만큼 증가시켰을 때 변화하는 Odds의 비율

- ✓ Odds ratio:

$$\frac{odds(x_1 + 1, \dots, x_d)}{odds(x_1, \dots, x_d)} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1(x_1 + 1) + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_d x_d}}{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_d x_d}} = e^{\hat{\beta}_1}$$

- ✓  $x_1$ 이 1 증가하게 되면 성공에 대한 승산 비율이  $e^{\hat{\beta}_1}$  만큼 변화함

- 회귀 계수가 양수 → 변수가 증가하면 성공 확률이 **증가** (성공범주와 양의 상관관계)
- 회귀 계수가 음수 → 변수가 증가하면 성공 확률이 **감소** (성공범주와 음의 상관관계)

$$\frac{p}{1 - p} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_d x_d}$$

# 로지스틱 회귀분석: 해석

- 로지스틱 회귀분석 결과 및 해석

- ✓ 로지스틱 회귀분석을 수행하고 나면 선형 회귀분석과 유사하게 다음과 같은 표를 결과로 얻을 수 있음

$$p = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d)}}$$

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

# 로지스틱 회귀분석: 해석

## • 로지스틱 회귀분석 결과 및 해석

### ✓ 회귀계수 Coefficient

- 로지스틱 회귀분석에서 각 변수에 대응하는 베타값임
- 선형회귀분석에서는 해당 변수가 1단위 증가할 때 종속변수의 변화량을 의미하나, 로지스틱 회귀분석에서는 해당 변수가 1단위 증가할 때 로그승산비의 변화량을 의미
- 양수이면 성공확률과 양의 상관관계, 음수이면 성공 확률과 음의 상관관계

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712



# 로지스틱 회귀분석: 해석

- 로지스틱 회귀분석 결과 및 해석

- ✓ 유의확률 p-value

- 로지스틱 회귀분석에서 해당 변수가 통계적으로 유의미한지 여부를 알려주는 지표
    - 0에 가까울수록 모델링에 중요한 변수이며, 1에 가까울수록 유의미하지 않은 변수임
    - 특정 유의수준( $\alpha$ )을 설정하여 해당 값 미만의 변수만을 사용하여 다시 로지스틱 회귀분석을 구축하는 것도 가능함 (주로  $\alpha = 0.05$  사용)

- ✓ 유의확률 p-value

- 로지스틱 회귀분석에서 해당 변수가 통계적으로 유의미한지 여부를 알려주는 지표
    - 0에 가까울수록 모델링에 중요한 변수이며, 1에 가까울수록 유의미하지 않은 변수임
    - 특정 유의수준( $\alpha$ )을 설정하여 해당 값 미만의 변수만을 사용하여 다시 로지스틱 회귀분석을 구축하는 것도 가능함 (주로  $\alpha = 0.05$  사용)

# 로지스틱 회귀분석: 해석

- 로지스틱 회귀분석 결과 및 해석

✓ 승산 비율 Odds Ratio

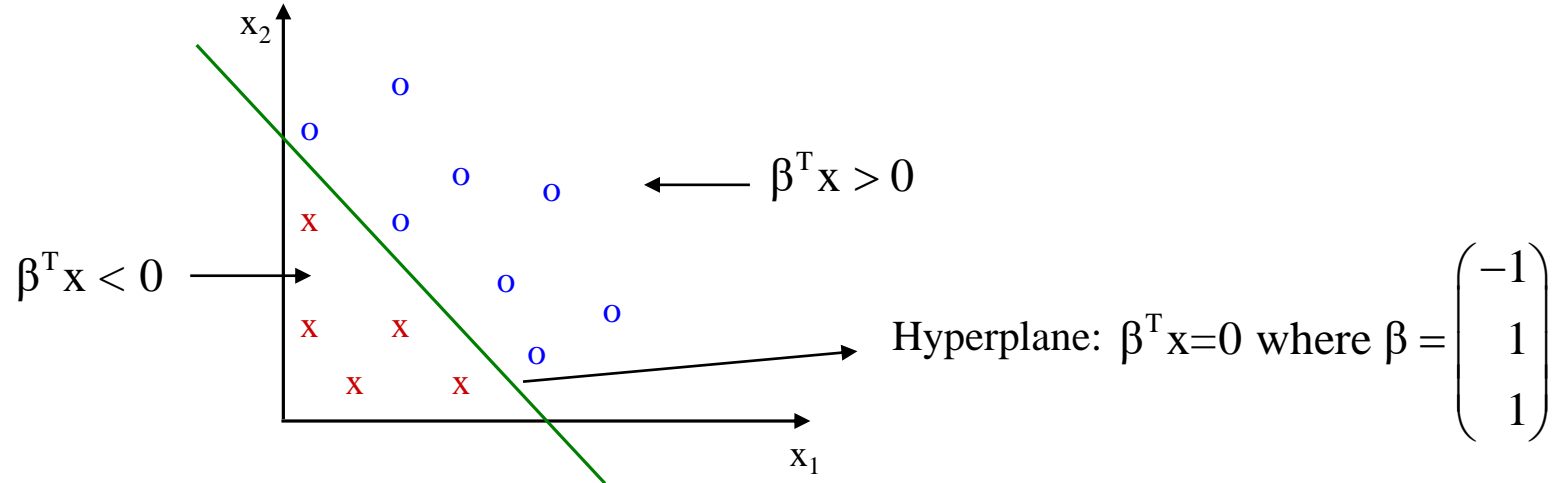
- 나머지 변수는 모두 고정시킨 상태에서 한 변수를 1만큼 증가시켰을 때 변화하는 Odds의 비율

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

# 로지스틱 회귀분석: 해석

- Geometric interpretation

✓ 로지스틱 회귀분석은  $d$ 차원의 데이터를 구분하는  $(d-1)$ 차원의 초평면을 찾는 것으로 이해할 수 있음



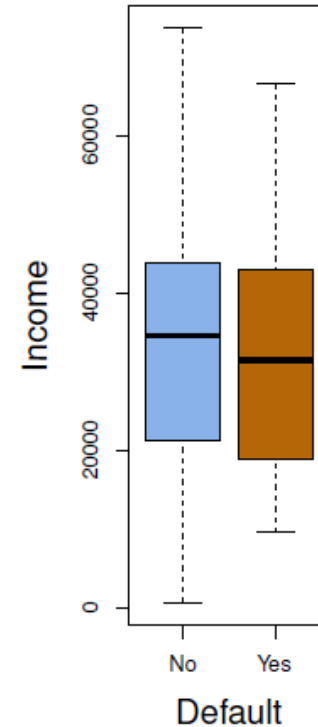
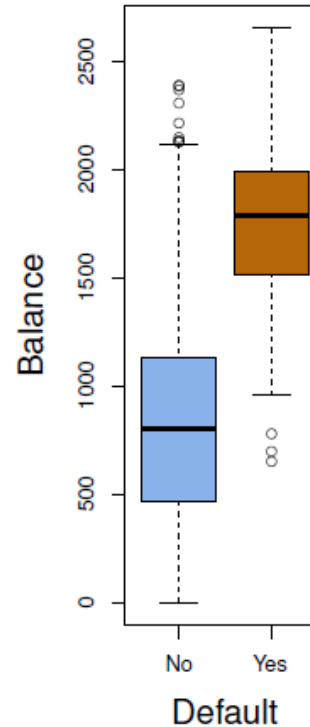
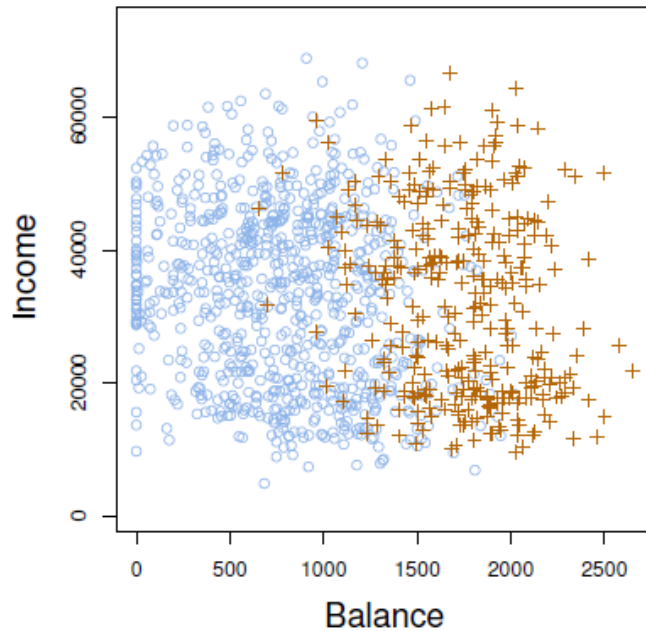
## Classifier

$$y = \frac{1}{(1 + \exp(-\beta^T \mathbf{x}))}$$

$$\begin{pmatrix} y \rightarrow 1 & \text{if } \beta^T \mathbf{x} \rightarrow \infty \\ y = \frac{1}{2} & \text{if } \beta^T \mathbf{x} = 0 \\ y \rightarrow 0 & \text{if } \beta^T \mathbf{x} \rightarrow -\infty \end{pmatrix}$$

# 로지스틱 회귀분석: 예시

- 신용카드 연체 예측



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

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

# 로지스틱 회귀분석: 예시

- 신용카드 연체 예측: 단변량 로지스틱 회귀분석

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

What is our estimated probability of **default** for someone with a balance of \$1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-10.6513 + 0.0055 \times 1000}}{1 + e^{-10.6513 + 0.0055 \times 1000}} = 0.006$$

With a balance of \$2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-10.6513 + 0.0055 \times 2000}}{1 + e^{-10.6513 + 0.0055 \times 2000}} = 0.586$$

# 로지스틱 회귀분석: 예시

- 신용카드 연체 예측: 다변량 로지스틱 회귀분석

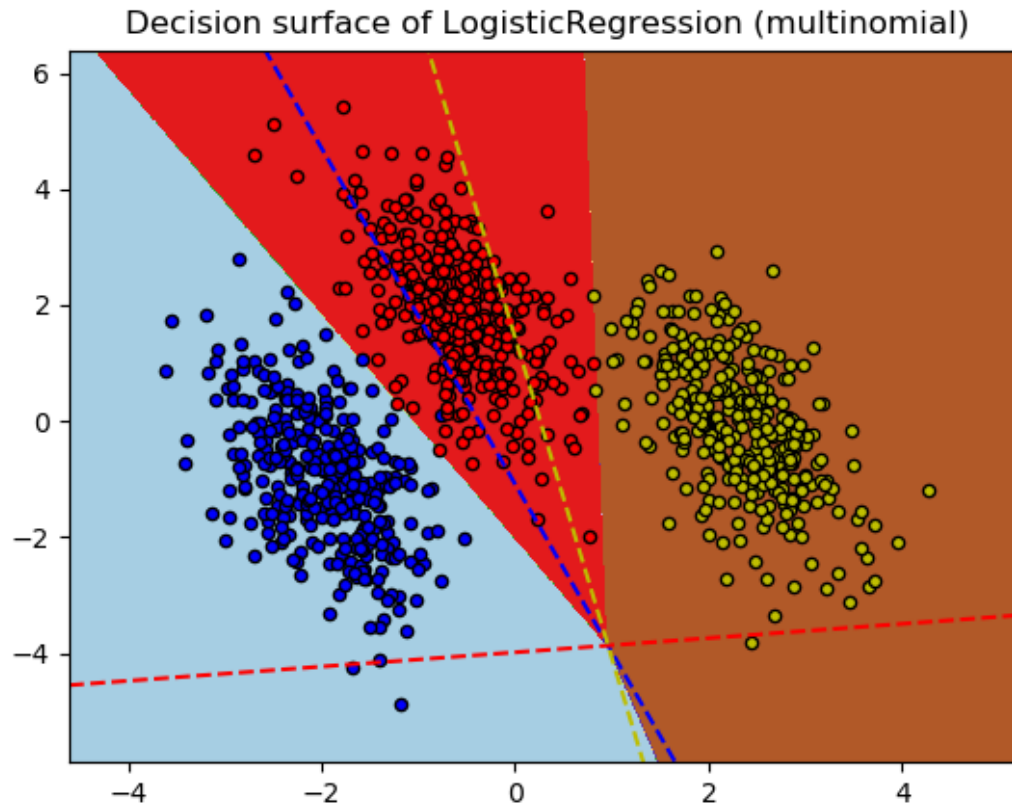
$$\log \left( \frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student [Yes]	-0.6468	0.2362	-2.74	0.0062

# 다항 로지스틱 회귀분석

- 지금까지의 로지스틱 회귀분석은 이범주 분류<sup>Binary classification</sup>를 풀기 위한 방식임
  - ✓ Q) 범주가 3개 이상인 다범주 분류에는 로지스틱 회귀분석을 어떻게 적용할 수 있을까?



# 다항 로지스틱 회귀분석

- 다항 로지스틱 회귀분석

- ✓ 기준<sup>Baseline</sup>이 되는 범주를 설정하고 이 범주 대비 다른 범주가 발생할 로그 승산을 회귀식으로 추정

- ✓ 예시) 범주가 3개인 분류 문제의 경우 아래 두 개의 회귀식에 대한 회귀 계수를 추정

- 범주 3 대비 범주 1의 발생 확률에 대한 로지스틱 회귀분석

$$\log\left(\frac{p(y=1)}{p(y=3)}\right) = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 \cdots + \beta_{1d}x_d = \beta_1^T \cdot \mathbf{x}$$

- 범주 3 대비 범주 2의 발생 확률에 대한 로지스틱 회귀분석

$$\log\left(\frac{p(y=2)}{p(y=3)}\right) = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 \cdots + \beta_{2d}x_d = \beta_2^T \cdot \mathbf{x}$$



# 다항 로지스틱 회귀분석

- 다항 로지스틱 회귀분석

✓ 왜 범주는 3개인데 2개의 모형만 학습하는가? (일반화하면 K개의 범주가 있을 때, (K-1)개의 모형만 학습하는 이유는?

- 각 범주에 속할 확률의 합은 항상 1이므로 나머지 K번째 범주에 대한 확률은 자동으로 산출됨

$$\frac{p(y=1)}{p(y=3)} = e^{\beta_1^T \cdot \mathbf{x}} \qquad \frac{p(y=2)}{p(y=3)} = e^{\beta_2^T \cdot \mathbf{x}}$$

$$p(y=1) + p(y=2) + p(y=3) = 1$$

$$p(y=3) \times e^{\beta_1^T \cdot \mathbf{x}} + p(y=3) \times e^{\beta_2^T \cdot \mathbf{x}} + p(y=3) = 1$$

$$p(y=3) = \frac{1}{1 + e^{\beta_1^T \cdot \mathbf{x}} + e^{\beta_2^T \cdot \mathbf{x}}}$$

# 다항 로지스틱 회귀분석

## • 다항 로지스틱 회귀분석에서의 회귀계수 분석

✓ 개별 모형에 대해서 회귀 계수와 이에 대한 유의확률을 산출할 수 있음

- Total phenols, Flavanoids, Monflavanoid penols, Hue, OD280~ 변수는 1 vs. 3, 2 vs. 3에서 모두 유의미한 변수로 나타남
- Ash., Proanthocyanins 변수는 범주 1과 3을 구분할 때는 유의미하지 않으나 2와 3을 구분할 때 매우 유의미함

	1 vs 3		2 vs 3	
	Coefficient	p-value	Coefficient	p-value
(Intercept)	-223.7894	0.0000	340.9326	0.0000
Alcohol.2	19.6193	0.7880	-35.2596	0.6828
Malic.acid.	1.0581	0.9228	-0.3022	0.9899
Ash.	14.6800	0.3881	-204.7437	0.0000
Alcalinity.of.ash.	-20.3881	0.8815	-2.2832	0.9864
Magnesium.	2.0553	0.9975	2.1132	0.9974
Total.phenols.	-169.4205	0.0000	-40.3325	0.0000
Flavanoids.	193.7935	0.0000	16.2013	0.0188
Nonflavanoid.phenols	93.5409	0.0000	214.1837	0.0000
Proanthocyanins.	15.5178	0.1453	115.3184	0.0000
Color.intensity.	-16.6775	0.4212	-11.5066	0.7671
Hue	-50.0008	0.0000	352.7617	0.0000
OD280.OD315.of.diluted.wines.	75.2435	0.0000	84.2914	0.0000
Proline.	-0.0120	1.0000	-0.2899	0.9999





# 분류 모형 성능 평가

- 예시: 성별 분류

✓ 한 사람의 체지방률만을 이용하여 남성/여성 분류

									
10.0	21.7	8.9	19.9	23.4	28.9	15.7	21.6	21.5	23.2

✓ 단순 분류기: 체지방률이 20보다 크면 여성으로, 작으면 남성으로 분류










									
10.0	21.7	8.9	19.9	23.4	28.9	15.7	21.6	21.8	23.2
M	F	M	M	F	F	M	F	F	F

✓ 위 분류기의 성능을 어떻게 평가할 것인가?

# 분류 모형 성능 평가

- 정오 행렬 Confusion Matrix

✓ 실제 범주와 예측된 범주를 이용하여 생성한 2X2 행렬

									
10.0	21.7	8.9	19.9	23.4	28.9	15.7	21.6	21.5	23.2
M	F	M	M	F	F	M	F	F	F

✓ 위 결과에 대한 정오 행렬은 다음과 같이 생성됨

Confusion Matrix		Predicted	
		F	M
Actual	F	4	1
	M	2	3

# 분류 모형 성능 평가

- 정오 행렬 Confusion Matrix

✓ 정오행렬을 통해 다음과 같이 다양한 분류 성능 평가 지표를 계산할 수 있음

Confusion Matrix		Predicted	
		1(+)	0(-)
Actual	1(+)	$n_{11}$	$n_{10}$
	0(-)	$n_{01}$	$n_{00}$

- 민감도(Sensitivity), true positive, 재현율(recall) =  $n_{11}/(n_{11}+n_{10})$
- 특이도(Specificity, true negative) =  $n_{00}/(n_{01}+n_{00})$
- 정밀도(Precision) =  $n_{11}/(n_{11}+n_{01})$
- 제1종 오류(Type I error, false negative) =  $n_{10}/(n_{11}+n_{10})$
- 제2종 오류(Type II error, false positive) =  $n_{01}/(n_{01}+n_{00})$

# 분류 모형 성능 평가

- 정오 행렬 Confusion Matrix

✓ 정오행렬을 통해 다음과 같이 다양한 분류 성능 평가 지표를 계산할 수 있음

Confusion Matrix		Predicted	
		1(+)	0(-)
Actual	1(+)	$n_{11}$	$n_{10}$
	0(-)	$n_{01}$	$n_{00}$

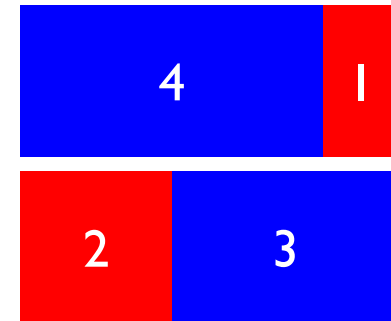
- 오분류율(Misclassification error) =  $(n_{01} + n_{10}) / (n_{11} + n_{10} + n_{01} + n_{00})$
- 정분류율(Accuracy = 1 - misclassification error) =  $(n_{11} + n_{00}) / (n_{11} + n_{10} + n_{01} + n_{00})$
- 균형 정확도 (Balanced correction rate) = 
$$\sqrt{\frac{n_{11}}{n_{11} + n_{10}} \cdot \frac{n_{00}}{n_{01} + n_{00}}}$$
- F1 measure (정밀도와 재현율의 조화평균) = 
$$F1 \text{ measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

# 분류 모형 성능 평가

- 정오 행렬 Confusion Matrix

✓ 이전 예시에서 여성(F)을 I(+) 범주로 정의할 경우

Confusion Matrix		Predicted	
		F	M
Actual	F	4	1
	M	2	3



- Sensitivity:  $4/5 = 0.8$ , Specificity:  $3/5 = 0.6$
- Recall:  $4/5 = 0.8$ , Precision:  $4/6 = 0.67$
- Type I error:  $1/5 = 0.2$ , Type II error:  $2/5 = 0.4$
- Misclassification error:  $(1+2)/(4+1+2+3) = 0.3$ , accuracy =  $0.7$
- Balanced correction rate:  $\sqrt{0.8 \times 0.6} = 0.69$
- F1 measure:  $(2 \times 0.8 \times 0.67) / (0.8 + 0.67) = 0.85$

# 분류 모형 성능 평가

- 분류 알고리즘의 Cut-off 설정

✓ 새로운 분류기: 체지방률이  $\theta$ 보다 크면 여성으로 분류



✓ 레코드들을 체지방률의 내림차순으로 정렬



✓ 분류를 위한 최적의 cut-off를 어떻게 설정할 것인가?



# 분류 모형 성능 평가

- 분류 알고리즘의 Cut-off 설정

✓ 다양한 Cut-off에 따른 분류 성능 비교

No.	체지방률	성별
1	28.6	F
2	25.4	M
3	24.2	F
4	23.6	F
5	22.7	F
6	21.5	M
7	19.9	F
8	15.7	M
9	10.0	M
10	8.9	M

▪ If  $\theta = 24$ ,

Confusion Matrix		Predicted	
		F	M
Actual	F	2	3
	M	1	4

- Misclassification error: 0.4
- Accuracy: 0.6
- Balanced correction rate: 0.57
- F1 measure = 0.5

# 분류 모형 성능 평가

- 분류 알고리즘의 Cut-off 설정

✓ 다양한 Cut-off에 따른 분류 성능 비교

No.	체지방률	성별
1	28.6	F
2	25.4	M
3	24.2	F
4	23.6	F
5	22.7	F
6	21.5	M
7	19.9	F
8	15.7	M
9	10.0	M
10	8.9	M

▪ If  $\theta = 22$ ,

Confusion Matrix		Predicted	
		F	M
Actual	F	4	1
	M	1	4

- Misclassification error: 0.2
- Accuracy: 0.8
- Balanced correction rate: 0.8
- F1 measure = 0.8

# 분류 모형 성능 평가

- 분류 알고리즘의 Cut-off 설정

✓ 다양한 Cut-off에 따른 분류 성능 비교

No.	체지방률	성별
1	28.6	F
2	25.4	M
3	24.2	F
4	23.6	F
5	22.7	F
6	21.5	M
7	19.9	F
8	15.7	M
9	10.0	M
10	8.9	M

▪ If  $\theta = 18$ ,

Confusion Matrix		Predicted	
		F	M
Actual	F	5	0
	M	2	3

- Misclassification error: 0.2
- Accuracy: 0.8
- Balanced correction rate: 0.77
- F1 measure = 0.83

# 분류 모형 성능 평가

- 분류 알고리즘의 Cut-off 설정

- ✓ 일반적으로 분류 알고리즘은 특정 범주에 속할 확률(probability)이나 우도(likelihood) 값을 생성함
- ✓ 동일한 확률값 하에서도 Cut-off가 어떻게 설정되느냐에 따라서 분류 성능이 크게 좌우되는 상황이 발생할 수 있음
- ✓ 분류 알고리즘간의 정확한 비교를 위해서는 Cut-off에 독립적인 측정 지표가 필요함
- ✓ 리프트 도표(Lift charts), receiver operating characteristic (ROC) curve 등이 사용

# 분류 모형 성능 평가

- ROC Curve 예시

- ✓ Glass 불량 진단 문제:

- Glass의 불량(NG) 여부를 판별
    - 총 100장의 Glass 중 20장의 Glass가 불량
    - 불량 확률: 0.2
    - Label: 1(NG), 0(G)

# 분류 모형 성능 평가

- 특정 분류 알고리즘에 의해 산출된 NG 범주에 속할 확률과 실제 Label 정보

Glass	P(NG)	Label	Glass	P(NG)	Label	Glass	P(NG)	Label	Glass	P(NG)	Label
1	0.976	1	26	0.716	1	51	0.41	0	76	0.186	0
2	0.973	1	27	0.676	0	52	0.406	1	77	0.183	0
3	0.971	0	28	0.672	0	53	0.378	0	78	0.178	0
4	0.967	1	29	0.662	0	54	0.376	0	79	0.176	0
5	0.937	0	30	0.647	0	55	0.362	0	80	0.173	0
6	0.936	1	31	0.64	1	56	0.355	0	81	0.17	0
7	0.929	1	32	0.625	0	57	0.343	0	82	0.133	0
8	0.927	0	33	0.624	0	58	0.338	0	83	0.12	0
9	0.923	1	34	0.613	1	59	0.335	0	84	0.119	0
10	0.898	0	35	0.606	0	60	0.334	0	85	0.112	0
11	0.863	1	36	0.604	0	61	0.328	0	86	0.093	0
12	0.862	1	37	0.601	0	62	0.313	0	87	0.086	0
13	0.859	0	38	0.594	0	63	0.285	1	88	0.079	0
14	0.855	0	39	0.578	0	64	0.274	0	89	0.071	0
15	0.847	1	40	0.548	0	65	0.273	0	90	0.069	0
16	0.845	1	41	0.539	1	66	0.272	0	91	0.047	0
17	0.837	0	42	0.525	1	67	0.267	0	92	0.029	0
18	0.833	0	43	0.524	0	68	0.265	0	93	0.028	0
19	0.814	0	44	0.514	0	69	0.237	0	94	0.027	0
20	0.813	0	45	0.51	0	70	0.217	0	95	0.022	0
21	0.793	1	46	0.509	0	71	0.213	0	96	0.019	0
22	0.787	0	47	0.455	0	72	0.204	1	97	0.015	0
23	0.757	1	48	0.449	0	73	0.201	0	98	0.01	0
24	0.741	0	49	0.434	0	74	0.2	0	99	0.005	0
25	0.737	0	50	0.414	0	75	0.193	0	100	0.002	0

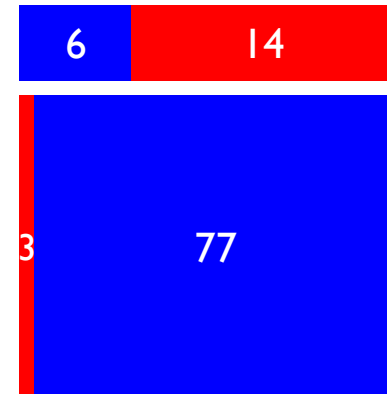
# 분류 모형 성능 평가

- 정오행렬

- ✓ Cut-off를 0.9로 설정할 경우

- NG if  $P(NG) > 0.9$ , else G

Confusion Matrix		Predicted	
		M	B
Actual	M	6	14
	B	3	77



- Misclassification error = 0.17
- Accuracy = 0.83

- ✓ 이 모델은 우수한 분류 모델인가?

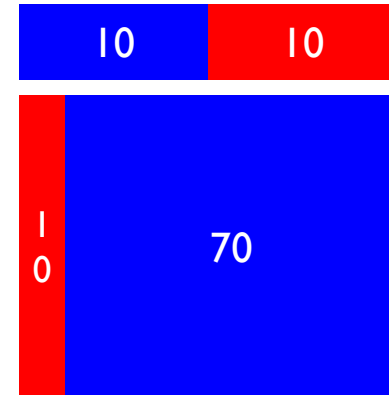
# 분류 모형 성능 평가

- 정오행렬

- ✓ Cut-off를 0.9로 설정할 경우

- NG if  $P(NG) > 0.8$ , else G

Confusion Matrix		Predicted	
		M	B
Actual	M	10	10
	B	10	70



- Misclassification error = 0.20
  - Accuracy = 0.80

- ✓ 이 모델은 이전 모델보다 열등한 모델인가?



# 분류 모형 성능 평가

- ROC 생성 절차

- ✓ 모든 개체를  $P(\text{interesting class})$ 를 기준으로 내림차순 정렬
- ✓ 가능한 모든 Cut-off 경우에 대해 True Positive Rate와 False Positive Rate를 계산
  - $P(NG)$ 에 동물이 없을 경우 이론적으로 101개의 cut-off 설정이 가능
- ✓ X축이 False Positive Rate, Y축이 True Positive Rate가 되는 2차원 그래프 도시

# 분류 모형 성능 평가

- ROC 생성 절차

✓ 첫 번째 Cut-off 설정

Glass	P(NG)	Label
1	0.976	1
2	0.973	1
3	0.971	0
4	0.967	1
5	0.937	0
⋮	⋮	⋮

Confusion Matrix		예측	
		NG	G
실제	NG	0	20
	G	0	80

$$\text{TPR} = \frac{0}{20} = 0$$

$$\text{FPR} = \frac{0}{80} = 0$$

# 분류 모형 성능 평가

- ROC 생성 절차

✓ 두 번째 Cut-off 설정

Glass	P(NG)	Label	TPR	FPR
			0	0
1	0.976	1		
2	0.973	1		
3	0.971	0		
4	0.967	1		
5	0.937	0		
⋮	⋮	⋮	⋮	⋮

Confusion Matrix		예측	
		NG	G
실제	NG	1	19
	G	0	80

$$\text{TPR} = \frac{1}{20} = 0.05$$

$$\text{FPR} = \frac{0}{80} = 0$$

# 분류 모형 성능 평가

- ROC 생성 절차

✓ 세 번째 Cut-off 설정

Glass	P(NG)	Label	TPR	FPR
			0	0
1	0.976	1	0.05	0
2	0.973	1		
3	0.971	0		
4	0.967	1		
5	0.937	0		
⋮	⋮	⋮	⋮	⋮

Confusion Matrix		예측	
		NG	G
실제	NG	2	18
	G	0	80

$$\text{TPR} = \frac{2}{20} = 0.10$$

$$\text{FPR} = \frac{0}{80} = 0$$

# 분류 모형 성능 평가

- ROC 생성 절차

✓ 네 번째 Cut-off 설정

Glass	P(NG)	Label	TPR	FPR
			0.00	0.00
1	0.976	1	0.05	0.00
2	0.973	1	0.10	0.00
3	0.971	0		
4	0.967	1		
5	0.937	0		
⋮	⋮	⋮	⋮	⋮

Confusion Matrix		예측	
		NG	G
실제	NG	2	18
	G	1	79

$$\text{TPR} = \frac{2}{20} = 0.10$$

$$\text{FPR} = \frac{1}{80} = 0.0125$$

# 분류 모형 성능 평가

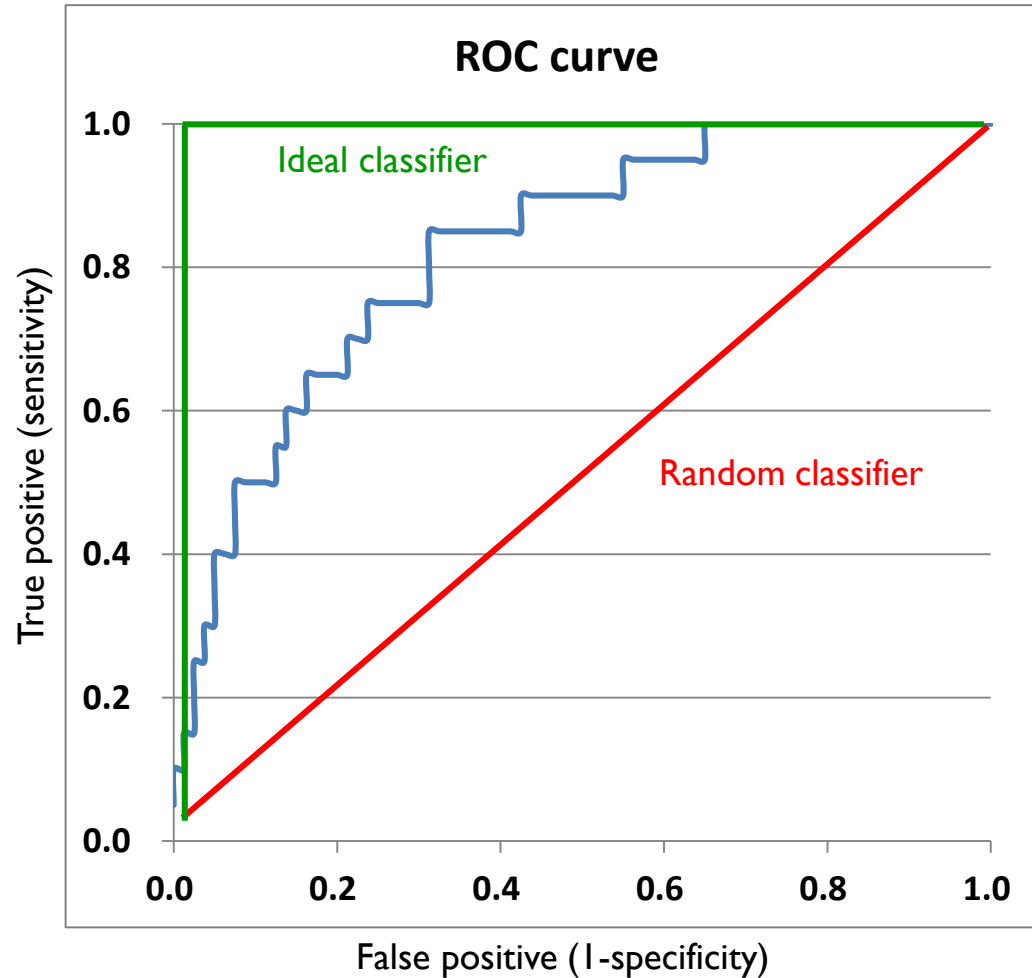
- ROC 생성 절차

- ✓ 모든 가능한 Cut-off 값에 대한 TPR/FPR 계산 완료
- ✓ FPR을 x축으로 하고, TPR을 y축으로 하는 그래프 생성

Glass	P(NG)	Label	TPR	FPR
			0.000	0.000
1	0.976	1	0.050	0.000
2	0.973	1	0.100	0.000
3	0.971	0	0.100	0.013
4	0.967	1	0.150	0.013
5	0.937	0	0.150	0.025
6	0.936	1	0.200	0.025
7	0.929	1	0.250	0.025
8	0.927	0	0.250	0.038
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
96	0.019	0	1.000	0.950
97	0.015	0	1.000	0.963
98	0.01	0	1.000	0.975
99	0.005	0	1.000	0.988
100	0.002	0	1.000	1.000

# 분류 모형 성능 평가

- ROC Curve 범위



# 분류 모형 성능 평가

- Area Under ROC Curve (AUROC)

- ✓ ROC curve 아래의 면적
- ✓ 이상적인 분류기는 1의 값을 갖고, 무작위 분류기는 0.5의 값을 가짐
- ✓ Cut-off에 독립적인 알고리즘 성능 평가 지표로 사용될 수 있음

