Chap.3 로지스틱 회귀 (Logistic Regression)

방 수 식 교수

(bang@tukorea.ac.kr)

한국공학대학교 전자공학부

2024년도 1학기

머신러닝실습 & 인공지능설계실습1

확률 기초

- 확률변수random variable
  - 예) 윷놀이

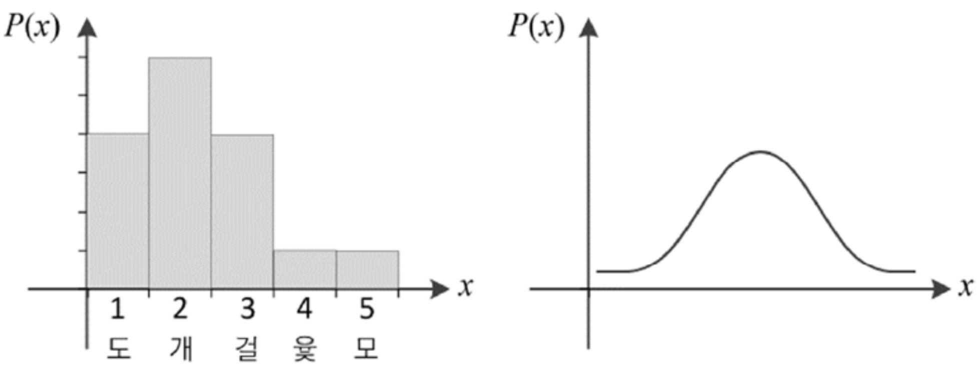
  - 다섯 가지 경우 중 한 값을 갖는 확률변수 x
  - x의 정의역은 {도, 개, 걸, 윷, 모}

확률 기초

확률분포probability distribution

윷의 앞면과 뒷면이 나올 확률이 각각 50%라고 가정

$$P(x=\text{도})=\frac{4}{16}, P(x=\text{개})=\frac{6}{16}, P(x=\text{걸})=\frac{4}{16}, P(x=\text{윷})=\frac{1}{16}, P(x=\text{모})=\frac{1}{16}$$



(a) 이산인 경우의 확률질량함수

(b) 연속인 경우의 확률밀도함수 *probability density function (PDF)*

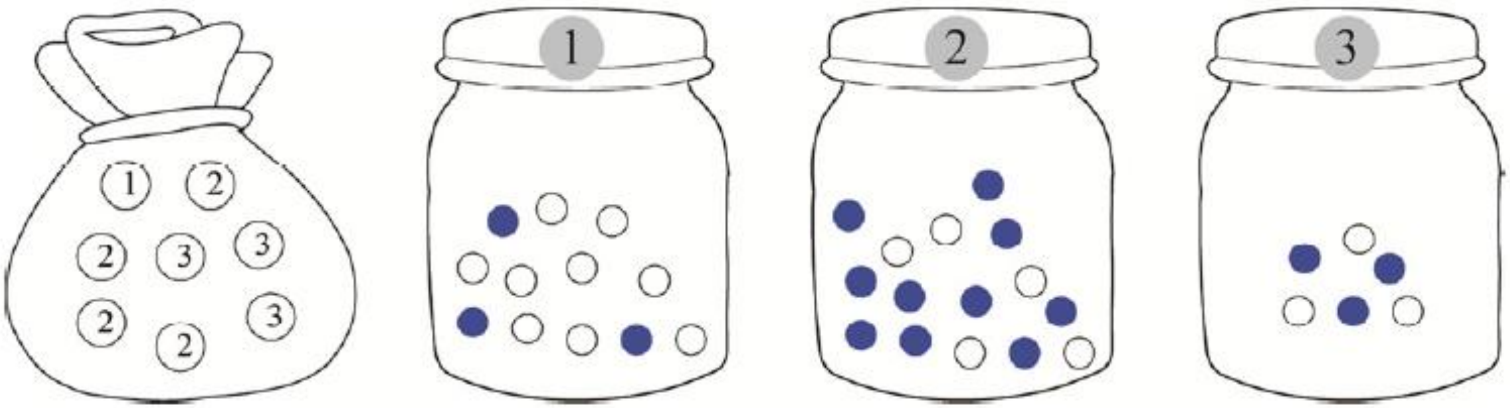
확률벡터random vector

- 예) Iris에서 확률벡터 $\mathbf{x}$는 4차원 $\mathbf{x}=(x_1,x_2,x_3,x_4)^{\mathrm{T}}=$ (꽃받침 길이, 꽃받침 너비$_1$, 꽃잎

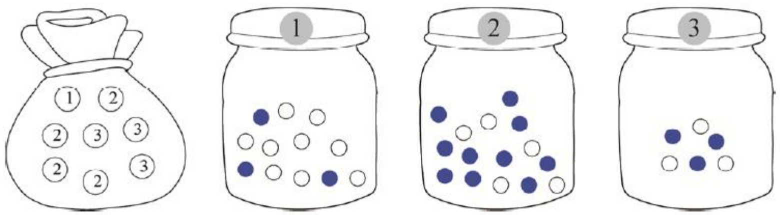
확률 기초

간단한 확률실험 장치

- 주머니에서 번호를 뽑은 다음, 번호에 따라 해당 병에서 공을 뽑고 색을 관찰함
- 번호를 y, 공의 색을 x라는 확률변수로 표현하면 정의역은 $y \in \{①,②,③\}$, $x \in \{파랑, 하양\}$



확률 기초



곱 규칙과 합 규칙

- ①번 카드를 뽑을 확률은 $P(y=①)=P(①)=1/8$
- **카드는 ①번, 공은 하양일 확률**은 $P(y=①, x=\text{하양})=P(①,\text{하양})$ ← 결합확률

$$P(y=①, x=\text{하양}) = P(x=\text{하양}|y=①)P(y=①) = \frac{9}{12}\frac{1}{8} = \frac{3}{32}$$

$P(y,x)$: 사건 y과 사건 x가 동시에 일어날 확률 (Joint probability, 결합 확률)

$P(x|y)$:사건 y가 일어난 상태에서 사건 x가 일어날 확률 (Conditional probability, 조건부 확률)

$P(y)$:사건 y가 일어날 확률 (Marginal probability, 주변 확률)

- 곱 규칙

$$\text{곱 규칙: } P(y,x) = P(x|y)P(y)$$

- **하얀 공이 뽑힐 확률**

$$P(\text{하양}) = P(\text{하양}|①)P(①) + P(\text{하양}|②)P(②) + P(\text{하양}|③)P(③)$$
$$= \frac{9}{12}\frac{1}{8} + \frac{5}{15}\frac{4}{8} + \frac{3}{6}\frac{3}{8} = \frac{43}{96}$$

- 합 규칙

$$\text{합 규칙: } P(x) = \sum_{y} P(y,x) = \sum_{y} P(x|y)P(y)$$

Bayes' Theorem

- Bayes' Theorem (베이즈 정리)

$$P(y,x) = P(x|y)P(y) = P(x,y) = P(y|x)P(x)$$

$$\longrightarrow \quad P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

"하얀 공이 나왔다는 사실만 알고 어느 병에서 나왔는지 모르는데, 어느 병인지 추정하라."

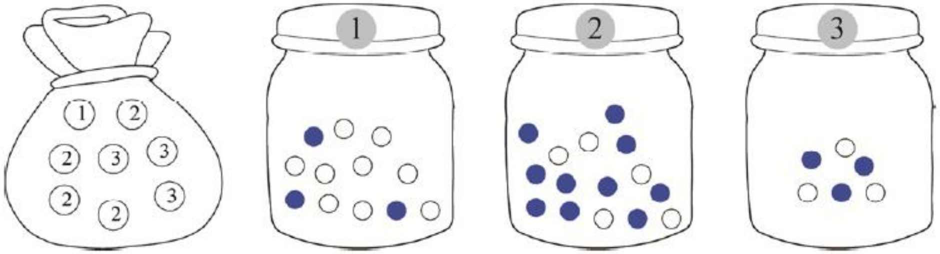
$P(y = ① \mid x = \text{하양})$

$P(y = ② \mid x = \text{하양})$ *중에 최대확률을 만드는 y가 가장 유력*

$P(y = ③ \mid x = \text{하양})$

Likelihood (우도, 가능도)

$$\hat{y} = \underset{y}{\operatorname{argmax}} \boxed{P(y|x)}$$



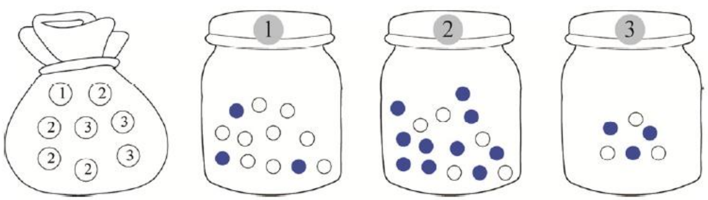
Bayes' Theorem

베이즈 정리

- 베이즈 정리를 적용하면, $\hat{y} = \underset{y}{\operatorname{argmax}} P(y|x = 하양) = \underset{y}{\operatorname{argmax}} \frac{P(x = 하양|y)P(y)}{P(x = 하양)}$

$$\boxed{P(y|x) = \frac{P(x|y)P(y)}{P(x)}}$$

- 세 가지 경우에 대해 확률을 계산하면,



$$P(①|하양) = \frac{P(하양|①)P(①)}{P(하양)} = \frac{\frac{9}{12}\frac{1}{8}}{\frac{43}{96}} = \frac{9}{43}$$

$$P(②|하양) = \frac{P(하양|②)P(②)}{P(하양)} = \frac{\frac{5}{15}\frac{4}{8}}{\frac{43}{96}} = \frac{16}{43}$$

$$P(③|하양) = \frac{P(하양|③)P(③)}{P(하양)} = \frac{\frac{3}{6}\frac{3}{8}}{\frac{43}{96}} = \frac{18}{43}$$

→ ③번 병일 확률이 가장 높음

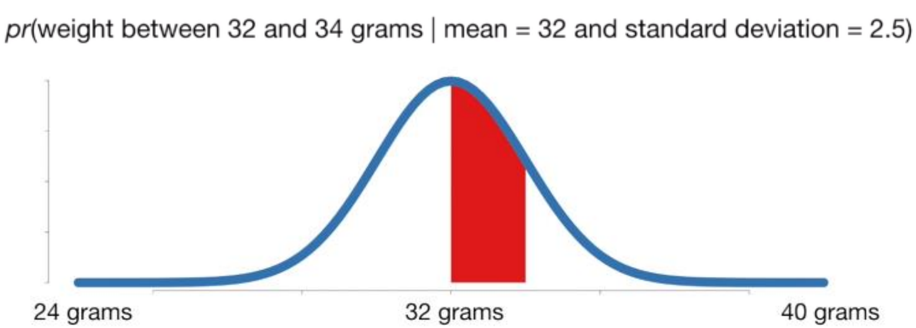
Probability vs Likelihood

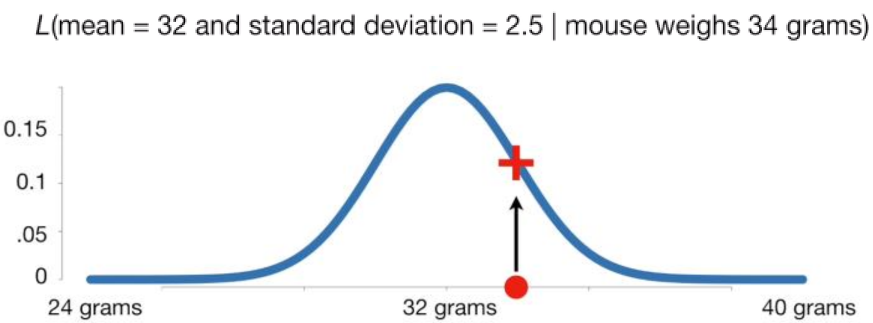
지능형 예측·진단 연구실
Intelligent Prognostics & Diagnostics Lab.

- X: 관측값, D: 확률분포 일 때,

- Conditional Probability: $P(X \mid D)$
  - D 확률 분포에서의 X 사건이 일어날 확률

- Likelihood: $P(D \mid X)$ or $L(D \mid X)$
  - X 사건이 D 확률분포에서 발생했을 확률

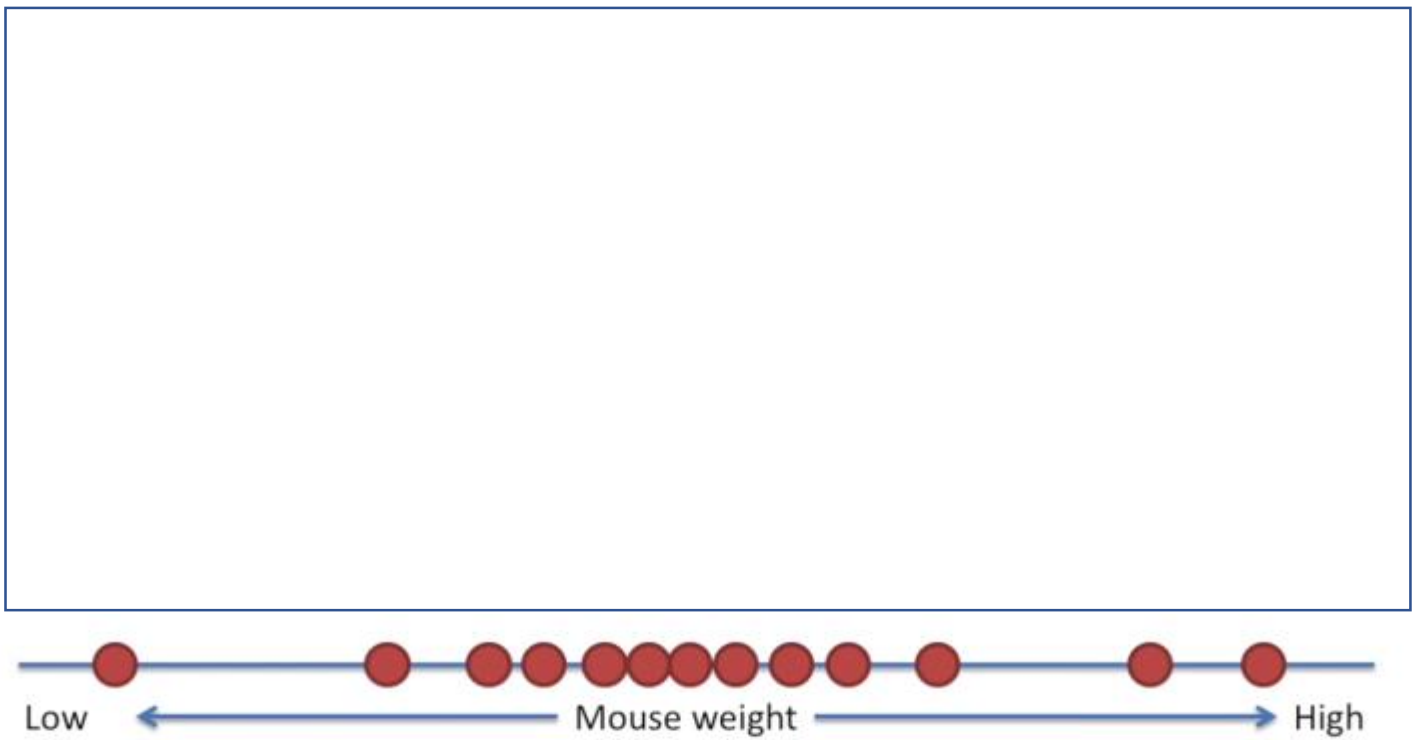




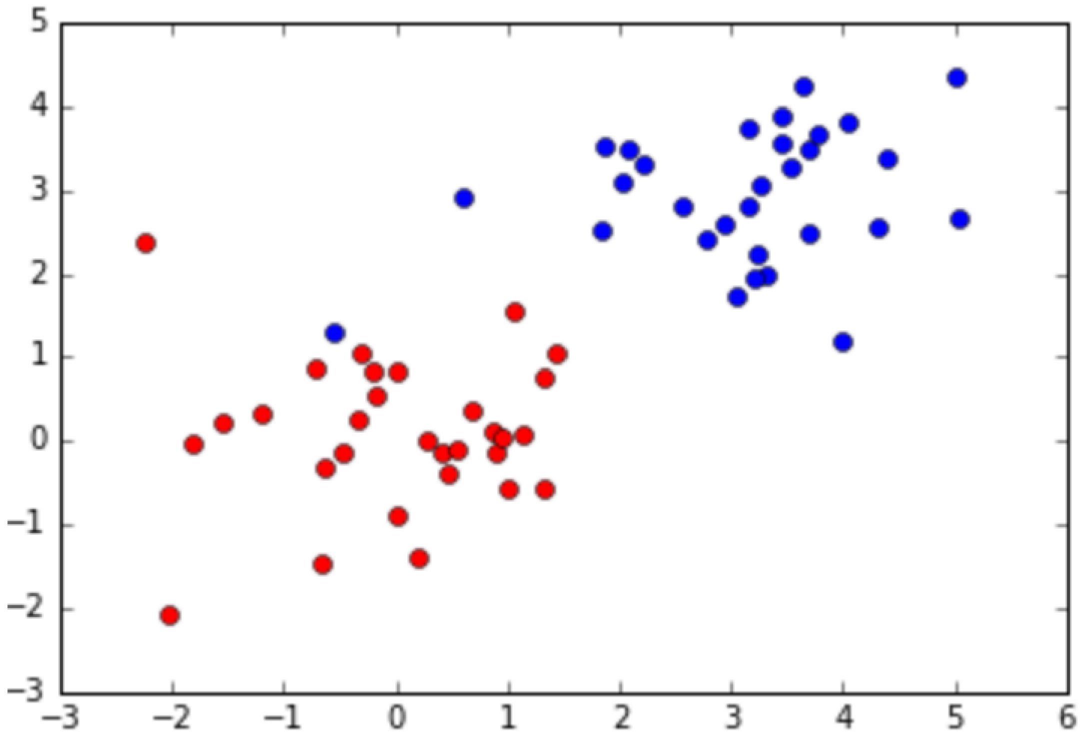
Maximum Likelihood

- Maximum Likelihood: 주어진 사건들(혹은 데이터)에 대해서 가장 합리적인 확률분포를 찾는 문제

Likelihood of observing the data



Regression vs Classification

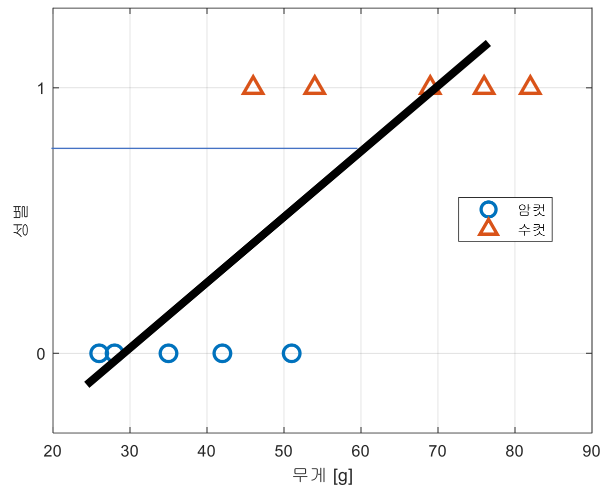


Classification Problem

지능형 예측·진단 연구실
Intelligent Prognostics & Diagnostics Lab.

- <u>입력에 대한 출력이 0 or 1 과 같이 특정 값으로 Discrete하게 나와야함.</u>
  - ➢ Linear Regression에서는 Continuous Space에서의 출력값을 가졌음.

Linear Regression

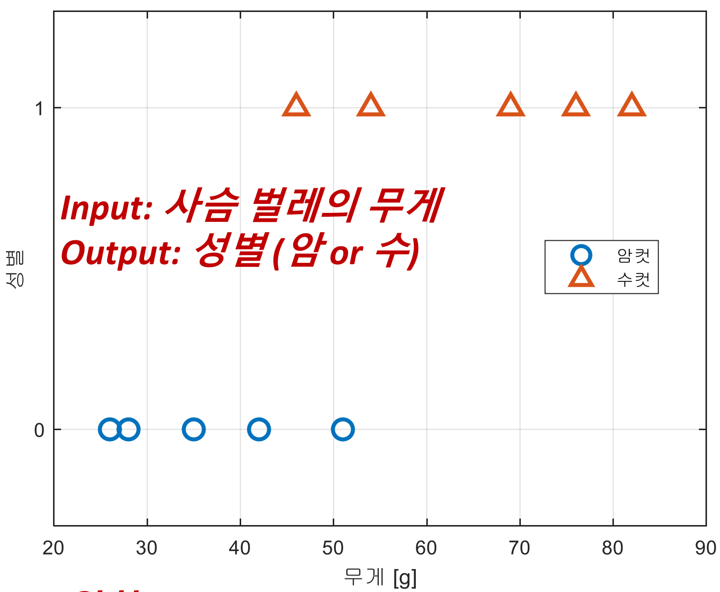


Input이 60g 일 때,
Linear Regression Model 출력 값은?
⇒ 0.85

어떻게 활용 가능할까?

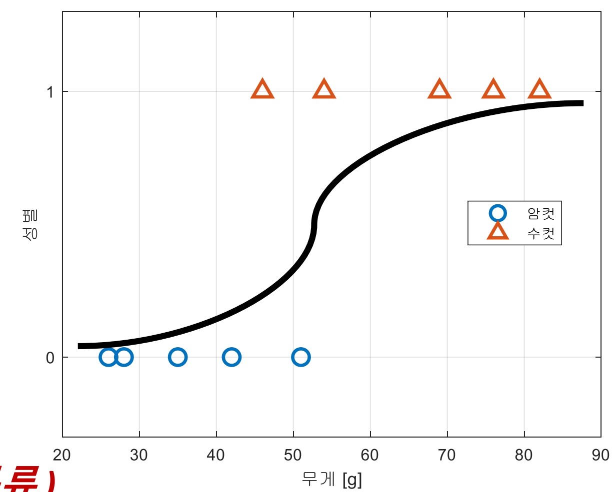
출력값 > 0.5 이면, 1
출력값 < 0.5 이면, 0

- 사슴벌레 10마리 관찰



원하는 ML Model?
⇒ Input을 넣으면 성별을 예측!!
⇒ 즉, 출력이 0 or 1: Binary Classification(이진 분류)

입력값에 대한 출력값을 확률로 생각하자!!!



확률개념을 적용하기 위해서는

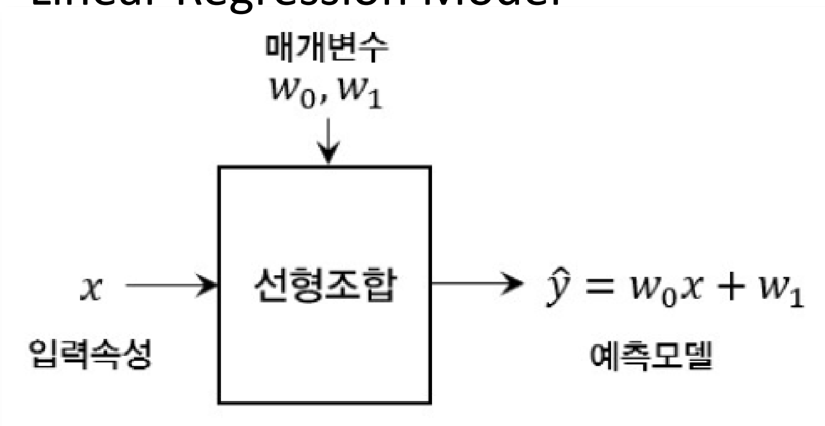
1) 실제 출력값(0,1)에 근접하면서
2) 0~1사이로 Bounded

출력값에 어떤 함수를 적용하면 좋을까?

Key Idea for Classification

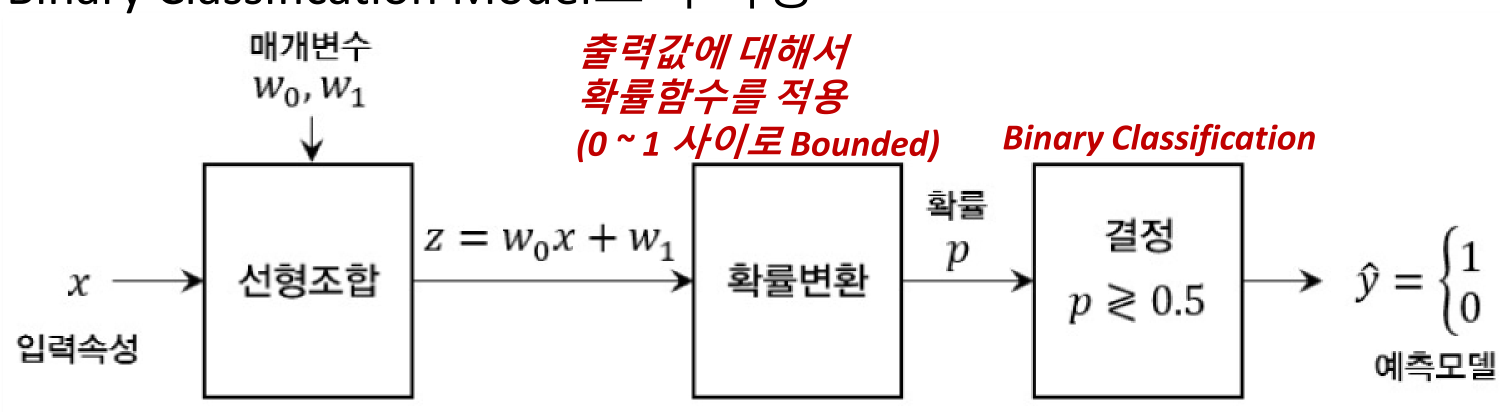
- Linear Regression Model을 Binary Classification Model로 확장

  - Linear Regression Model



Linear Model의 출력값은 -∞ ~ ∞

  - Binary Classification Model로의 확장



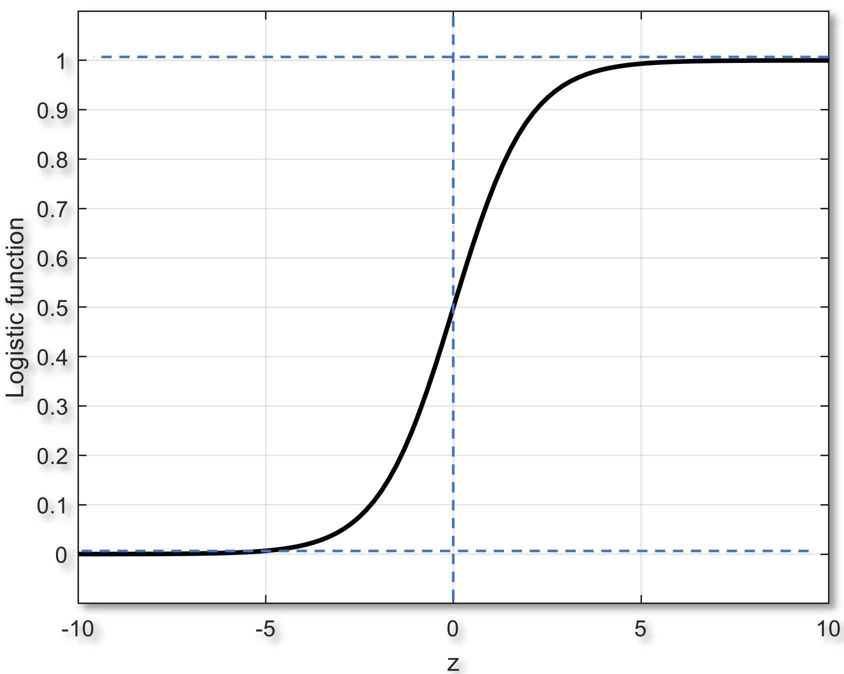
Sigmoid Function (Logistic Function)

지능형 예측·진단 연구실
Intelligent Prognostics & Diagnostics Lab.

- Sigmoid Function (Logistic function)
  - 정의
    - $f(z) = \frac{1}{1+e^{-z}}$

$$f(z) = \begin{cases} 0, & z \rightarrow -\infty \\ 0.5, & z = 0 \\ 1, & z \rightarrow \infty \end{cases}$$

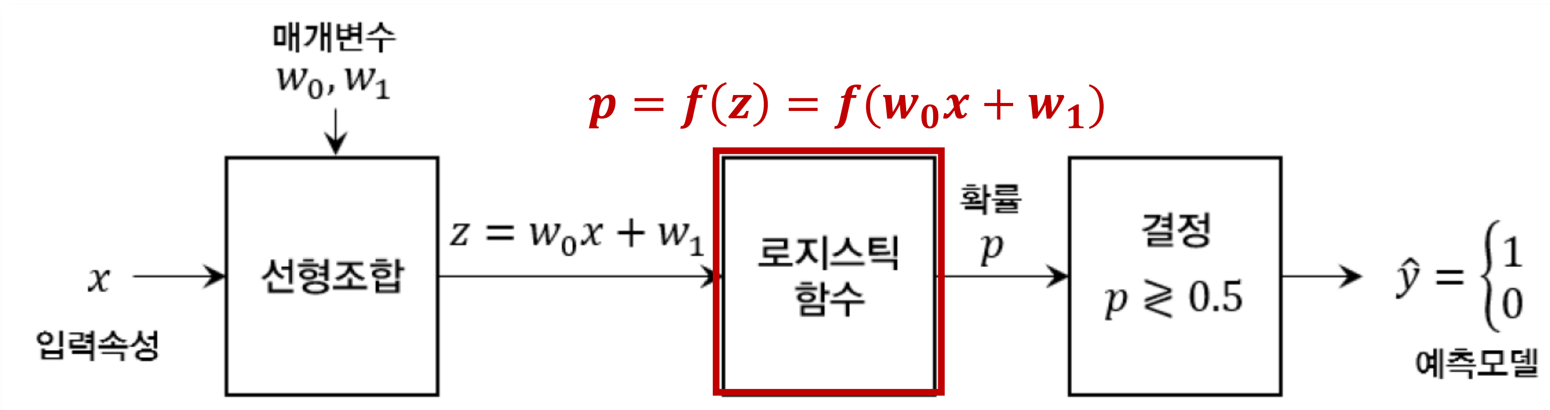
  - 특징
    - 임의의 값을 [0,1]의 값으로 변환
    - 우수한 미분 특성을 가짐
    - $\frac{df(z)}{dz} = f(z)(1 - f(z))$



Sigmoid Function을
"확률변환" 함수로 활용하자!

Logistic Regression (로지스틱 회귀)

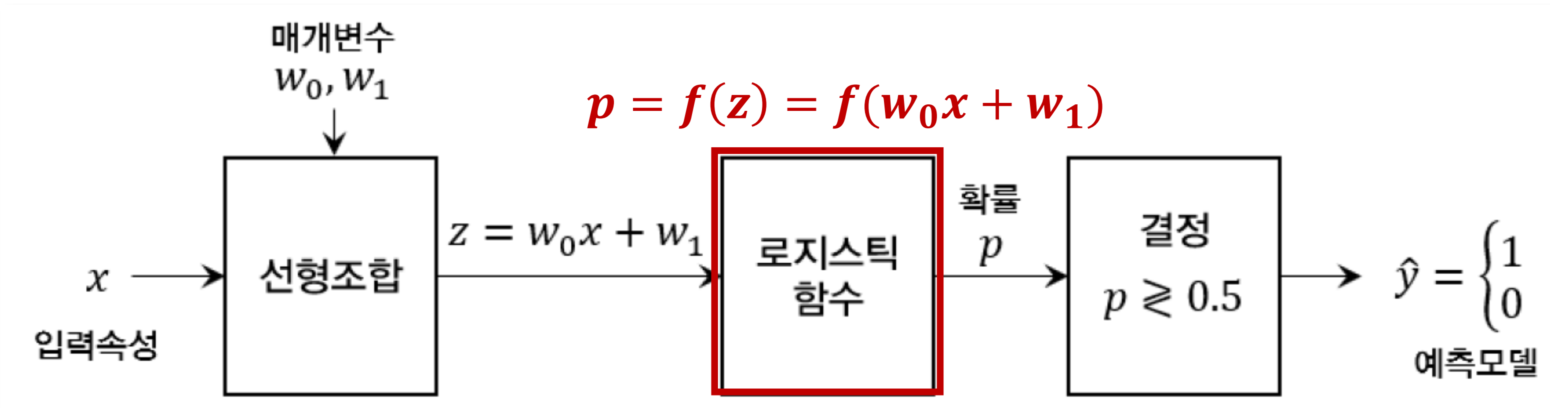
- **Logistic Regression Model의 목표**
  - ➢ 모든 입력 데이터들에 대한 최종 예측값 $\hat{y}$이 실제값 y와 같아야 한다.



- **Sigmoid 함수로 변환한 예측 확률 p**
  - 입력 x에 대한 출력 y가 “1”일 확률을 의미
    - ✓ $p = P(y = 1|x) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-w_0x-w_1}}$
  - 입력 x에 대한 출력 y가 “0”일 확률
    - ✓ $1 - p = P(y = 0|x) = \frac{e^{-z}}{1+e^{-z}} = \frac{e^{-w_0x-w_1}}{1+e^{-w_0x-w_1}}$

Logistic Regression (로지스틱 회귀)

- Logistic Regression Model의 목표
  - 모든 입력 데이터들에 대한 최종 예측값 $\hat{y}$이 실제값 y와 같아야 한다.



- **실제 y가 1일 때 => 확률 p는 1에 가까워야 한다. => p^y가 커야 한다.**
- **실제 y가 0일 때 => 확률 p는 0에 가까워야 한다. => $(1-p)^{1-y}$ 가 커야 한다.**

=> w_0, w_1, x 가 주어졌을 때, 예측값 $\hat{y}$이 실측값 y 일 확률: $P(y|x, w)$ = $p^y(1-p)^{1-y}$

조건부 확률

- 위의 모델에서 우리가 바꿀 수 있는 변수는?
  - Weight (w_0, w_1)

즉, 우리는 조건부 확률 $p^y(1-p)^{1-y}$ 이 **<u>최대</u>**가 되는 **<u>Weight 을 찾는 것이 목표</u>**

Logistic Regression (로지스틱 회귀)

지능형 예측·진단 연구실
Intelligent Prognostics & Diagnostics Lab.

- 다수의 데이터에 대한 조건부 확률
  - 개별 데이터에 대한 조건부 확률
    - $P(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = p_n^{y_n}(1-p_n)^{1-y_n},\ n = 0, 1, \cdots, N-1$

데이터 번호	입력	출력	사후확률	상태확률
0	$\boldsymbol{x}_0$	y_0	p_0	$P(y_0 \mid \boldsymbol{x}_0, \boldsymbol{w}) = p_0^{y_0}(1-p_0)^{1-y_0}$
1	$\boldsymbol{x}_1$	y_1	p_1	$P(y_1 \mid \boldsymbol{x}_1, \boldsymbol{w}) = p_1^{y_1}(1-p_1)^{1-y_1}$
⋮	⋮	⋮	⋮	⋮
$N-1$	$\boldsymbol{x}_{N-1}$	y_{N-1}	p_{N-1}	$P(y_{N-1} \mid \boldsymbol{x}_{N-1}, \boldsymbol{w}) = p_{N-1}^{y_{N-1}}(1-p_{N-1})^{1-y_{N-1}}$

  - 데이터 세트에 대한 조건부 확률
    - 개별 조건부 확률의 곱

개별 데이터에 대해 ***<u>서로 독립적</u>***
=> 확률 간 곱으로 표현 가능

$$\boxed{P(\boldsymbol{y}|X, \boldsymbol{w}) = \prod_{n=0}^{N-1} p_n^{y_n}(1-p_n)^{1-y_n}}$$

Likelihood(우도): 모든 입력값에 대해서 ***<u>출력 분포</u> y****가 나올 확률*

다수의 데이터에 대해 ***y****는 분포로 볼 수 있음*
=> 0 1 1 1 0 0 0 1 0 0 1 0 1 1 0 …

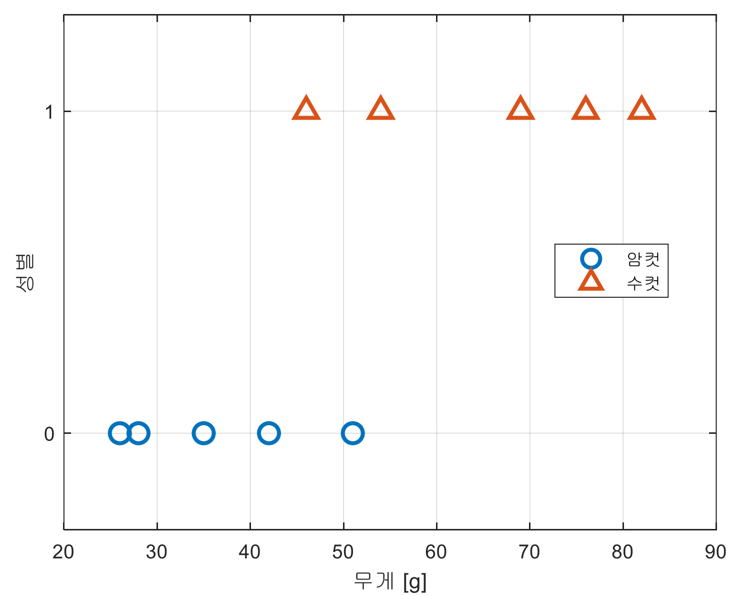
Likelihood 예시

지능형 예측·진단 연구실 Intelligent Prognostics & Diagnostics Lab.

- 사슴벌레 데이터
  - 최적이 아닌 **임의의 매개변수**에 대한 likelihood 계산
  - $(w_0, w_1) = (\ a\ ,\ \ b\ )$, likelihood = ?

$$p = P(y=1|x) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-w_0 x - w_1}}$$

번호 (n)	무게 (x)	성별 (y)	사후확률 (p)	상태확률 $p^y(1-p)^{1-y}$
0	26	0		
1	28	0		
2	35	0		
3	42	0		
4	51	0		
5	46	1		
6	54	1		
7	69	1		
8	76	1		
9	82	1		



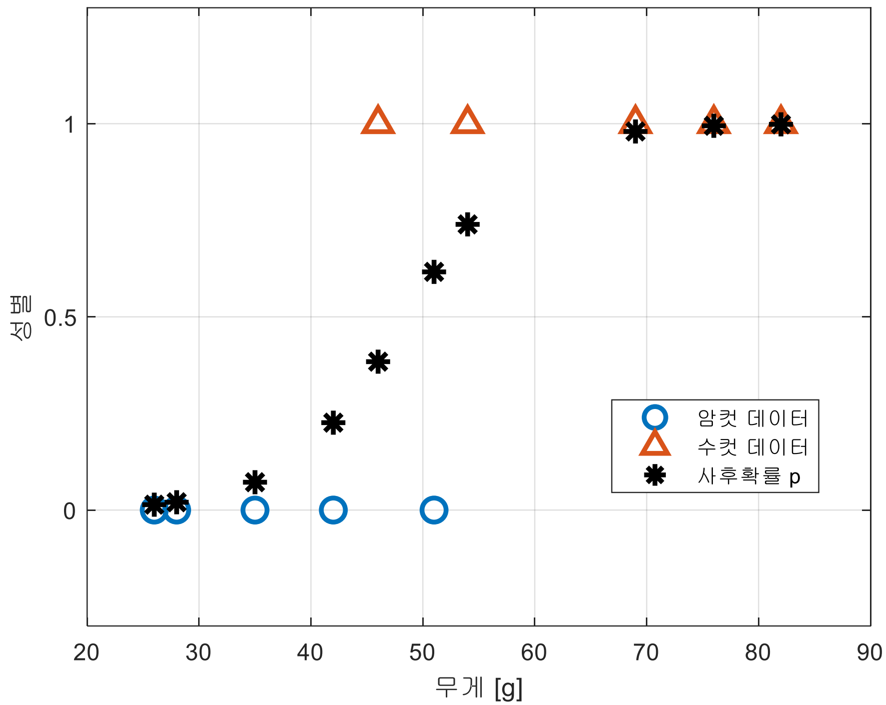
Likelihood 예시

○ 사슴벌레 데이터

- **최적 매개변수**에 대한 likelihood 계산
- $(w_0, w_1) = (0.1894, -9.1848)$, likelihood =0.0735

$$p = P(y = 1|x) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-w_0 x - w_1}}$$

번호 (n)	무게 (x)	성별 (y)	사후확률 (p)	상태확률 $p^y(1-p)^{1-y}$
0	26	0	0.0139	0.9861
1	28	0	0.0202	0.9798
2	35	0	0.0720	0.9280
3	42	0	0.2262	0.7738
4	51	0	0.6165	0.3835
5	46	1	0.3840	0.3840
6	54	1	0.7394	0.7394
7	69	1	0.9798	0.9798
8	76	1	0.9946	0.9946
9	82	1	0.9982	0.9982



Logistic Regression에서의 Cost Function

지능형 예측·진단 연구실
Intelligent Prognostics & Diagnostics Lab.

- 목표: $P(\boldsymbol{y}|X,\boldsymbol{w}) = \prod_{n=0}^{N-1} p_n^{y_n}(1-p_n)^{1-y_n}$ 를 최대로 만드는 $\boldsymbol{w}$

$$\boldsymbol{w}^* = \arg\max \prod_{n=0}^{N-1} p_n^{y_n}(1-p_n)^{1-y_n}$$

- *머신러닝에서의 weight 학습 방향성: Cost Function 이* ***최소****가 되는 방향*

$$\boldsymbol{w}^* = \arg min - \prod_{n=0}^{N-1} p_n^{y_n}(1-p_n)^{1-y_n}$$

- *경사하강법의 원리: Cost Function 미분 값의 – 방향으로 update*

❖ **문제점: *"곱"으로 이루어진 식에 대한 미분의 복잡성***

⇒ *어떻게 하면 Linear Regression처럼 "**합**"으로 표현할 수 있을까?*

⇒ *Log를 취하자!! Log는 단조증가함수이므로 최소와 최대값을 찾는데 영향을 주지 않는다.*

$$\boldsymbol{w}^* = \arg\min - \sum_{n=0}^{N-1} \{y_n \ln p_n + (1-y_n)\ln(1-p_n)\}$$

- *개별 오차에 대한 합이니깐 평균을 취하는 것이 타당하다.*

Logistic Regression에서의 Cost Function

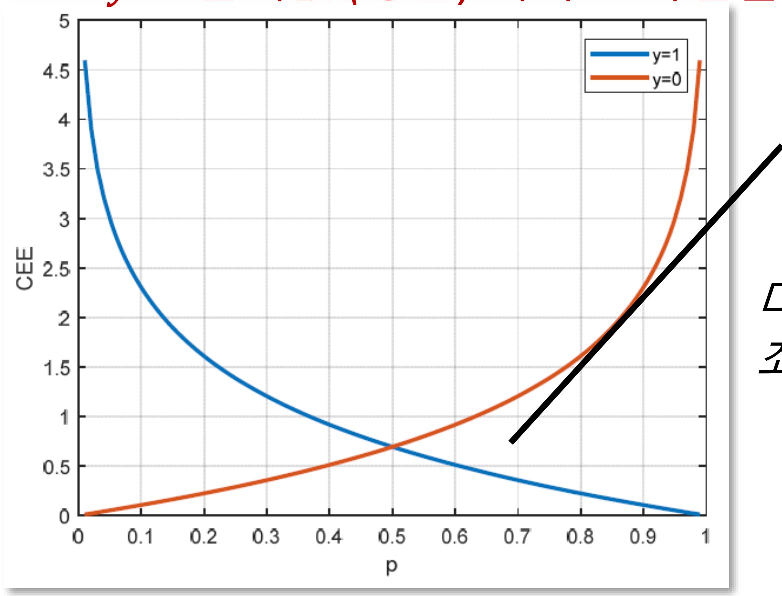
지능형 예측·진단 연구실
Intelligent Prognostics & Diagnostics Lab.

- Logistic Regression에서의 Cost Function
  - 교차 엔트로피 오차 (Cross Entropy Error)

$$\epsilon_{CEE} = -\frac{1}{N}\sum_{n=0}^{N-1}\{y_n \ln p_n + (1-y_n)\ln(1-p_n)\}$$

예측값 $\hat{y} = \begin{cases} 1 & (p > 0.5) \\ 0 & (p < 0.5) \end{cases}$

y는 실제값(정답) 이라는 사실을 잊지 말자



미분(기울기)의 – 방향이 최소값이 있는 방향이다!

*즉, CEE가 최소가 되는 **w** 를 찾으면 된다. => 경사하강법 적용*

[참고] 정보이론

- 메시지가 지닌 정보를 수량화할 수 있나?
  - “고비 사막에 눈이 왔다”와 “대관령에 눈이 왔다”라는 두 메시지 중 어느 것이 더 많은 정보를 가지나?
  - 정보이론의 기본 원리 → 확률이 작을수록 많은 정보
- 자기 정보self information
  - 사건(메시지) e_i의 정보량 (단위: 비트 또는 나츠)

$$h(e_i) = -\log_2 P(e_i) \quad \text{또는} \quad h(e_i) = -\log_e P(e_i)$$

- 엔트로피entropy
  - 확률변수 x의 불확실성을 나타내는 엔트로피

이산 확률분포 $H(x) = -\sum_{i=1,k} P(e_i)\log_2 P(e_i)$ 또는 $H(x) = -\sum_{i=1,k} P(e_i)\log_e P(e_i)$

연속 확률분포 $H(x) = -\int_{\mathbb{R}} P(x)\log_2 P(x)$ 또는 $H(x) = -\int_{\mathbb{R}} P(x)\log_e P(x)$

[참고] 정보이론

지능형 예측·진단 연구실
Intelligent Prognostics & Diagnostics Lab.

■ 자기 정보와 엔트로피 예제

윷을 나타내는 확률변수를 x라 할 때 x의 엔트로피는 다음과 같다.

$$H(x) = -\left(\frac{4}{16}\log_2\frac{4}{16} + \frac{6}{16}\log_2\frac{6}{16} + \frac{4}{16}\log_2\frac{4}{16} + \frac{1}{16}\log_2\frac{1}{16} + \frac{1}{16}\log_2\frac{1}{16}\right) = 2.0306\text{비트}$$

주사위는 눈이 6개인데 모두 1/6이라는 균일한 확률을 가진다. 이 경우 엔트로피를 계산하면 다음과 같다.

$$H(x) = -\left(\frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) = 2.585\text{ 비트}$$

주사위가 윷보다 엔트로피가 높은 이유는?

■ 교차 엔트로피cross entropy

- 두 확률분포 P와 Q 사이의 교차 엔트로피

$$H(P,Q) = -\sum_{x} P(x)\log_2 Q(x) = -\sum_{i=1,k} P(e_i)\log_2 Q(e_i)$$

Gradient Descent Method

- 경사하강법
  - Cost Function의 미분에 마이너스 값을 취해 Update

$$\boxed{\boldsymbol{w}_{new} = \boldsymbol{w}_{old} - \alpha \frac{\partial}{\partial \boldsymbol{w}_{old}} \epsilon_{CEE}(\boldsymbol{w}_{old})}$$

CEE에 마이너스 부호가 있다는 것을 잊지 말자

$$\epsilon_{CEE} = -\frac{1}{N}\sum_{n=0}^{N-1}\{y_n \ln p_n + (1-y_n)\ln(1-p_n)\}$$

Sigmoid Function의 미분은 매우 깔끔하게 정리된다.

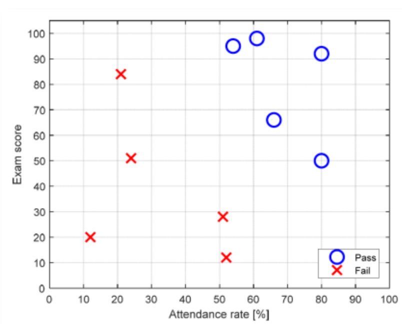
$$\frac{\partial}{\partial \boldsymbol{w}}\epsilon_{CEE}(\boldsymbol{w}) = \frac{1}{N}\sum_{n=0}^{N-1}(p_n - y_n)\boldsymbol{x}_n$$

이때 , $\boldsymbol{x}_n = [\, \boldsymbol{x}_{0,n} \;\; \boldsymbol{x}_{1,n} \;\; \boldsymbol{x}_{2,n} \;\; \dots \;\; \boldsymbol{x}_{M-1,n} \;\; \mathbf{1} \,]^{\mathrm{T}}$

- $\boldsymbol{M}$=2 일 때, (Input 속성이 2개)

$$p_n = \frac{1}{1+e^{-z_n}} = \frac{1}{1+e^{-(w_0 x_{0,n} + w_1 x_{1,n} + w_2)}}$$

데이터 번호	입력		출력 y (이수 여부, 1=pass, 0=fail)
	x_0 (출석률)	x_1 (시험성적)	
0	21	84	0
1	54	95	1
2	80	50	1
3	51	28	0
4	66	66	1
5	80	92	1
6	24	51	0
7	61	98	1
8	12	20	0
9	52	12	0



$$\frac{\partial}{\partial w_0}\epsilon_{CEE}(w_0, w_1, w_2) = \frac{1}{N}\sum_{n=0}^{N-1}(p_n - y_n)x_{0,n}$$

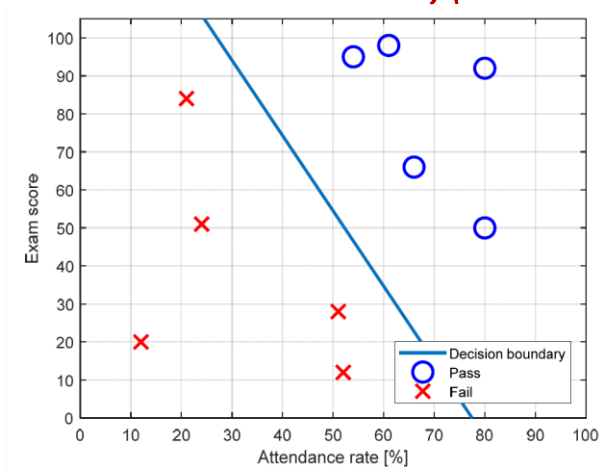
$$\frac{\partial}{\partial w_1}\epsilon_{CEE}(w_0, w_1, w_2) = \frac{1}{N}\sum_{n=0}^{N-1}(p_n - y_n)x_{1,n}$$

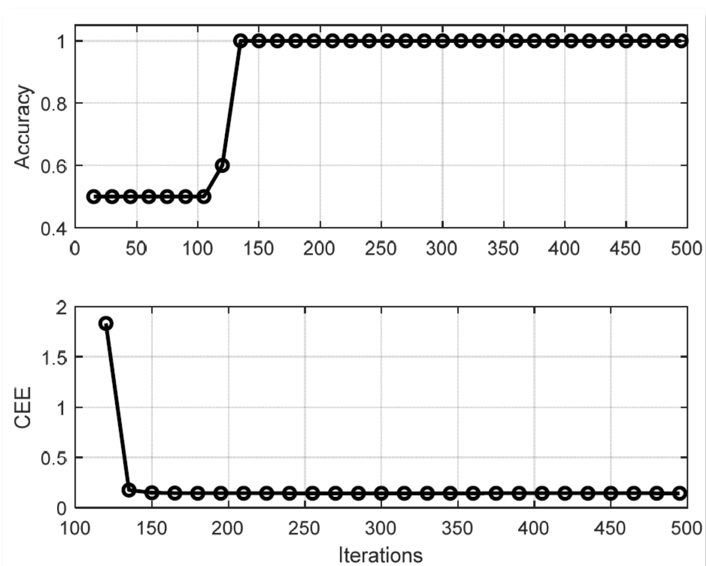
$$\frac{\partial}{\partial w_2}\epsilon_{CEE}(w_0, w_1, w_2) = \frac{1}{N}\sum_{n=0}^{N-1}(p_n - y_n)$$

Logistic Regression의 결과

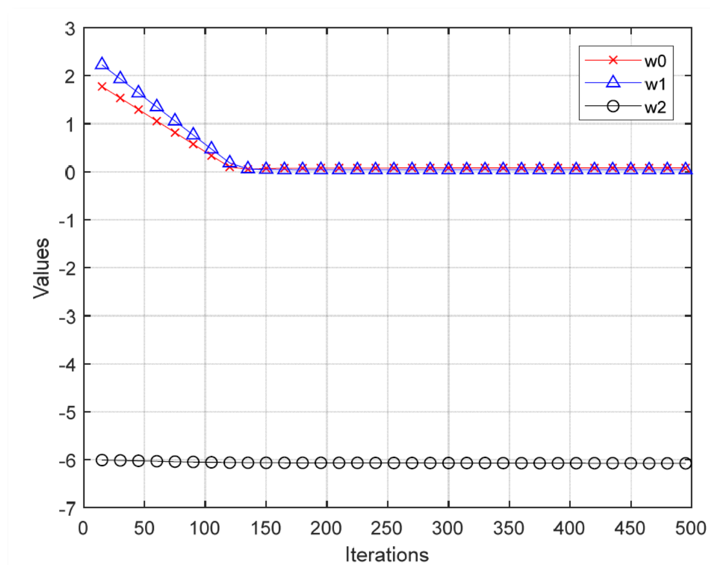
데이터 번호	입력		출력 y (이수 여부, 1=pass, 0=fail)
	x_0 (출석률)	x_1 (시험성적)	
0	21	84	0
1	54	95	1
2	80	50	1
3	51	28	0
4	66	66	1
5	80	92	1
6	24	51	0
7	61	98	1
8	12	20	0
9	52	12	0

Decision Boundary(결정경계)





경사하강법 적용

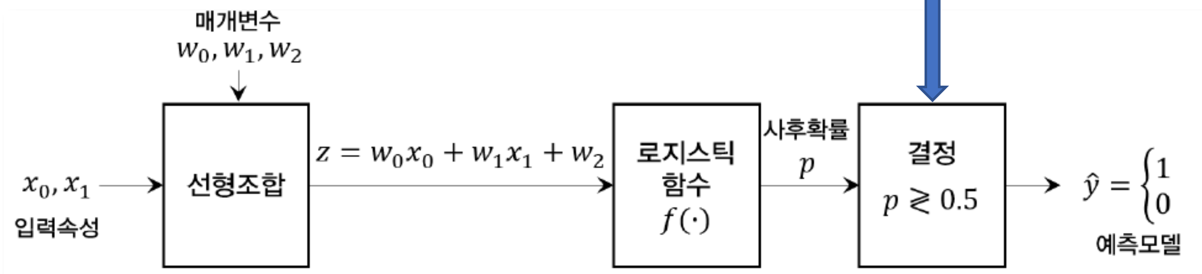


Weight의 변화

Decision Boundary

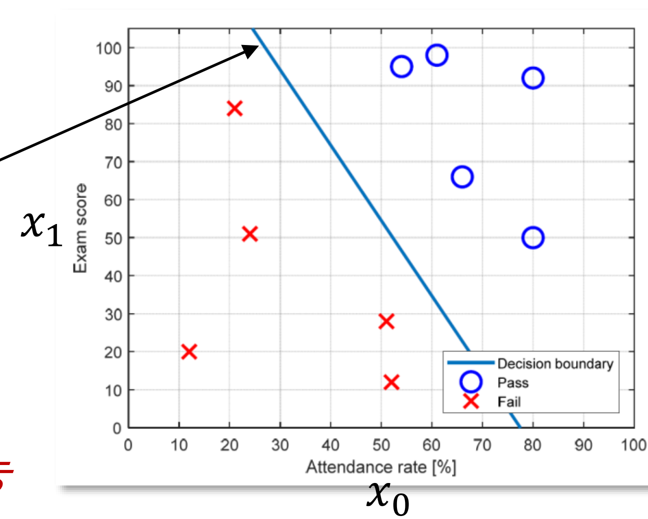
지능형 예측·진단 연구실
Intelligent Prognostics & Diagnostics Lab.

Decision Boundary(결정경계)



- binary classification에서의 Decision Boundary는 0.5
  - Sigmoid Function (Logistic Function)을 통과한 값이 0.5
- Sigmoid Function의 입력이 0이면 => 출력 0.5
  - $z = w_0x_0 + w_1x_1 + w_2 = 0$

$$x_1 = -\frac{w_0}{w_1}x_0 - \frac{w_2}{w_1}$$



[한계점]

1) 직선(평면) 형태로만 공간을 분할 한다.

2) Binary Classification (0 or 1) 에만 적용가능

Chap.3 실습

지능형 예측·진단 연구실
Intelligent Prognostics & Diagnostics Lab.

각 실습문제에 대해 code(+주석), 그래프, 분석 내용 등은 필수적으로 포함시킬 것

1) Chap.1 에서 구현한 linear regression에 대한 경사하강법 사용자 지정함수를 응용 및 수정하여, logistic regression을 위한 경사하강법 사용자 지정함수를 구현하라.

2) Chap.2에서 구현한 데이터 set 분할 함수를 활용하여, "logistic_regression_data.csv" 데이터 set을 Training : Test = 7:3으로 분할하라.

3) 1)에서 구현한 경사하강법 함수를 이용해 Training set을 활용하여 logistic regression 모델 학습을 진행하고, 학습 진행(epoch)에 따른 $\boldsymbol{w}$ / CEE / training set에 대한 분류 정확도를 각각 그래프(본 강의자료 23page 하단 그림 참조)로 나타내라

   실습은 정확도 그래프만 검사

4) 학습이 완료된 모델을 활용하여, Test set을 분류하고, 분류 정확도를 나타내라. (만약, 분류 정확도가 낮을 경우, validation set 활용 및 Hyper-parameter 변경 등을 통한 재학습을 수행)

5) 학습이 끝난 모델을 활용하여, Training set 및 Test set에 대한 Decision Boundary (본 강의자료 23page 우측 상단 그림 참조) 그래프를 각각 그려라.