Artificial Intelligence

How to binary classification using linear regression?

☆ 흘러가는 흐름

Linear regression → Logistic regression

↓

multi class classification ⟸ Binary classification

Review: Linear Regression

↳데이터기반 → 특정값 예측

x1 (hours)	x2 (attendance)	y (score)
10	5	90
9	5	80
3	2	50
2	4	60
11	1	40

가설설정

- Hypothesis: $\underline{H(X) = WX}$

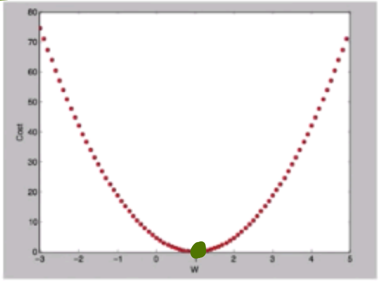
예측값

- Cost: $cost(W) = \frac{1}{m}\sum(WX - y)^2$

실제값

비용함수

최소화



(학습률)

Learning rate

☆

- Gradient decent: $W := W - \alpha \underline{\frac{\partial}{\partial W} cost(W)}$

경사하강

기울기

Today's Goal: Binary Classification 이진분류

- Spam Detection: Spam or Ham
- Facebook feed: show or hide
- Credit Card Fraudulent Transaction detection: legitimate/fraud

범주형

: 사기거래

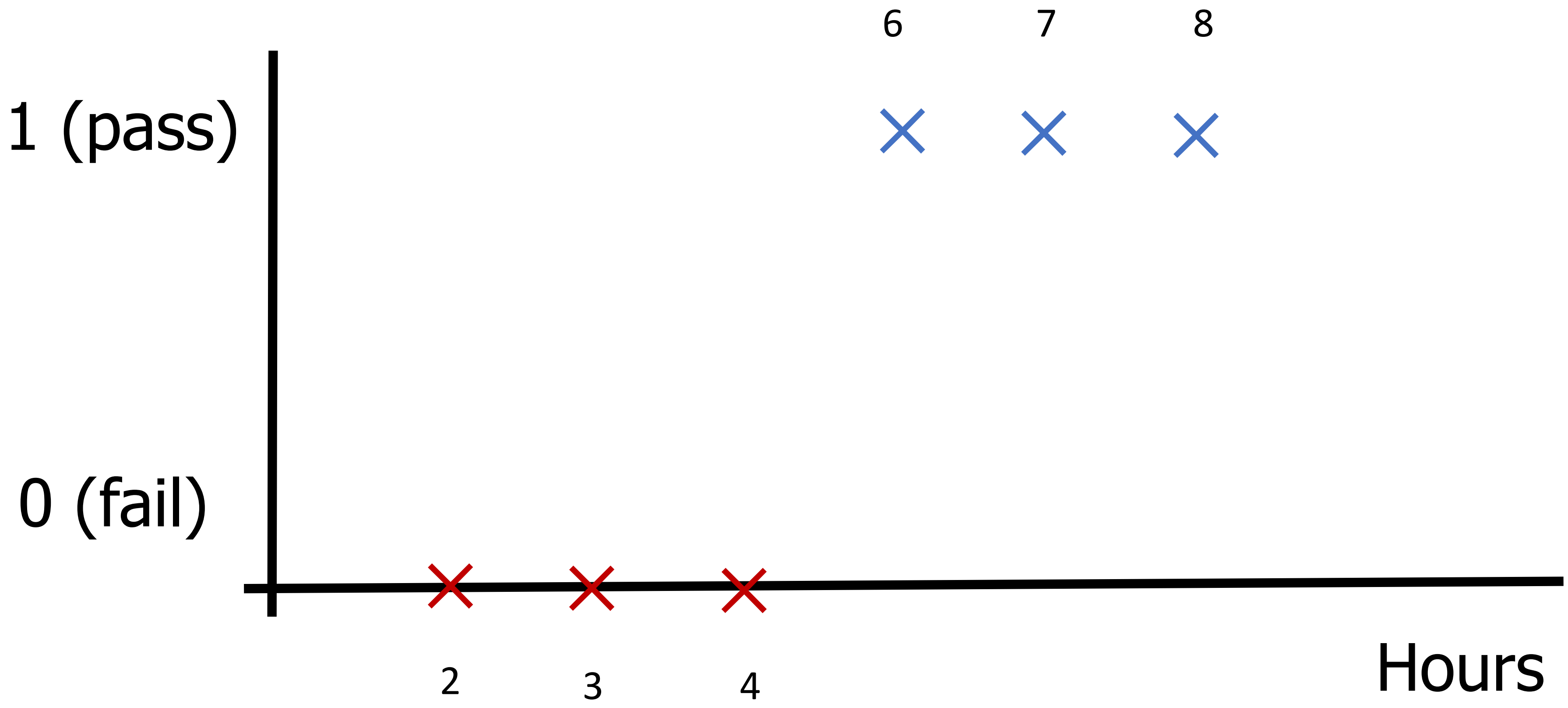
합법 불법(사기)

Binary Label Encoding ➔ "0" or "1" → "라벨링"

- Spam Detection: Spam(1) or Ham(0)
- Facebook feed: show(1) or hide(0)
- Credit Card Fraudulent Transaction detection: legitimate(1)/fraud(0)

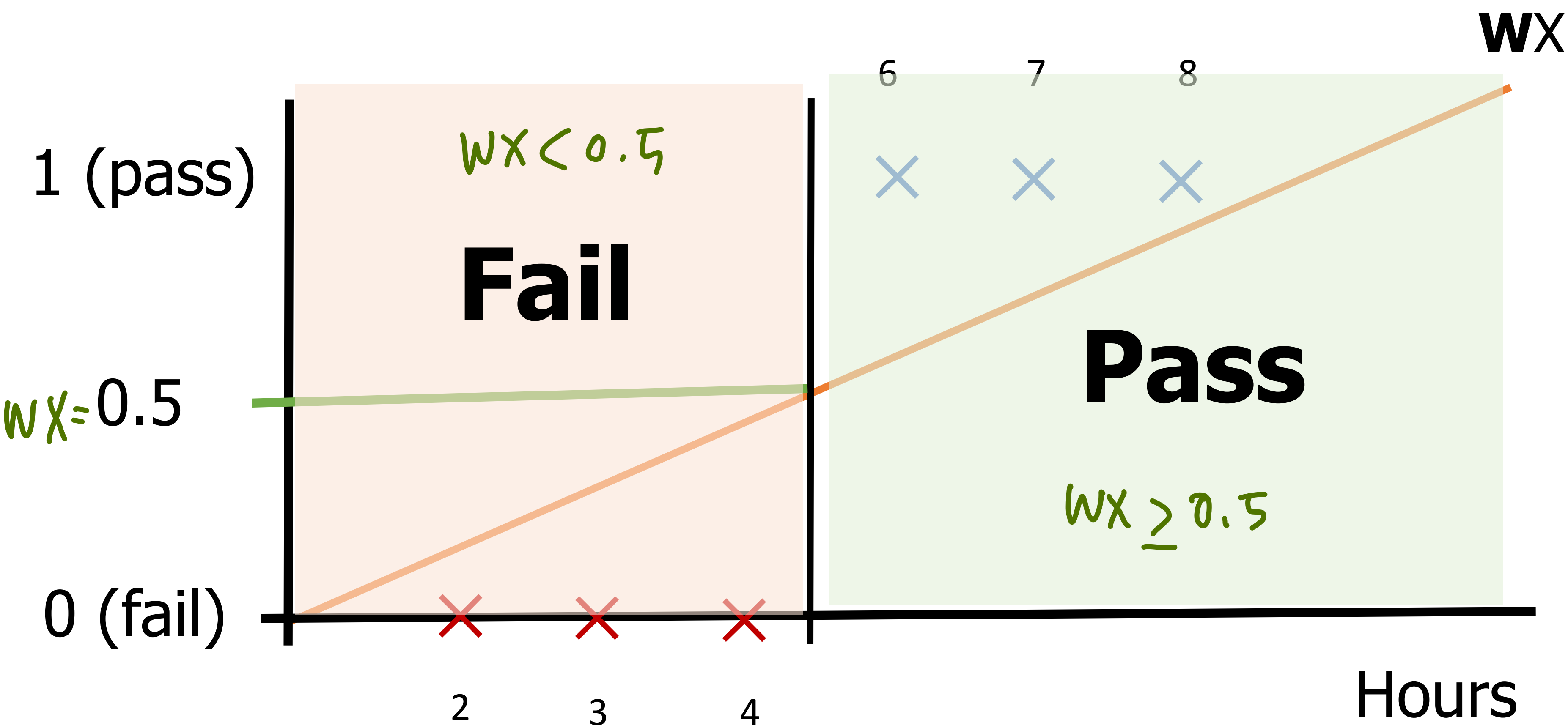
Pass or Fail based on study hours

- Pass(1) / Fail(0)



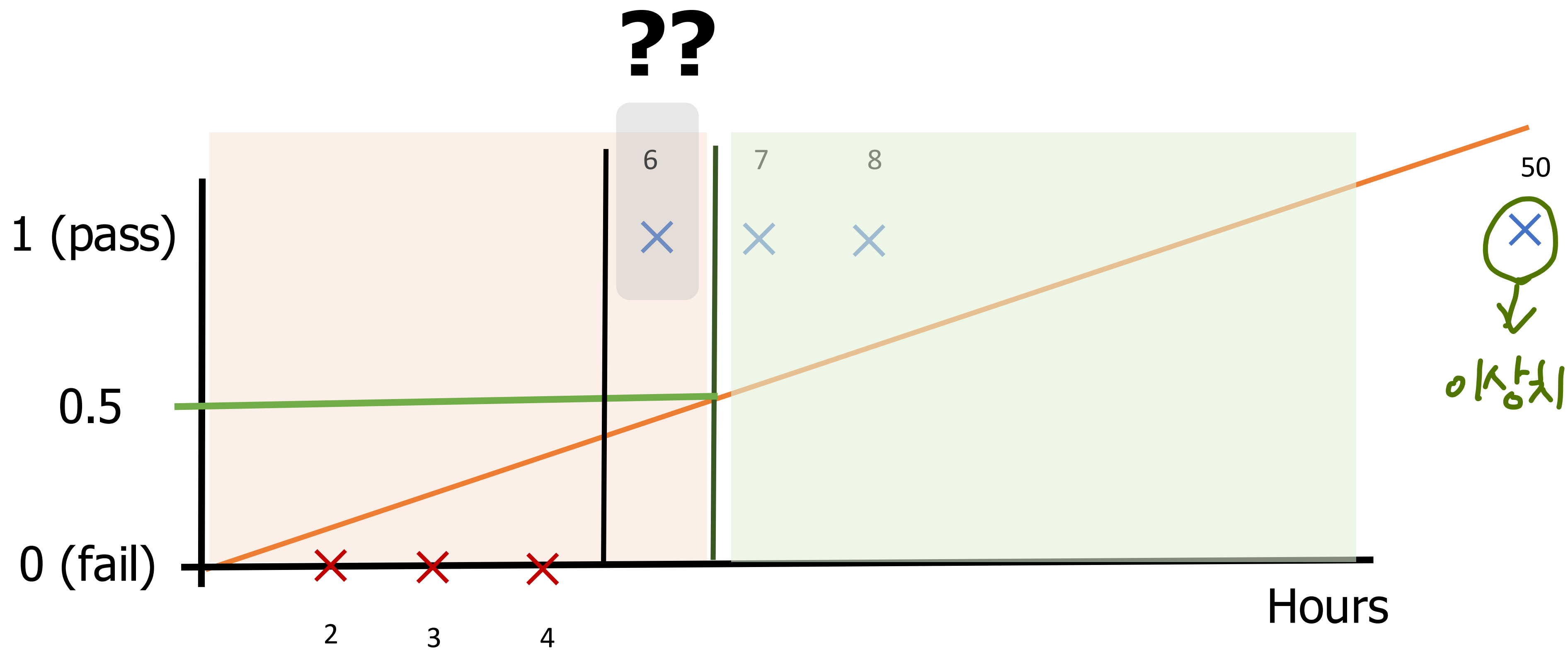
Binary Classification suing Linear Regression?

- Pass(1) / Fail(0)



Binary Classification suing Linear Regression?

- **Disadvantage #1**



Binary Classification suing Linear Regression?

- We know Y is 0 or 1

- **H**(x) = **W**x+**b**
- (+) This hypothesis is simple and easy to use
- (-) This hypothesis can give values large than 1 or less than 0

- **Example)**

  x = [1;2;5;10], W=0.5, b = 0 ➔ 0 <= y <= 1

  But!!

  if x = 100, y = 50 >> 1

Logistic Hypothesis

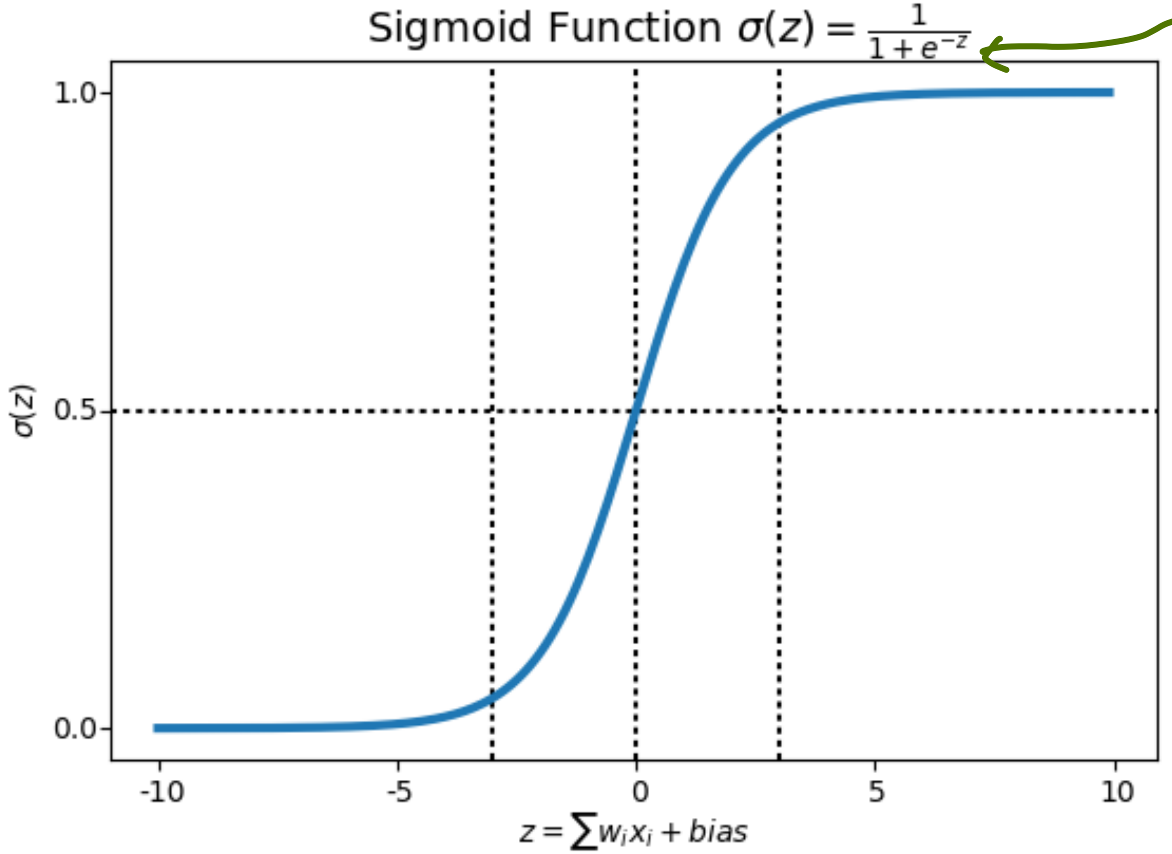
$$H(x) = \mathbf{W}x+\mathbf{b}$$

How???

$$0 <= H(x) <= 1$$

Logistic Hypothesis

- Sigmoid: Curved in two directions, like the letter “S”



Z = WX

$H(x) = \frac{1}{1+e^{-wx}}$

Logistic function := sigmoid function

Logistic Hypothesis

- Sigmoid 함수 덕분에 H(x)가 Bound 되었다.

$H(x) = WX$

↓ 변화

$$H(X) = \frac{1}{1 + e^{-(W^T X)}}$$

0<= H(x) <= 1

Cost function

- 이제 Cost function 에 적용해보자.

Cost function

- 기존 linear regression cost function에 적용하니, local minima에 빠진다.

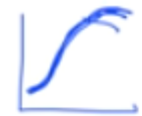
지역 극소점

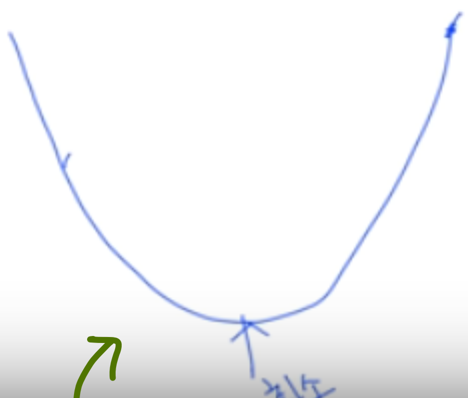
$$cost(W, b) = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$

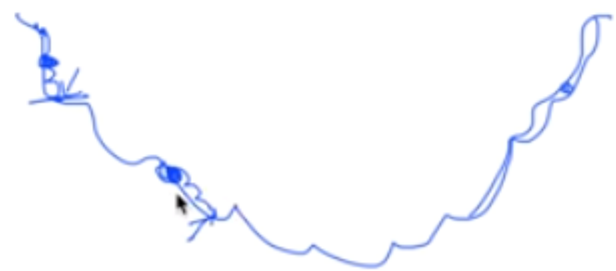
0< ~ <1

$$H(x) = Wx + b$$

$$H(X) = \frac{1}{1 + e^{-W^TX}}$$







Global minimum 극소점 찾을수 0

Cost function

가설설정 변경 → cost 함수도 재설계필요

- What is a cost function?
  - **우리의 "예측 값"(가설 값)이 얼마나 "정답"에 가까운가를 측정하는 척도!**

설계 팁:

정답에 가까워 질수록 Cost function 값 은 작고

정답에서 멀어질 수록 Cost function 값은 크게!

설계하면 된다.

Cost function for logistic regression

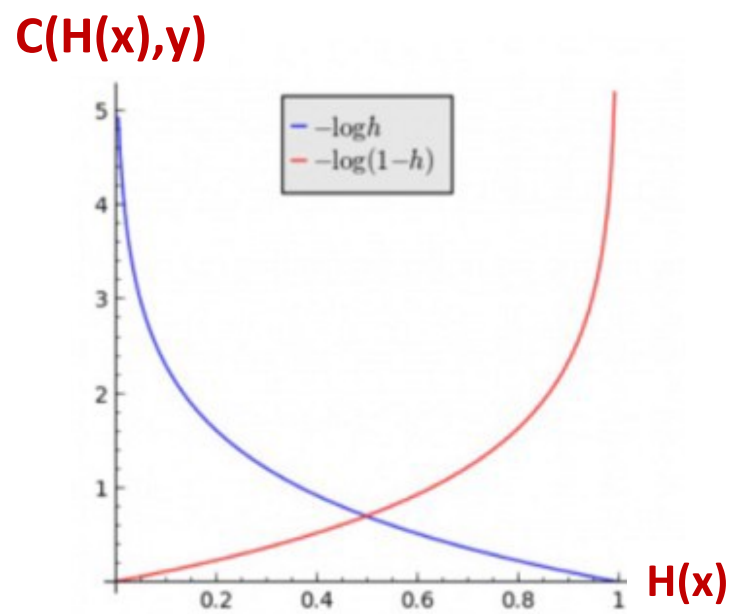
- Logistic regression
  - **H(x) ➔ [0, 1]**
  - if H(x) < 0.5, y = 0 else y = 1

  - 정답에 가까워 질수록 Cost function 값 은 작고, 정답에서 멀어질 수록 Cost function 값은 크게! 설계하면 된다.

$$H(X) = \frac{1}{1 + e^{-(W^T X)}}$$

$$cost(W) = \frac{1}{m} \sum c(H(x), y)$$

$$c(H(x), y) = \begin{cases} -log(H(x)) & : y = 1 \\ -log(1 - H(x)) & : y = 0 \end{cases}$$



Cost function for logistic regression

$$cost(W) = \frac{1}{m}\sum c(H(x), y)$$

$$c(H(x), y) = \begin{cases} -log(H(x)) & : y = 1 \\ -log(1 - H(x)) & : y = 0 \end{cases}$$

$$\Downarrow$$

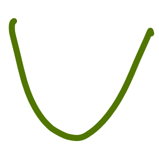
하나의 equation으로 잘 표현하면

$$\boldsymbol{C(H(x), y) = -ylog\big(H(x)\big) - (1 - y)\mathbf{log}(1 - H(x))}$$

Gradient decent algorithm

$$H(X) = \frac{1}{1 + e^{-W^T X}}$$

$$cost(W) = -\frac{1}{m} \sum y log\big(H(x)\big) + (1 - y)\log(1 - H(x))$$



형태

$$W := W - \alpha \frac{\partial}{\partial W} \mathrm{cost}(W)$$

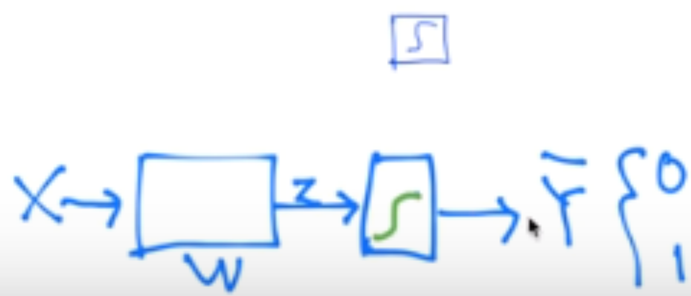
각자 해보자!!

(Summary) Logistic regression for binary classification

$$H_L(x) = Wx$$

$$z = H_L(x), \quad g(z)$$

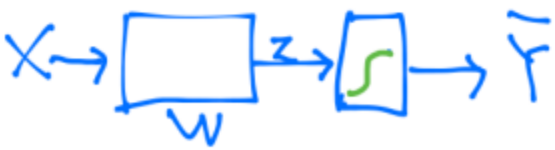
$$g(z) = \frac{1}{1+e^{-z}}$$

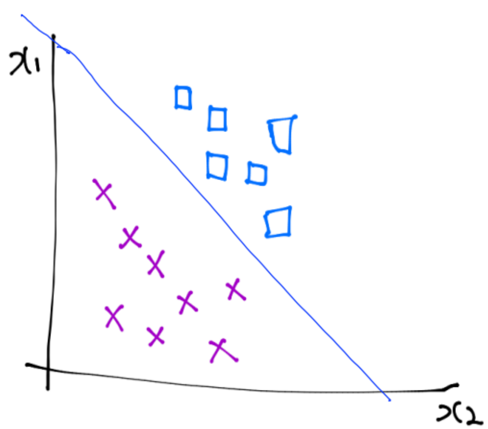


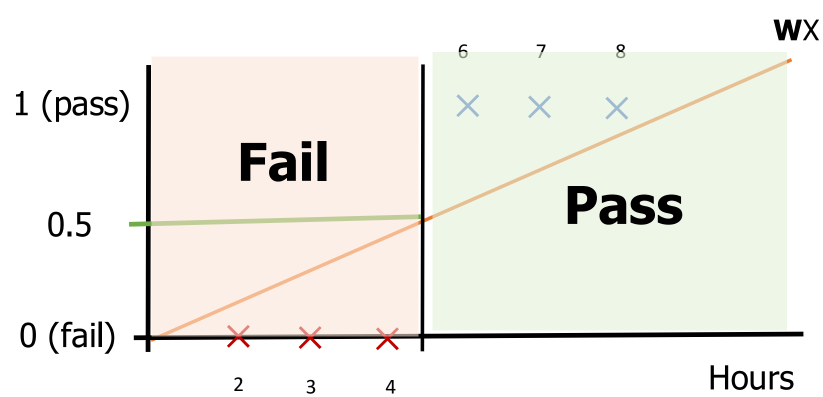
$$H_R(x) = g(H_L(x))$$

(Summary) Logistic regression for binary classification

$$g(z) = \frac{1}{1+e^{-z}} \qquad H_R(x) = g(H_L(x))$$







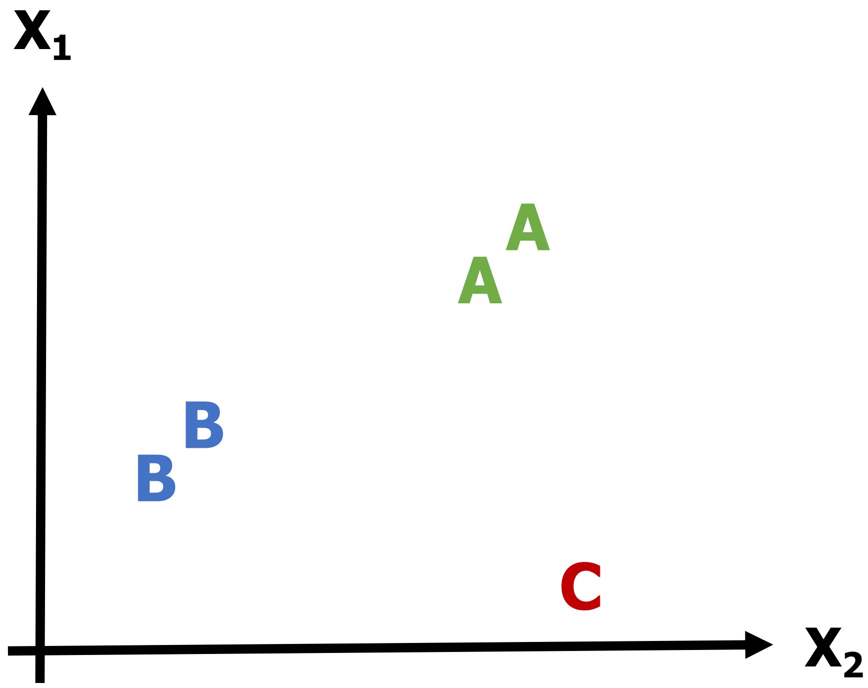
How to multi-class classification using binary classification?

Multinomial classification := Softmax classification

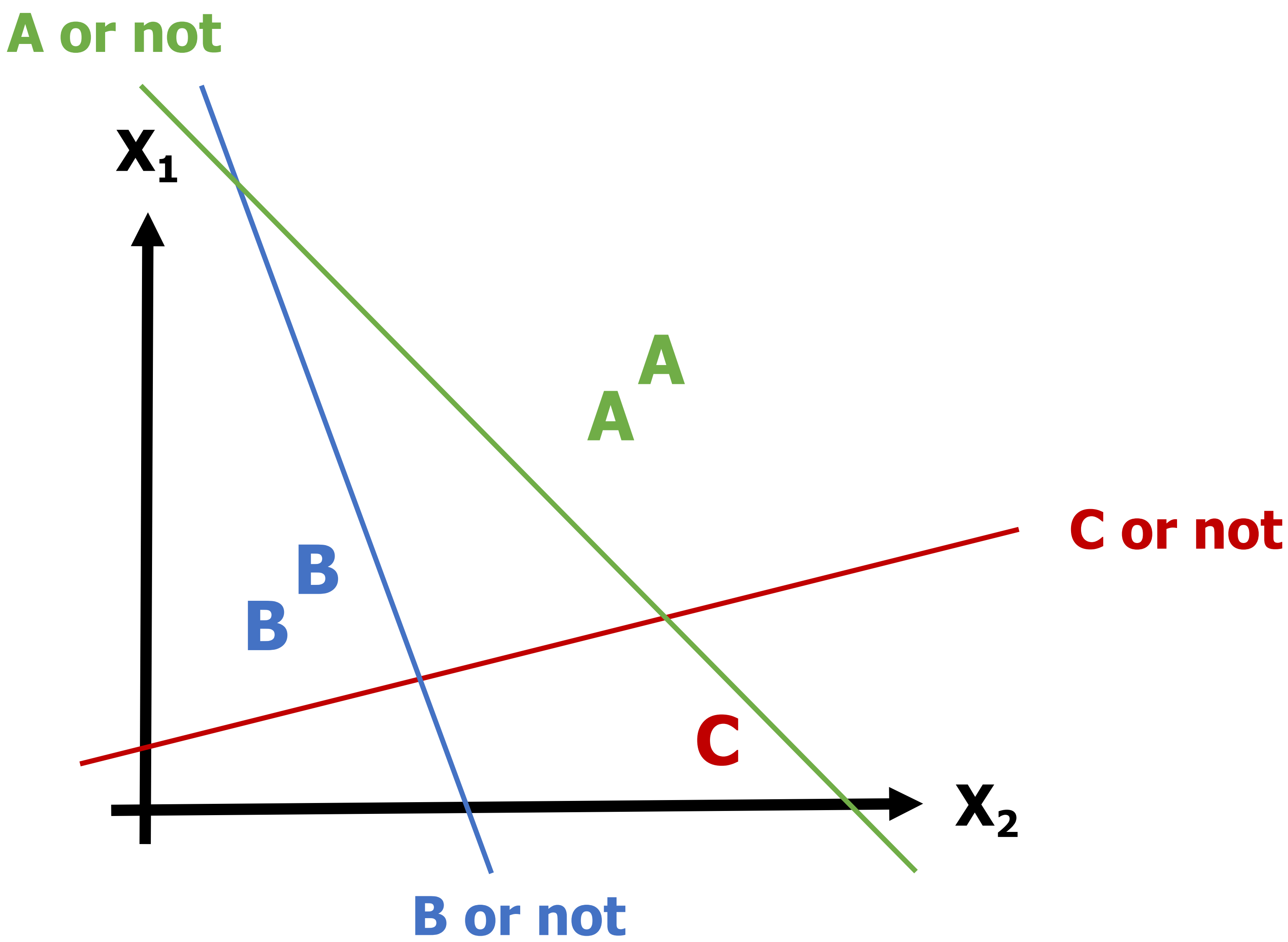
Multinomial classification (여러개의 클래스!)

x1 (hours)	x2 (attendance)	y (grade)
10	5	A
9	5	A
3	2	B
2	4	B
11	1	C

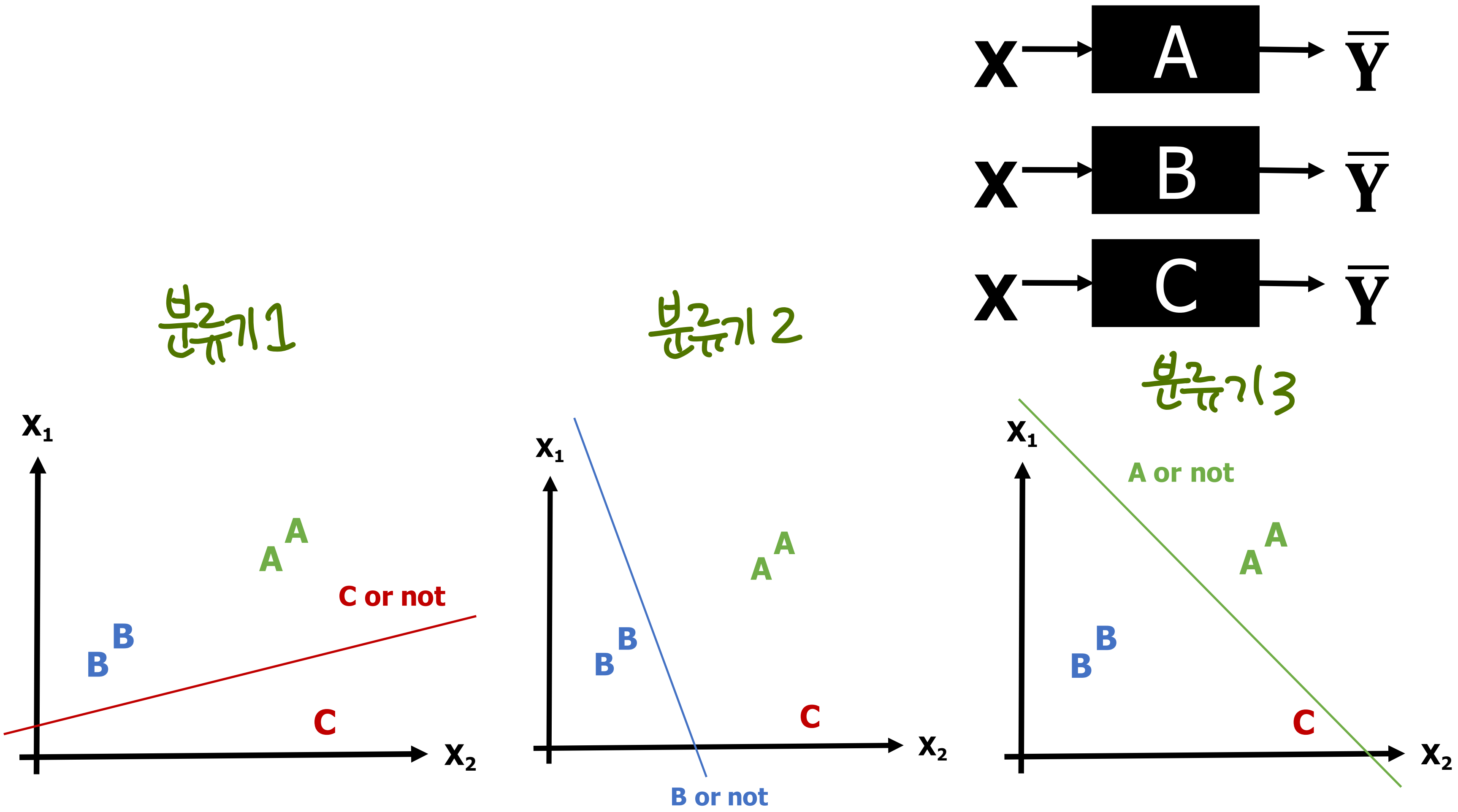
클래스 3개



Multinomial classification using binary classification



Multinomial classification using binary classification



Multinomial classification

Using Matrix Operation

A

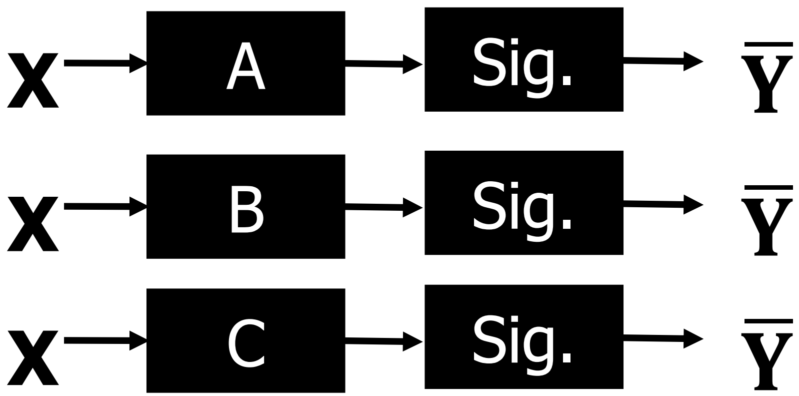
$$[w_1 \quad w_2 \quad w_3]\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} = [w_1x_1 + w_2x_2 + w_3x_3]$$

B

$$[w_1 \quad w_2 \quad w_3]\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} = [w_1x_1 + w_2x_2 + w_3x_3]$$

C

$$[w_1 \quad w_2 \quad w_3]\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} = [w_1x_1 + w_2x_2 + w_3x_3]$$

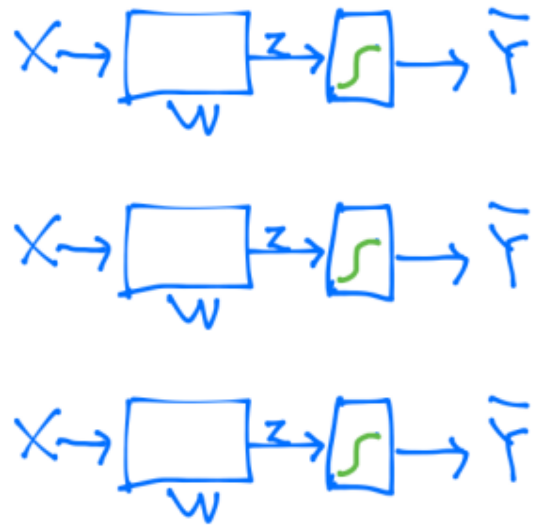


Multinomial classification

하나로 표현해서 복잡도를 내려보자. 코딩할 때도 편하게!

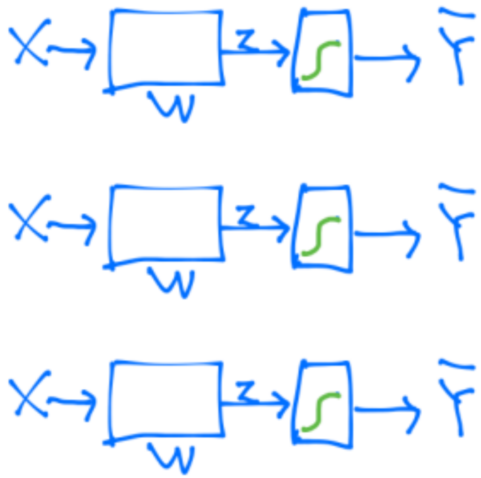
$$\begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} w_1 x_1 + w_2 x_2 + w_3 x_3 \end{bmatrix}$$

$$\begin{bmatrix} w_{A1} & w_{A2} & w_{A3} \\ w_{B1} & w_{B2} & w_{B3} \\ w_{C1} & w_{C2} & w_{C3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} =$$



Multinomial classification

$$\begin{bmatrix} w_{A1} & w_{A2} & w_{A3} \\ w_{B1} & w_{B2} & w_{B3} \\ w_{C1} & w_{C2} & w_{C3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} w_{A1}x_1 + w_{A2}x_2 + w_{A3}x_3 \\ w_{B1}x_1 + w_{B2}x_2 + w_{B3}x_3 \\ w_{C1}x_1 + w_{C2}x_2 + w_{C3}x_3 \end{bmatrix}$$



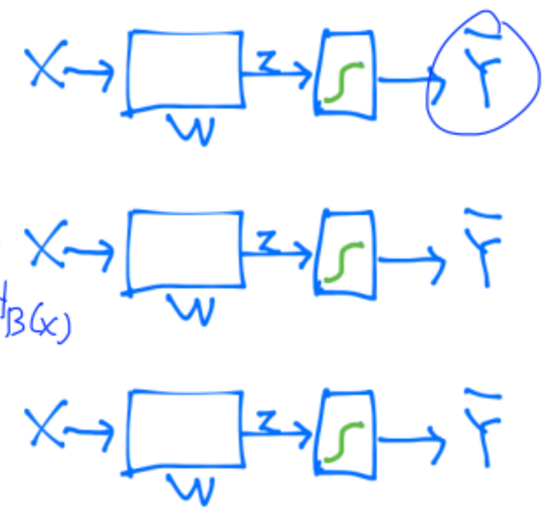
Multinomial classification

$H_A(x)$

$$\begin{bmatrix} w_{A1} & w_{A2} & w_{A3} \\ w_{B1} & w_{B2} & w_{B3} \\ w_{C1} & w_{C2} & w_{C3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} w_{A1}x_1 + w_{A2}x_2 + w_{A3}x_3 \\ w_{B1}x_1 + w_{B2}x_2 + w_{B3}x_3 \\ w_{C1}x_1 + w_{C2}x_2 + w_{C3}x_3 \end{bmatrix} = \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix}$$

$H_B(x)$

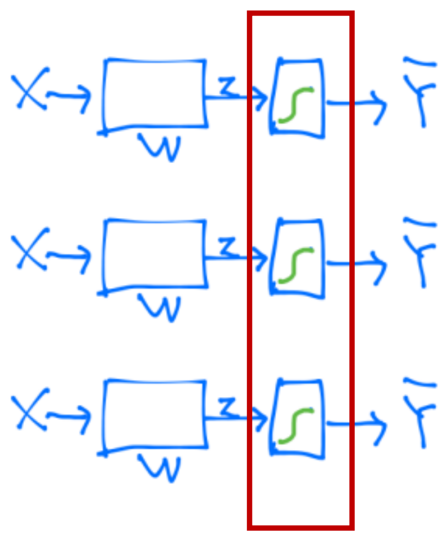
$H_C(x)$



Where is sigmoid?

$$\begin{bmatrix} w_{A1} & w_{A2} & w_{A3} \\ w_{B1} & w_{B2} & w_{B3} \\ w_{C1} & w_{C2} & w_{C3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} w_{A1}x_1 + w_{A2}x_2 + w_{A3}x_3 \\ w_{B1}x_1 + w_{B2}x_2 + w_{B3}x_3 \\ w_{C1}x_1 + w_{C2}x_2 + w_{C3}x_3 \end{bmatrix} = \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix}$$

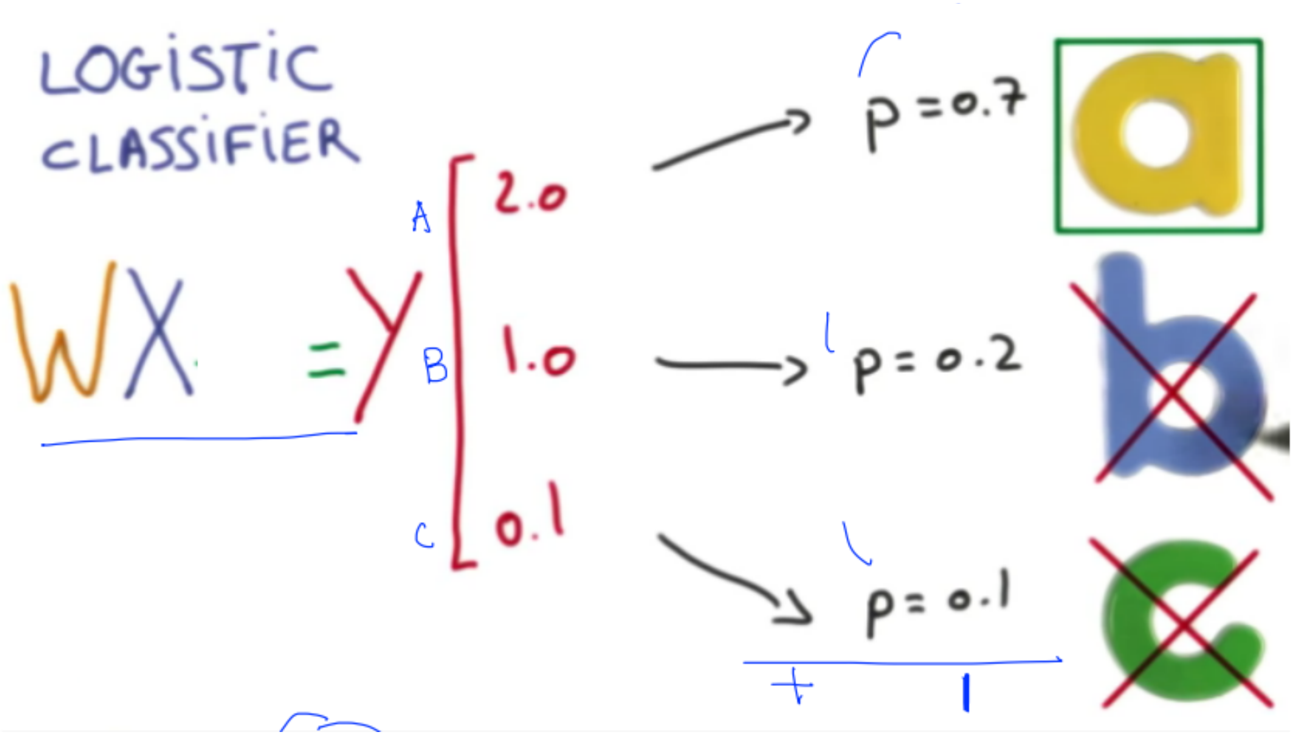
$$\begin{bmatrix} 2.0 \\ 1.0 \\ 0.1 \end{bmatrix} \xrightarrow{0\sim 1} \begin{matrix} 1 \\ 1 \\ 0 \end{matrix}$$



두개가
1이라
문제가 생김

Sigmoid?

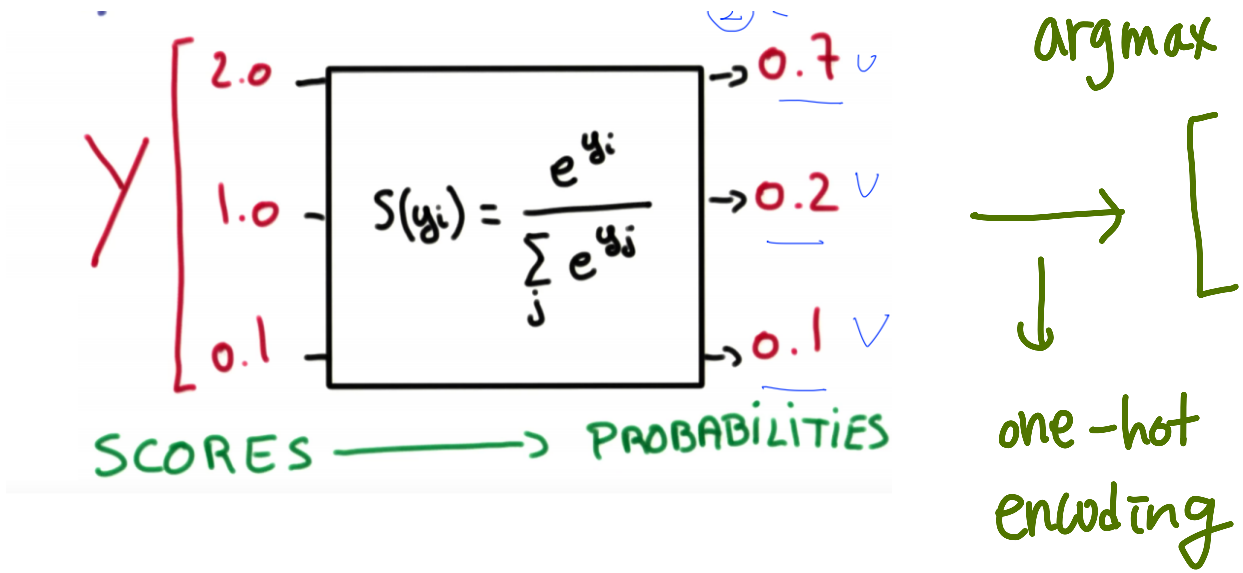
Multiclass classification의 Hypothesis를 **효율적으로 [0,1]** 로 제한하는 방법을 알아보자.



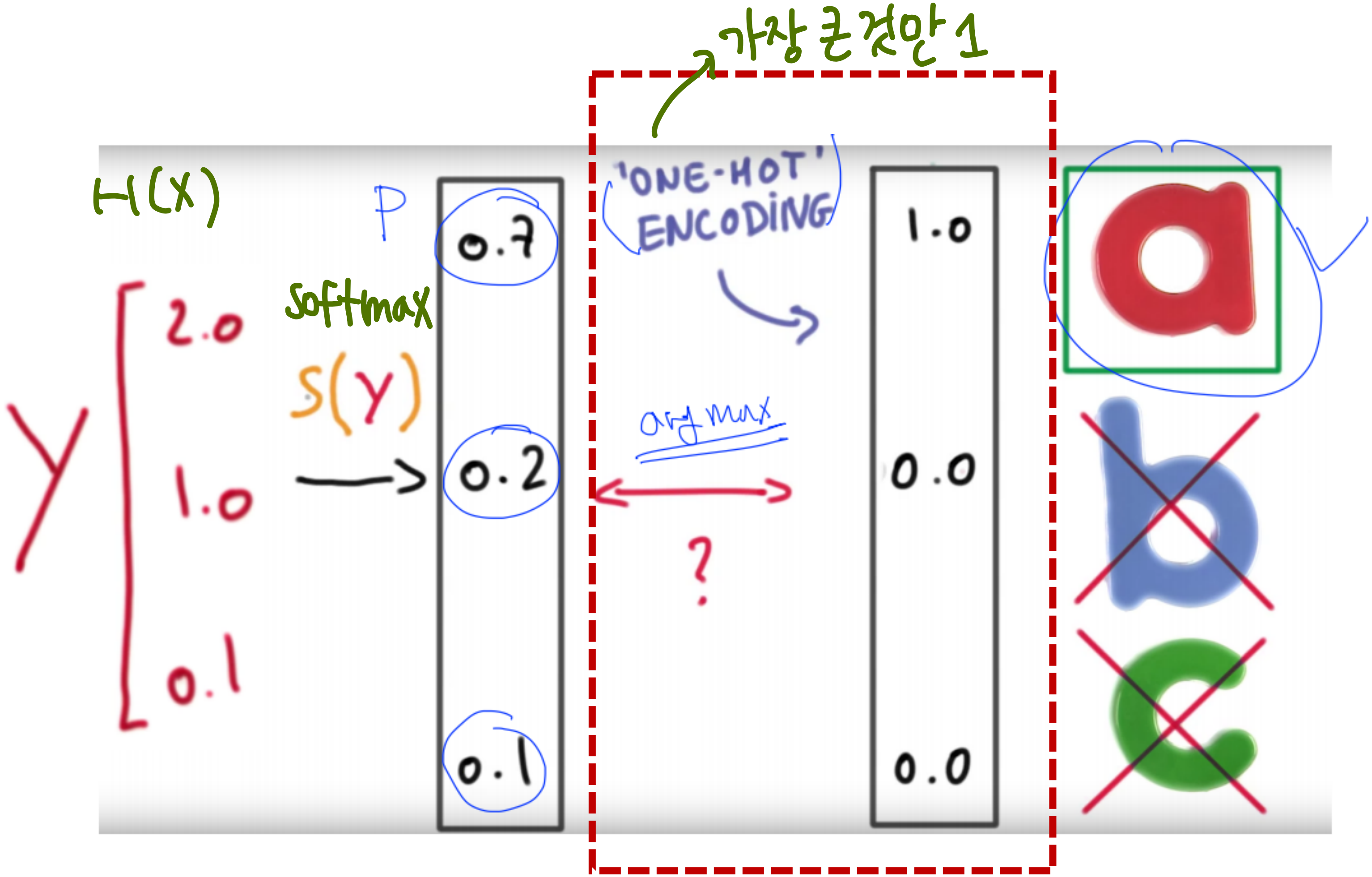
Softmax function instead of sigmoid function

Softmax function

1) 모든 값이 0~1 사이
2) 전체 합이 1 (확률 정규화!!)

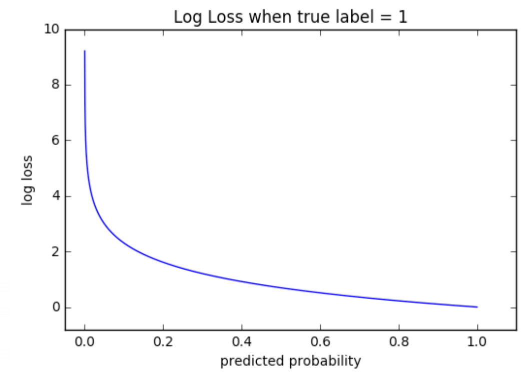


Softmax function (:=Probability)



Cost function for multinomial classification

- **Cross-entropy**



$S(Y) = \bar{Y}$

$L = Y$

$$\begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} \qquad D(S, L) = \boxed{-\sum_i L_i \log(S_i)} \qquad \begin{bmatrix} 1.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

~~**Sigmoid(Y)**~~

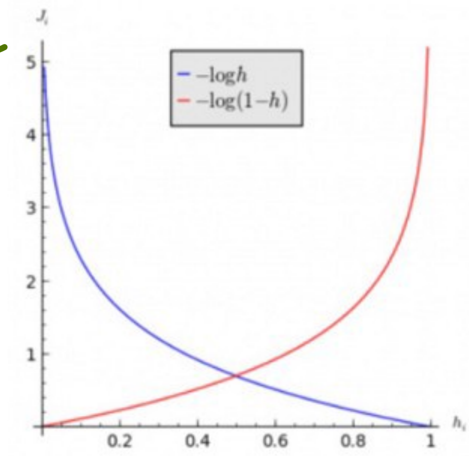
Softmax(Y)

GT: Y

Cross-entropy cost function

$$-\sum_i L_i \log(S_i) = -\sum_i L_i \log(\bar{y}_i) = \sum_i L_i * -\log(\bar{y}_i)$$

predict label



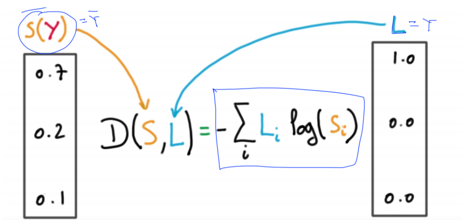
element-wise 곱

정답	$Y = \begin{bmatrix}0\\1\end{bmatrix} = B$
가설이 참 일때	$\bar{Y} = \begin{bmatrix}0\\1\end{bmatrix} = B, \begin{bmatrix}0\\1\end{bmatrix} \odot -log\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}0\\1\end{bmatrix} \odot \begin{bmatrix}\infty\\0\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix} => 0$
가설이 거짓일 때	$\bar{Y} = \begin{bmatrix}1\\0\end{bmatrix} = A, \begin{bmatrix}0\\1\end{bmatrix} \odot -log\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}0\\1\end{bmatrix} \odot \begin{bmatrix}0\\\infty\end{bmatrix} = \begin{bmatrix}0\\\infty\end{bmatrix} => \infty$

Cost function

$$\text{Loss function} = \frac{1}{N}\sum_i D(S(WX_i + b), L_i)$$

$$\mathrm{D(S, L)} = -\sum_i L_i \log(S_i)$$



$$S(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}}$$

$$y = H(X) = \frac{1}{1 + e^{-W^T X}}$$