Artificial Intelligence (AI)

Lec06: Logistic Regression Part 1

충북대학교

문성태 (지능로봇공학과)

stmoon@cbnu.ac.kr

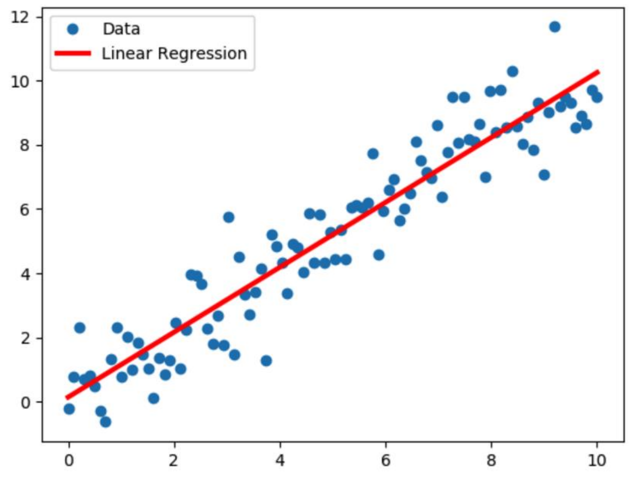
01

Logistic Regression for binary classification

Recap: Linear Regression

❖ Objective Function

$$f_{w,b}(X) = WX + b$$



Let’s apply linear regression

- Example

Application	Observation	0	1
Medical Diagnosis	Patient	Healthy	Diseased
Email Analysis	Email	Not Spam	Spam
Financial Data Analysis	Transaction	Not Fraud	Fraud
Marketing	Website visitor	Won't Buy	Will Buy
Image Classification	Image	Hotdog	Not Hotdog

Binary Classification

❖ Let’s try to use linear regression!

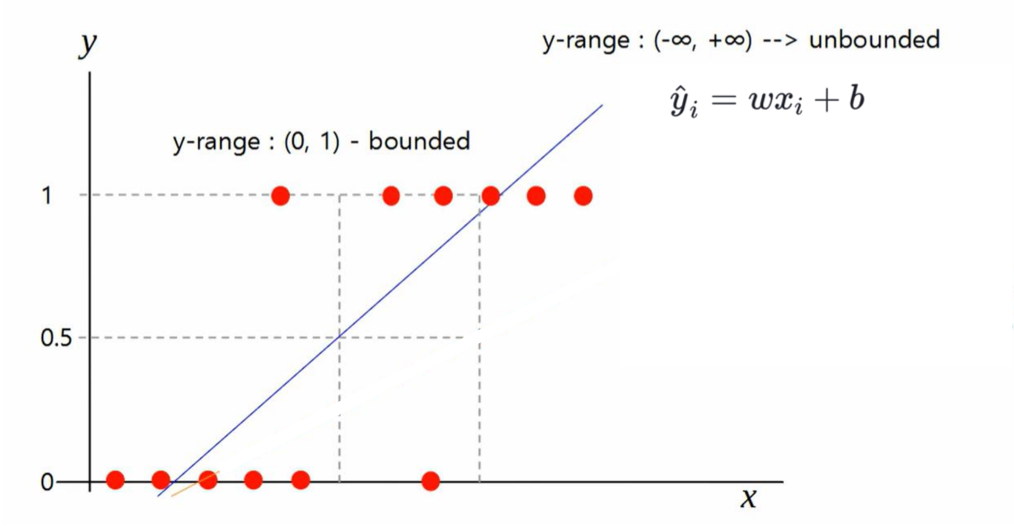
Training data set

no	x	y
1	1.2	0
2	10.0	1
3	3.5	0
4	8.2	1
5	2.0	0
6	9.4	1
7	1.5	0
8	7.8	1
...	...	...
N	2.5	0

Test data set

no	x	y
1	1.8	?
2	6.2	?

y_pred
0.10
0.68



Linear Regression

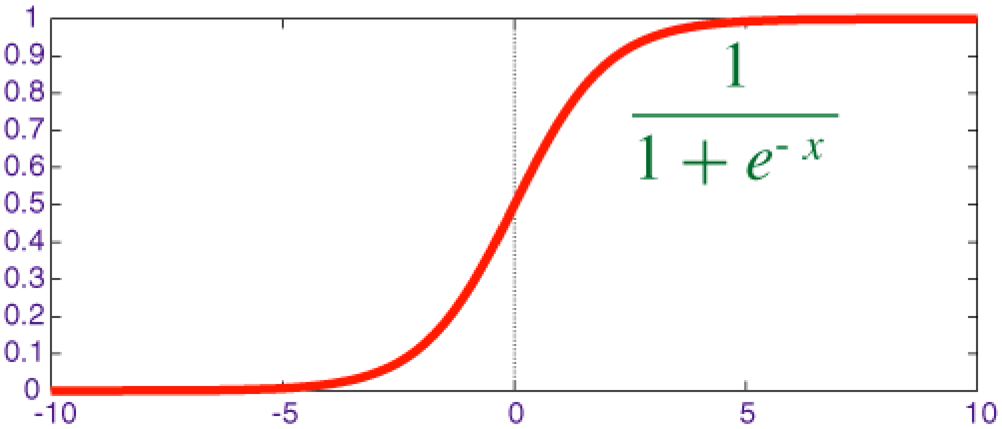
- Let's use linear regression!
- Problem
  - Y : 0 ~ 1 (bounded) but prediction: $-\infty \sim +\infty$ (unbounded)
    - Not match!
  - wrong regression because of some data

Logistic Regression

- Transform y
  - $0 \sim 1$ (bounded) ➔ $-\infty \sim +\infty$ (unbounded)
  - Use log odds (logit) in linear regression ➔ logistic regression

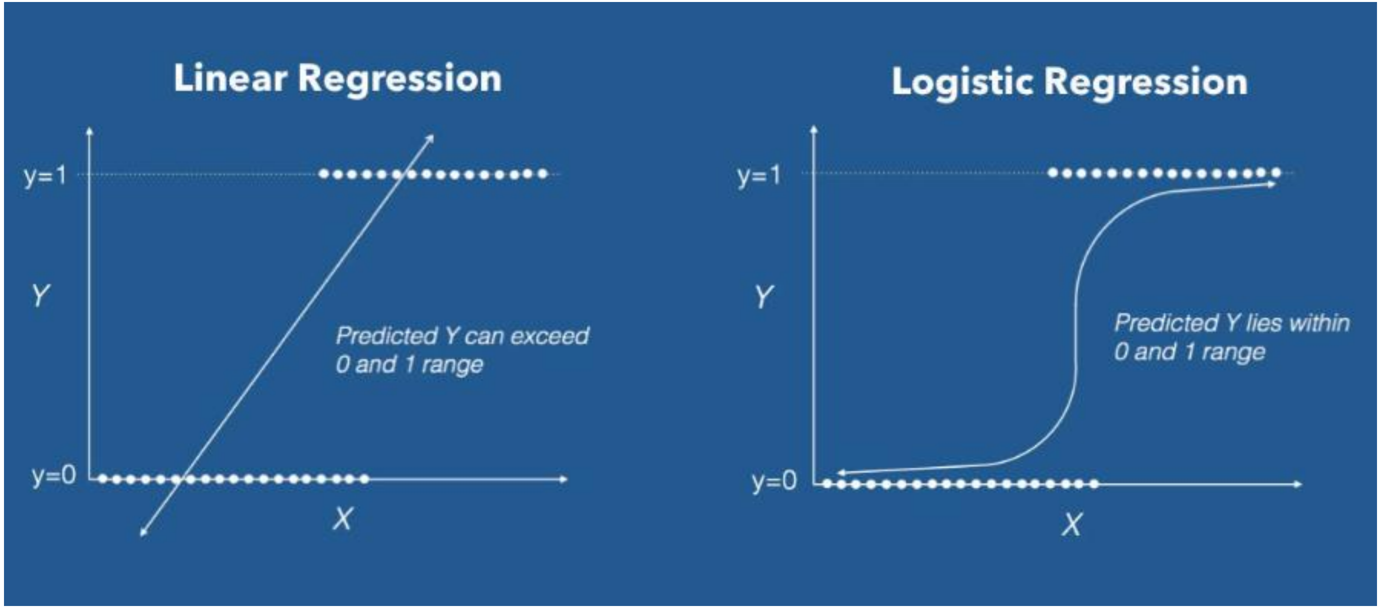
Sigmoid or Logistic Function

- Curved in two directions like the letter “S” or the Greek ς (sigma)



Linear Regression vs Logistic Regression

- ❖ Linear regression: 확률 예측 불가 (y: -inf ~ inf)
- ❖ Logistic regression: 확률 예측 (y: 0 ~ 1)

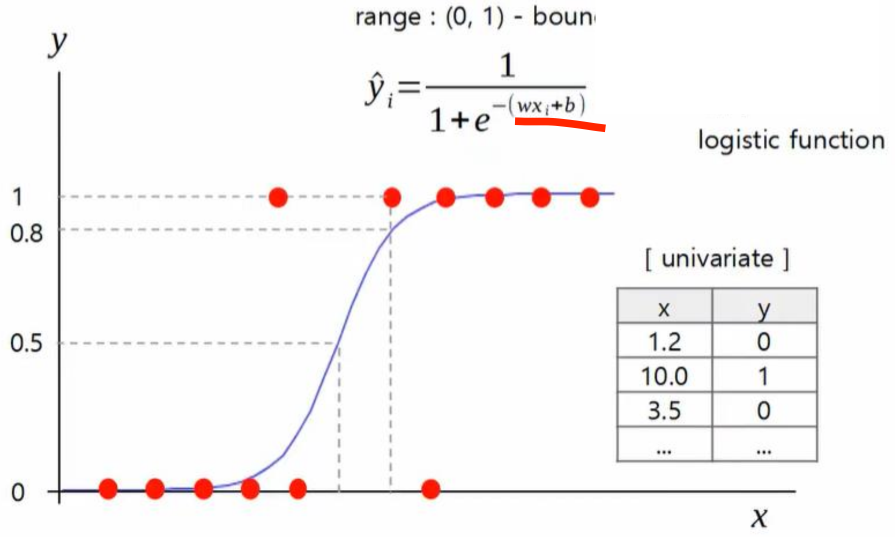


Sigmoid or Logistic Function

❖ 이진 분류 (binary classification) 데이터를 학습하고 추정하려면, sigmoid (logistic function)을 추정식으로 설정

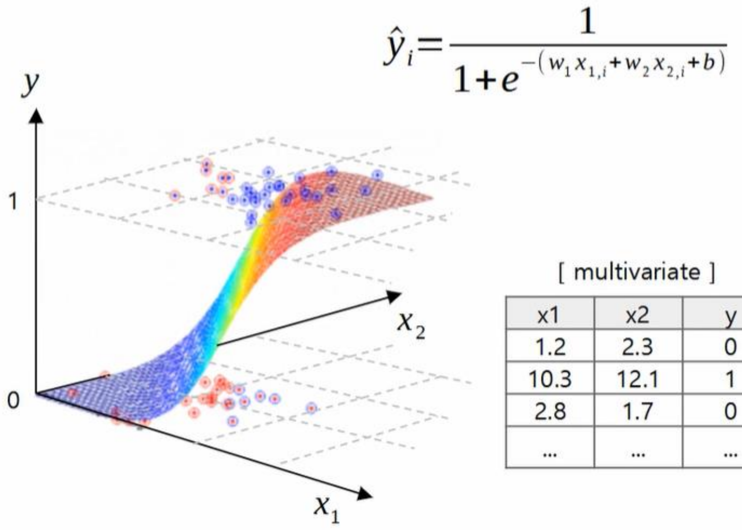
- 이진 분류: 0 / 1 □ 확률로 표현

❖ 단일변수 (univariate)나 다중 변수 (multivariate)를 갖는 데이터에 동일하게 적용 가능



[univariate]

x	y
1.2	0
10.0	1
3.5	0
...	...

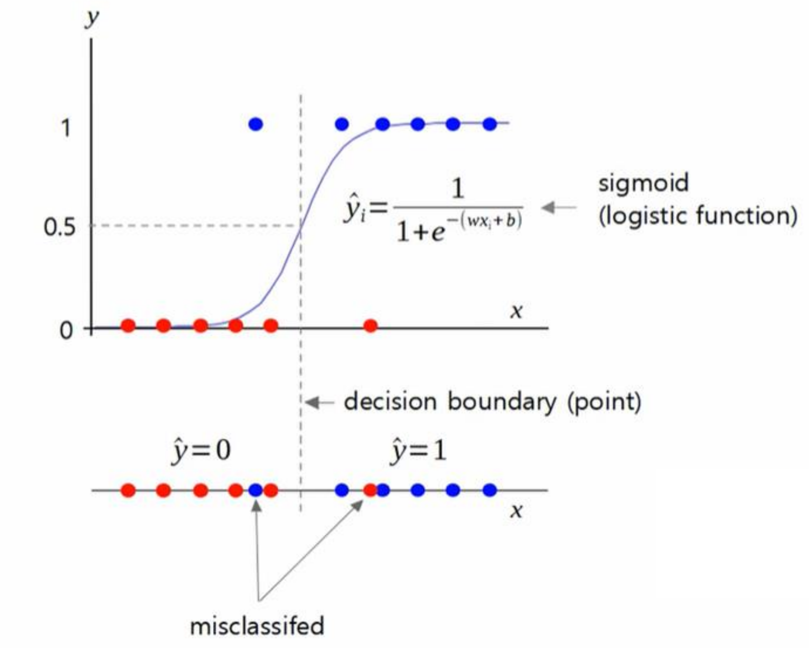


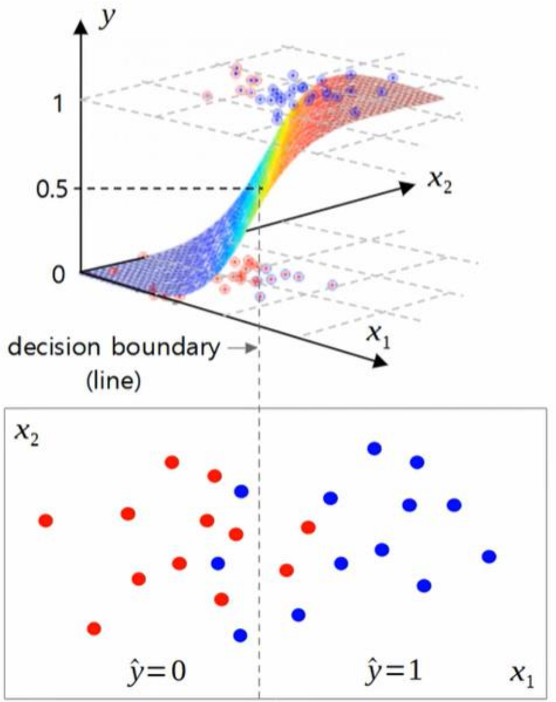
[multivariate]

x1	x2	y
1.2	2.3	0
10.3	12.1	1
2.8	1.7	0
...	...	...

Decision Boundary

❖ Decision boundary is 0.5 in general (not always)





Objective Function

- (BCE) Binary cross entropy by MLE (maximum likelihood estimation)

Cost Function

- Binary Cross Entropy (BCE)

$$J(H(x), y) = -ylog\left(H(x)\right) - (1-y)\log(1-H(x))$$

02

Entropy

정보량

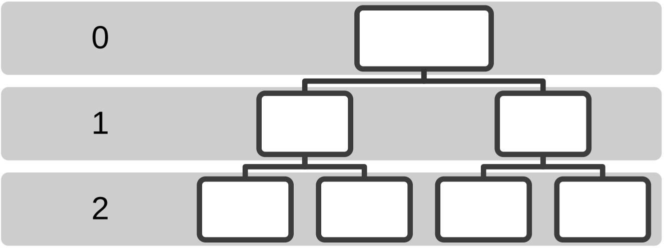
- ❖ 임의의 알파벳(26)을 맞추기 위해 필요한 질문의 수는?

- ❖ 6개 알파벳을 맞추기 위해 필요한 질문의 수는?

  - □ 6(n)개 알파벳(s=26)을 맞추기 위해서는 28.2개의 질문 (정보, H)가 필요하다

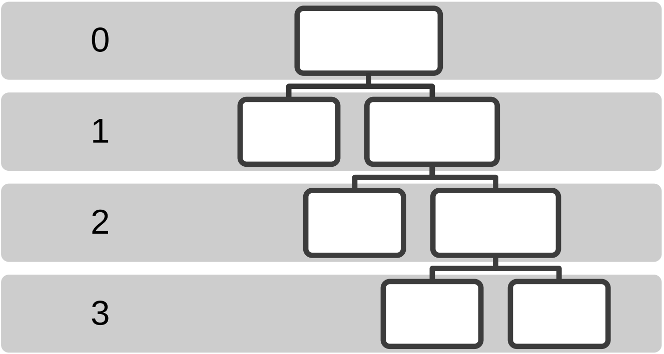
Entropy

- Machine X (A, B, C, D를 각각 0.25의 확률로 출력)
  - P(A): 0.25, P(B): 0.25, P(C): 0.25, P(D): 0.25



Machine A

- Machine Y
  - P(A): 0.5, P(B): 0.125, P(C): 0.125, P(D): 0.25



Machine B

- A : 0.5 = 1/2
- B: 0.125 = 1/8
- C: 0.125 = 1/8
- D: 0.25 =1/4

$$\log_2(\frac{1}{1/2}) = 1$$

$$\log_2(\frac{1}{1/8}) = 3$$

$$\log_2(\frac{1}{1/4}) = 2$$

Entropy

- 정보를 정량화하는 단위
  - 정보가 클수록 엔트로피가 크고, 정보가 작으면 엔트로피가 작다.

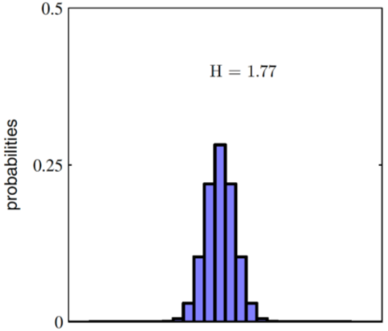
$$H = \sum (\text{사건 발생확률}) \cdot \boxed{\log_2\left(\frac{1}{\text{사건 발생확률}}\right)}$$

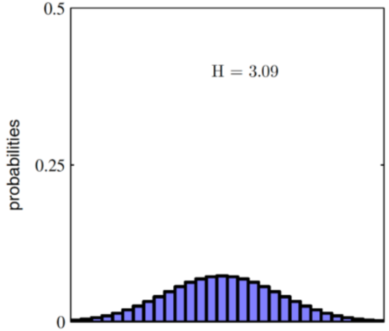
$$= \sum_i p_i \log_2\left(\frac{1}{p_i}\right)$$

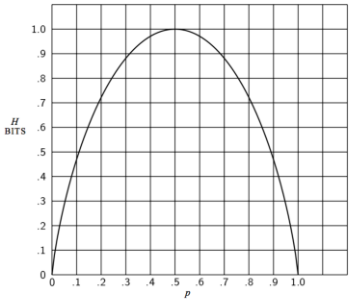
$$= -\sum_i p_i \log_2(p_i)$$

information

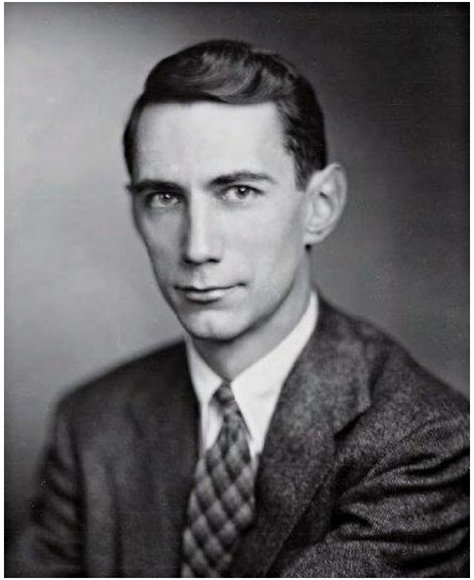
*) unit of H: bit







Entropy function



Claude Shannon
(1961~2001)

A Mathematical Theory of Communication

By C. E. SHANNON

Reprinted with corrections from *The Bell System Technical Journal,*
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

Cross-Entropy

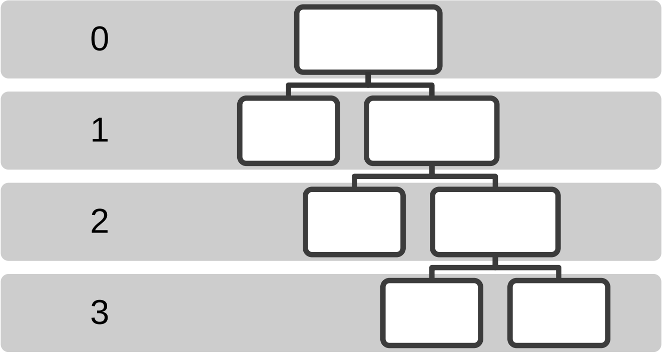
- H(Machine B) = 1.75
- H(Strategy for Machine B) = 2 □ Cross Entropy
  - Cross Entropy: 어떤 문제에 대해 특정 전략을 사용할 때 예상되는 질문개수에 대한 기댓값

    □ 확률분포로 된 어떤 문제 p에 대해 확률분포로 된 어떤 전략 q를 사용할 때 예상되는 질문개수에 대한 기댓값

$$H(p,q) = \sum_i p_i \ \log_2 \frac{1}{q_i} = -\sum_i p_i \ \log_2 q_i$$

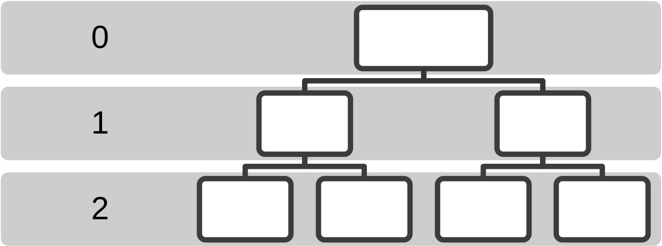
$p(A) \times 1 + p(B) \times 3 + p(C) \times 3 + p(D) \times 2 = 1.75$

$p(A) \times 2 + p(B) \times 2 + p(C) \times 2 + p(D) \times 2 = 2$



Machine B

- A : 0.5 = 1/2
- B: 0.125 = 1/8
- C: 0.125 = 1/8
- D: 0.25 =1/4



Strategy S

Cross-Entropy

- The cross-entropy of the distribution *p* relative to a distribution *q* over a given set

$$H(p,q) = \sum_i p_i \log_2 \frac{1}{q_i}$$
$$= -\sum_i p_i \log_2 q_i$$

- Example
  - Probability of ball in the bag (*p*): red(80%), green(10%), blue(10%)
  - Prediction of ball (*q*): red(20%), green(20%), blue(60%)

Binary Cross Entropy vs Cross Entropy

- If $i \in \{0, 1\}$,
  - $p(i = 0) = y,\ p(i = 1) = 1 - y$ ➔ $p = [y, 1 - y]$
  - $q(i = 0) = \hat{y},\ q(i = 1) = 1 - \hat{y}$ ➔ $q = [\hat{y}, 1 - \hat{y}]$

Kullback–Leibler(KL) Divergence

두 분포가 유사 할 수록, KL-divergence 값이 작고, 두 분포의 차이가 클수록 KL-divergence의 값은 크다

KL Divergence vs Cross Entropy

Cross entropy 두 분포 사이에 존재하는 정보량

$$\boxed{H(P,Q)} = KL(P||Q) + H(P)$$

$$J(H(x),y) = -y\log\left(H(x)\right) - (1-y)\log(1-H(x))$$

KL divergence 두 분포간의 정보 엔트로피 차이 :information gain

$$\boxed{KL(P||Q)} = H(P,Q) - H(P)$$