2일차 : 8월 15일

https://bit.ly/2023어쩌다텍스트분석

Time Table

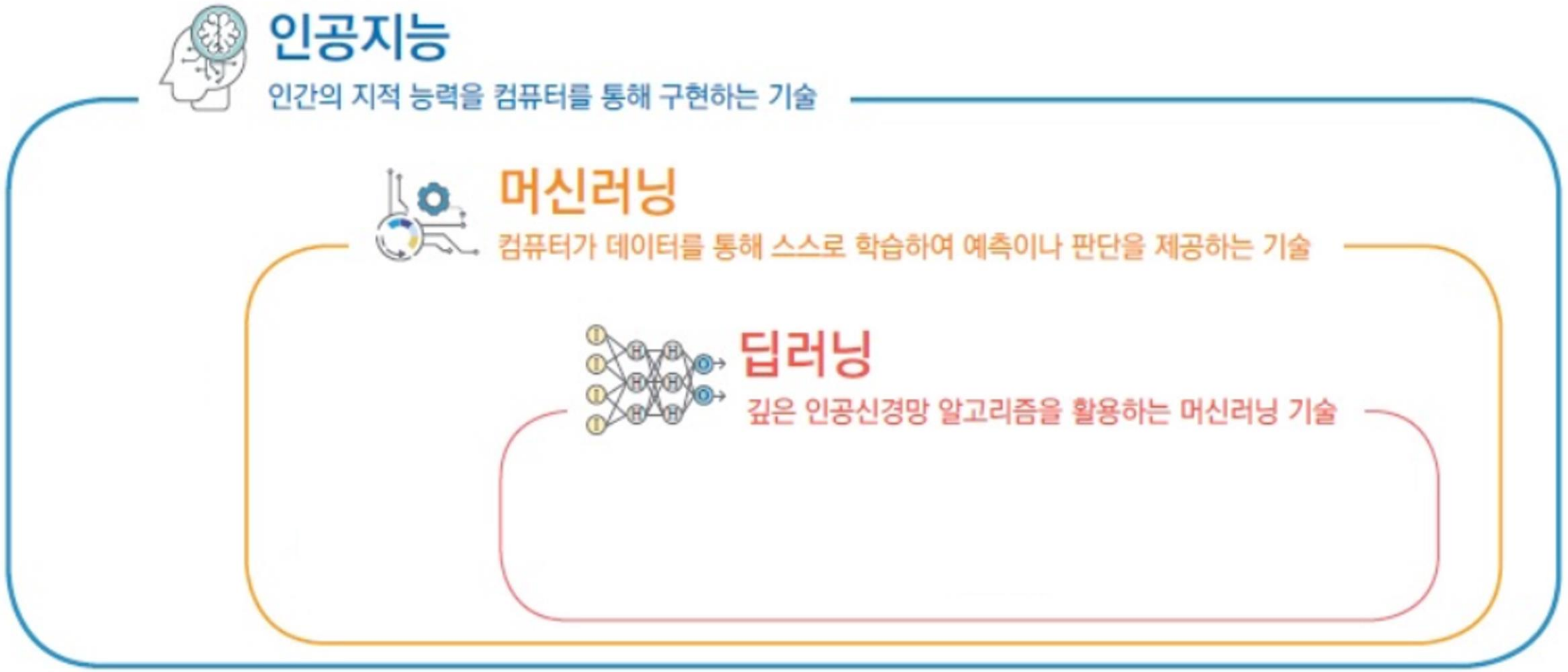
8월 13일	오전 (09:00~12:00)	1. 오리엔테이션 2. 오늘, 우리에게 필요한 머신러닝 1
	오후 (13:00~16:00)	3. 5장 연합뉴스 타이틀 주제 분류
8월 14일	@home	개별 프로젝트 구상 및 발표준비 (생성형 인공지능을 많이 많이 활용해 보세요! ChatGPT, Bing, wrtn, bard 중 하나를 선택하여 활용하기)
8월 15일	오전 (09:00~12:00)	1. 오늘, 우리에게 필요한 머신러닝2 2. 6장 연합뉴스 타이틀 주제 분류 3. 7장 '120다산콜재단' 토픽 모델링과 RNN, LSTM 4. 8장 인프런 이벤트 댓글분석
	오후 (13:00~16:00)	4. 개별 프로젝트 발표 자료 정리 및 발표★

오늘, 우리에게 필요한 머신러닝2

경복고 김선경

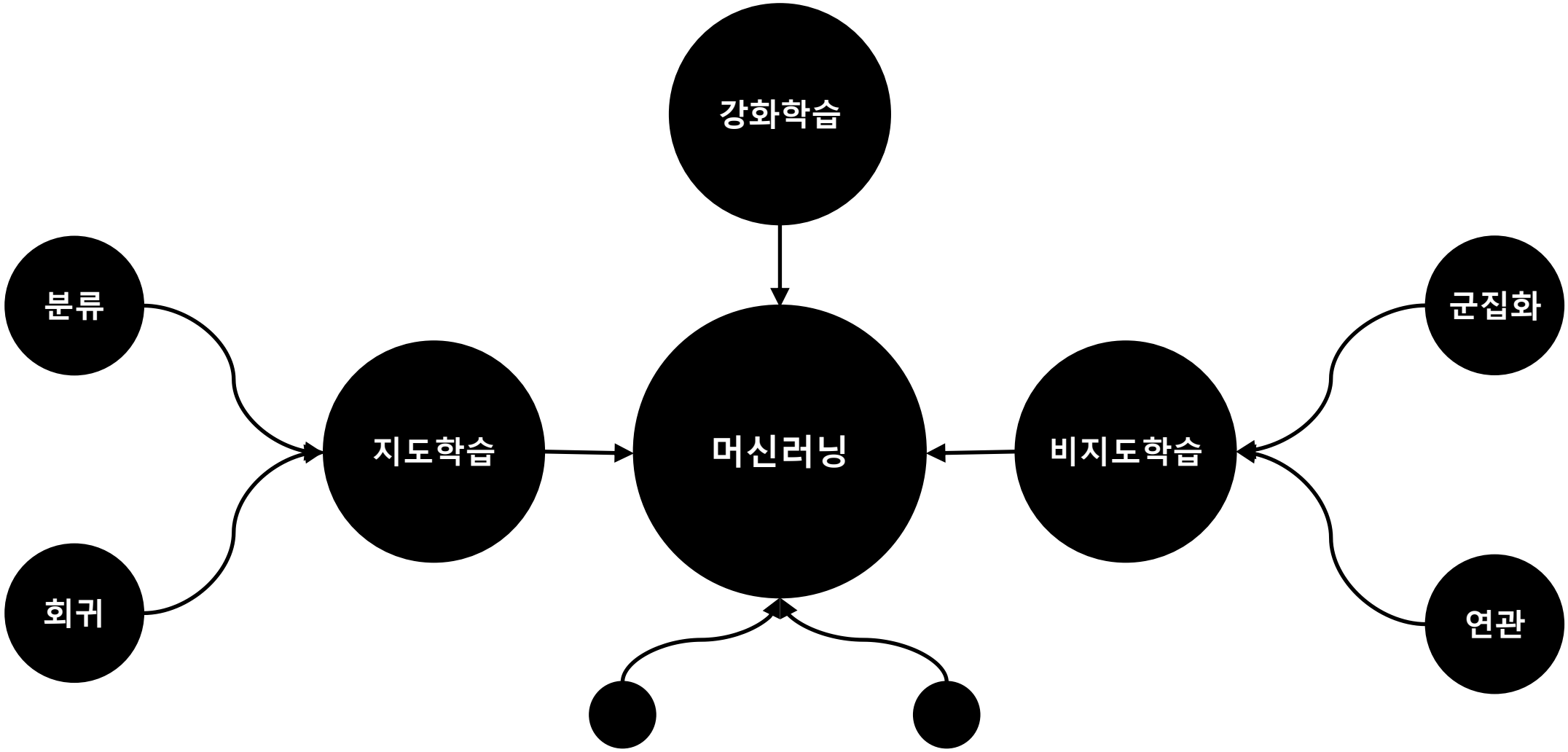
이미지 출처 : https://www.sedaily.com/NewsView/29S01GJ0YH

인공지능, 머신 러닝, 딥러닝



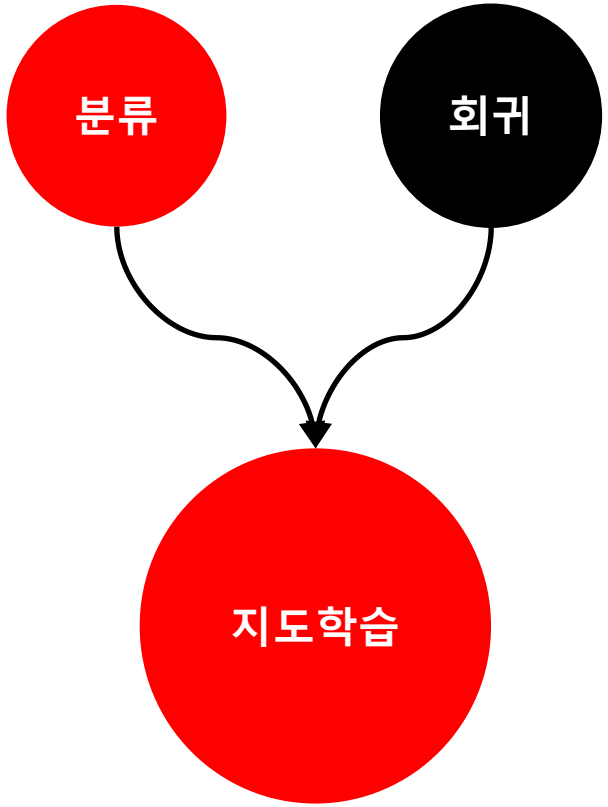
이미지 출처 : https://samstory.coolschool.co.kr/zone/story/modi/streams/76601

머신 러닝의 분류



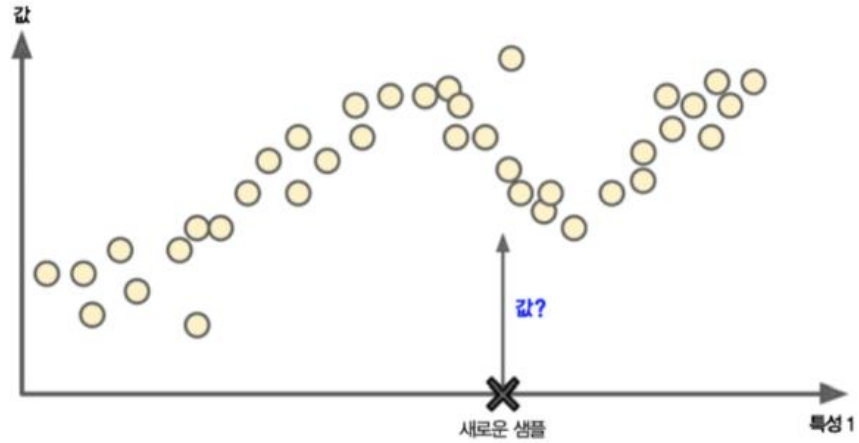
이미지 출처 : https://www.opentutorials.org/course/4548/28934

지도학습, 분류 Vs. 회귀



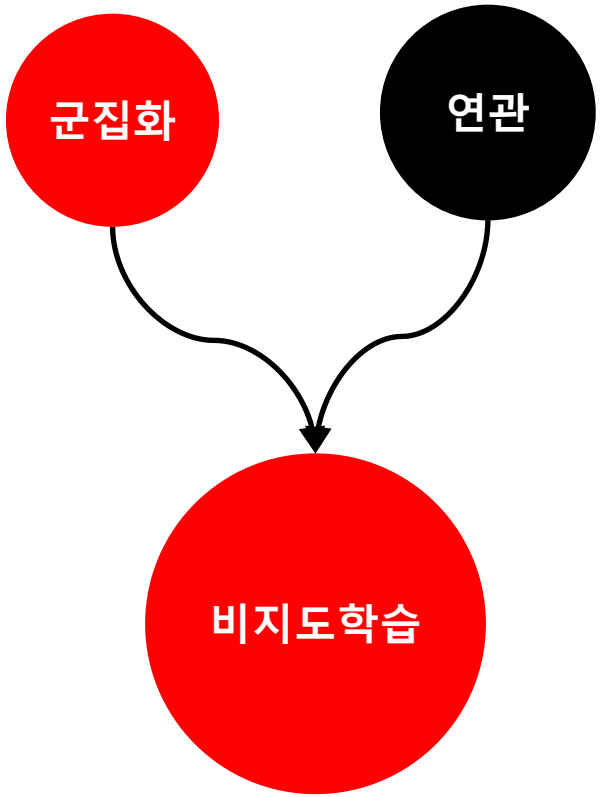
분류(Classification)

회귀(regression)

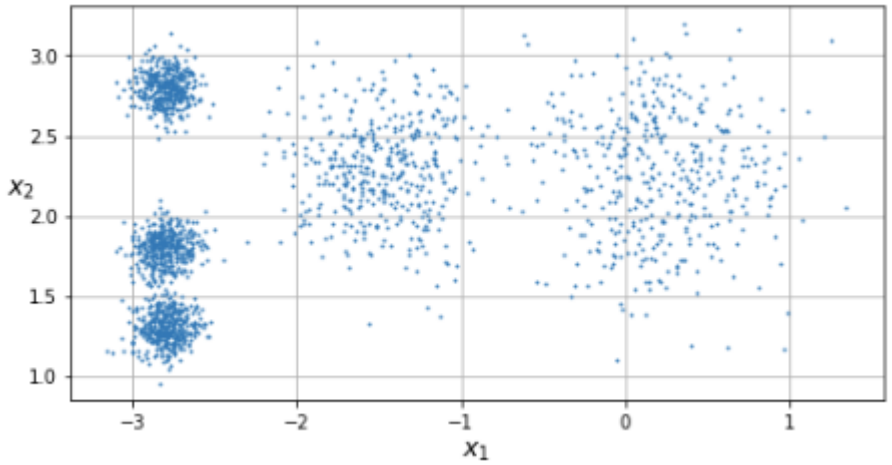


이미지 출처 : https://ittrue.tistory.com/35

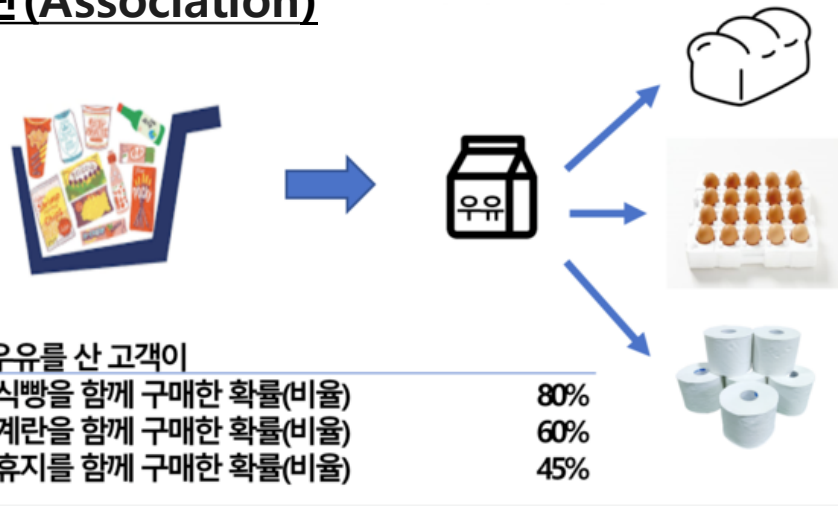
비지도학습, 군집



군집(Clustering)



연관(Association)



우유를 산 고객이

식빵을 함께 구매한 확률(비율)	80%
계란을 함께 구매한 확률(비율)	60%
휴지를 함께 구매한 확률(비율)	45%

이미지 출처 : https://codingalzi.github.io/handson-ml3/unsupervised_learning.html ,
http://www.sbr.ai/news/articleView.html?idxno=1582

우리가 공부할 5,6,7,8장의 목차를 살펴보았더니,

5장 연합뉴스 타이틀 주제 분류

- 학습 세트와 시험 세트 분리하기 / 랜덤포레스트

6장 국민청원 데이터 시각화와 분류

- 학습 세트와 시험 세트 분리하기 / 이진 분류, LightGBM

7장 '120다산콜재단' 토픽 모델링과 RNN, LSTM

- 학습-시험 데이터 세트 분리하기 / Bidirectional LSTM

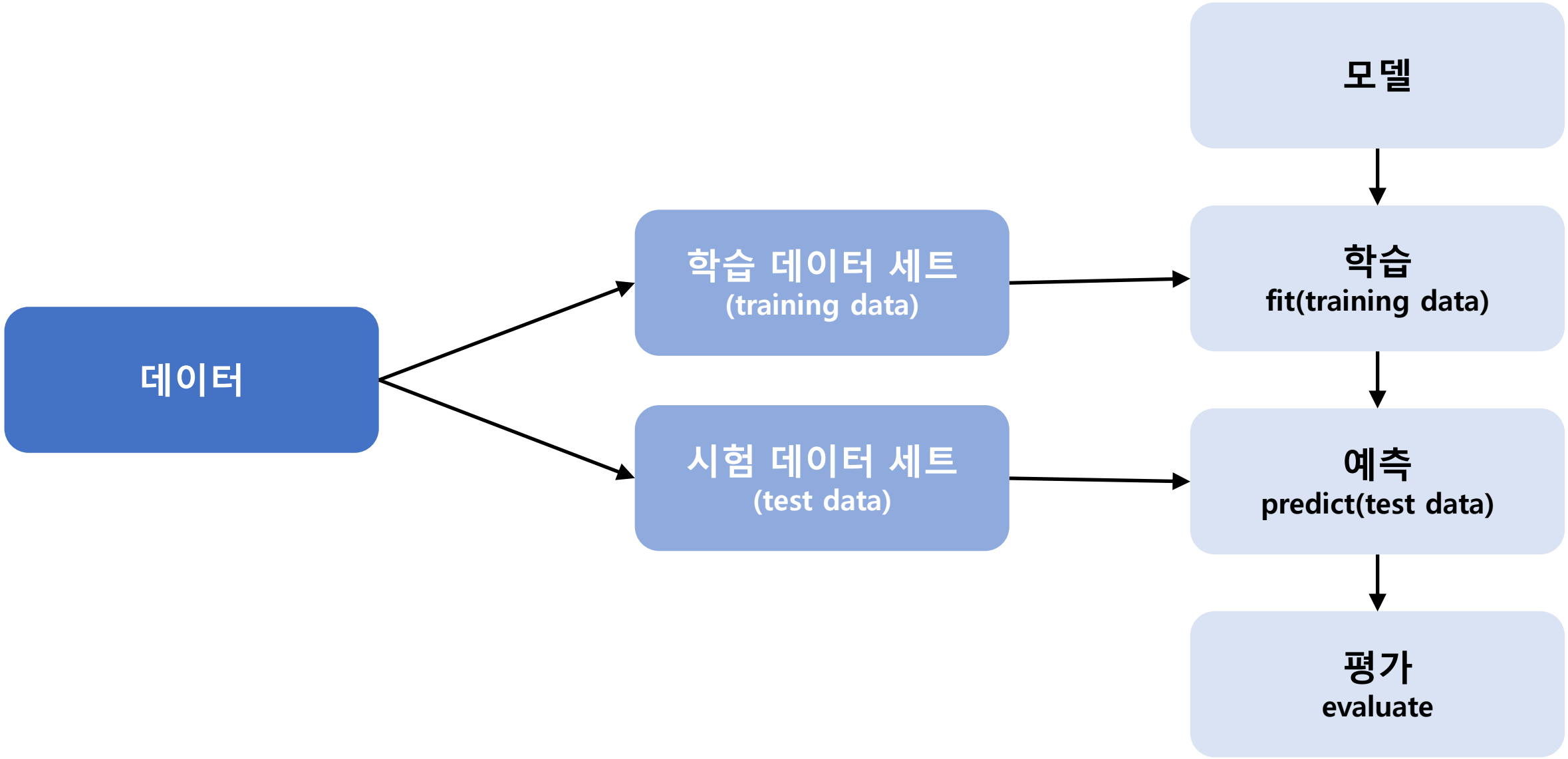
지도학습

8장 인프런 이벤트 댓글 분석

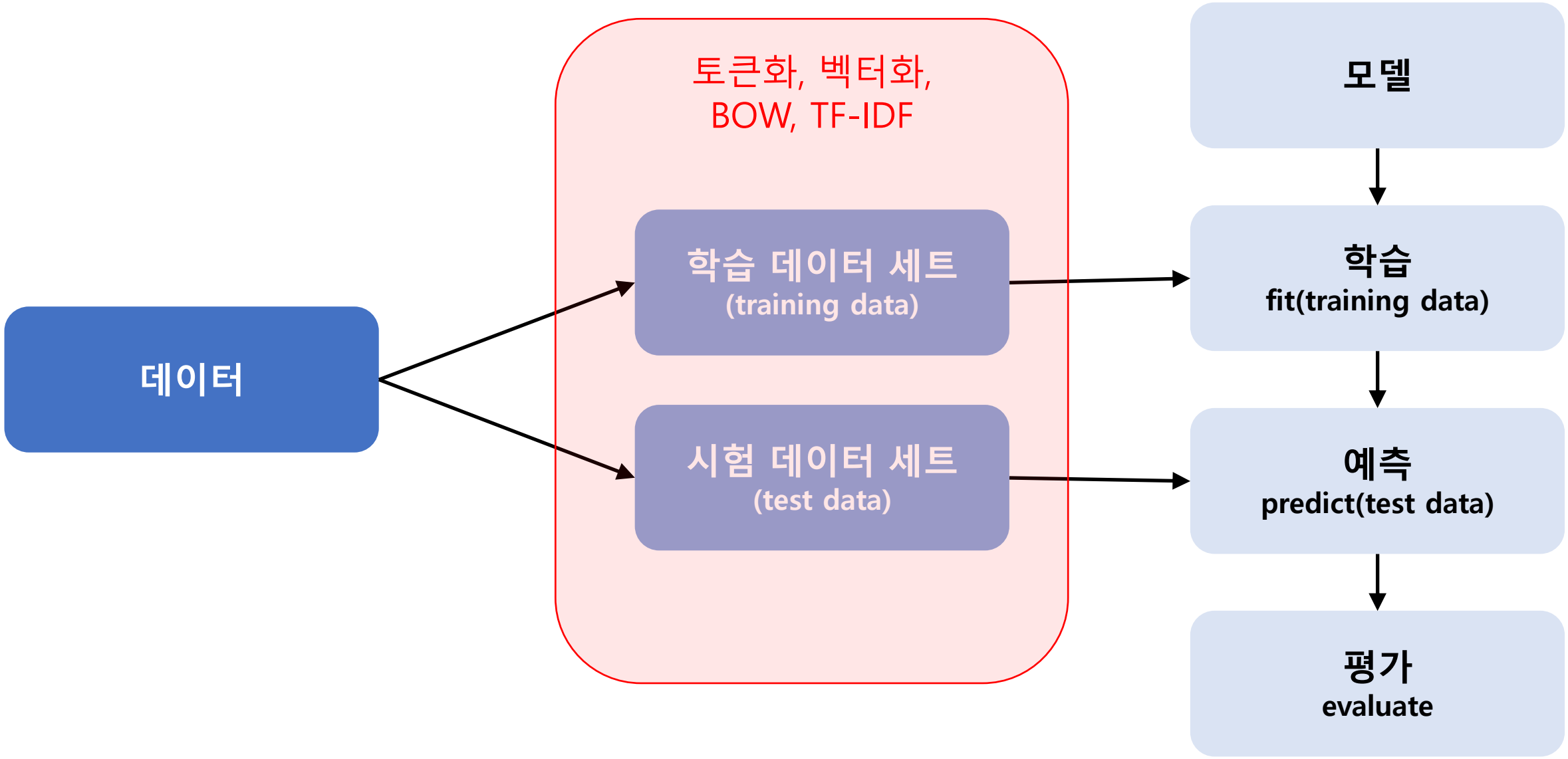
- 군집화 하기 / KMeans

비지도학습

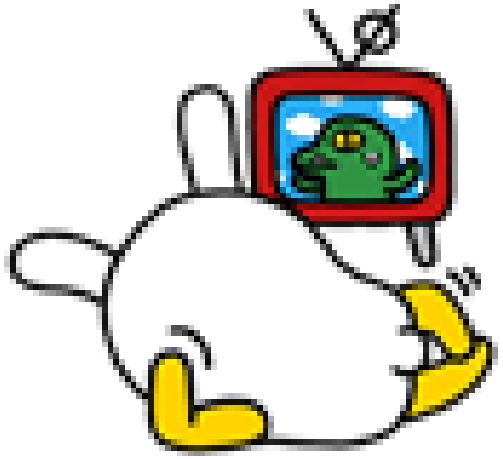
모델 생성, 훈련, 예측, 평가


모델
학습
fit(training data)
예측
predict(test data)
평가
evaluate
데이터
학습 데이터 세트
(training data)
시험 데이터 세트
(test data)

모델 생성, 훈련, 예측, 평가


토큰화, 벡터화,
BOW, TF-IDF
데이터
학습 데이터 세트
(training data)
시험 데이터 세트
(test data)
모델
학습
fit(training data)
예측
predict(test data)
평가
evaluate

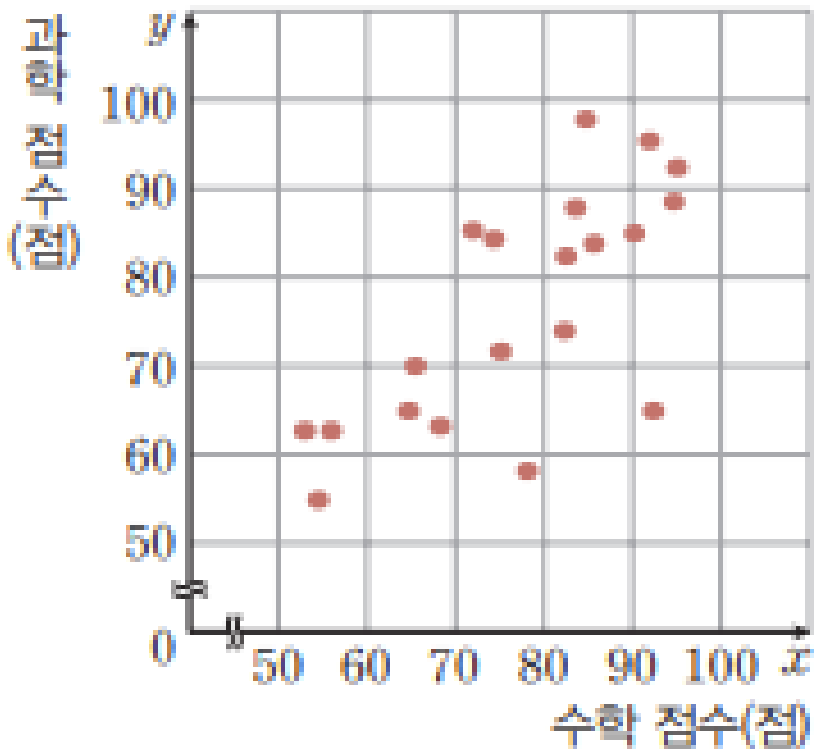
지금부터, 편안하게 들어주세요.



단, 어려워보이는 수학식이 나오면 해석을 부탁할 수도 있습니다. 🙏

자료의 경향성과 예측

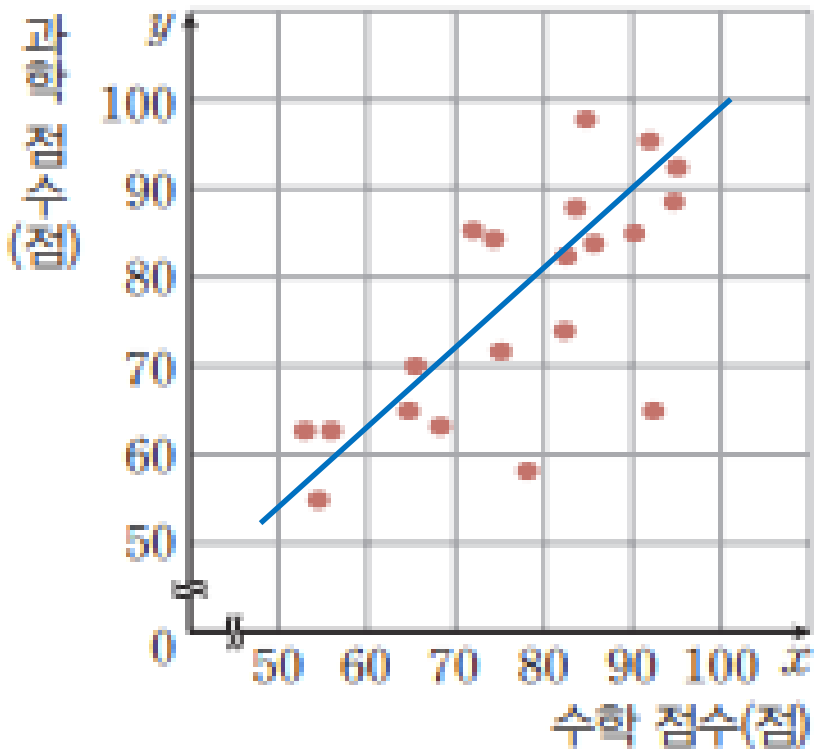
번호	수학	과학	번호	수학	과학
1	90	86	11	73	84
2	64	65	12	82	83
3	94	89	13	78	59
4	57	62	14	68	64
5	82	74	15	54	55
6	92	64	16	92	95
7	84	98	17	65	70
8	72	85	18	76	72
9	52	62	19	83	88
10	86	84	20	95	93



수학점수(x)와 과학점수(y)에 대한 산점도

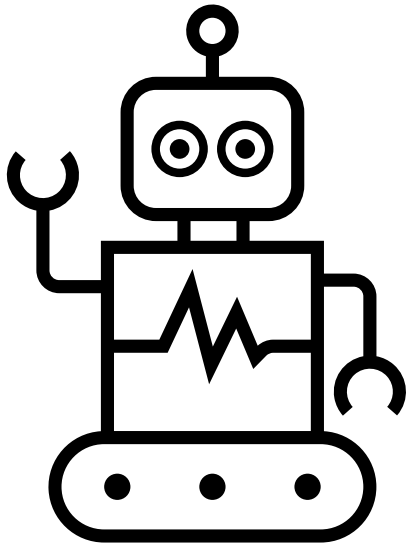
자료의 경향성과 예측

번호	수학	과학	번호	수학	과학
1	90	86	11	73	84
2	64	65	12	82	83
3	94	89	13	78	59
4	57	62	14	68	64
5	82	74	15	54	55
6	92	64	16	92	95
7	84	98	17	65	70
8	72	85	18	76	72
9	52	62	19	83	88
10	86	84	20	95	93



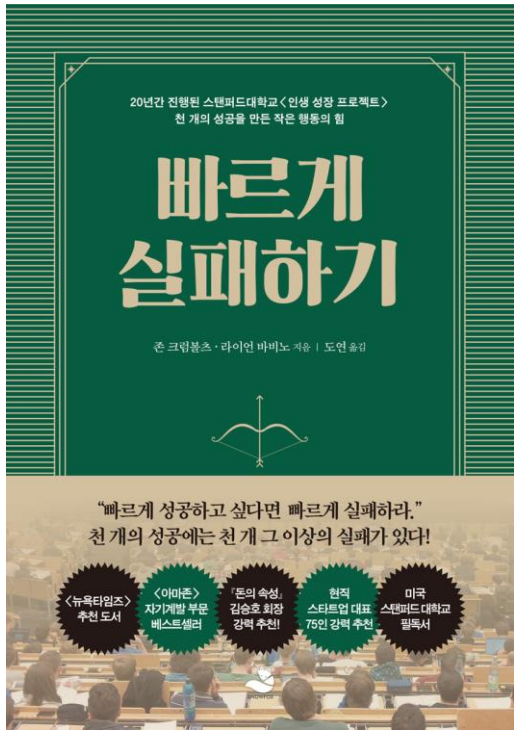
산점도에 선을 하나 그으면?

쌤은,



Can machines think?

Yes, But machines can’t think as people do.

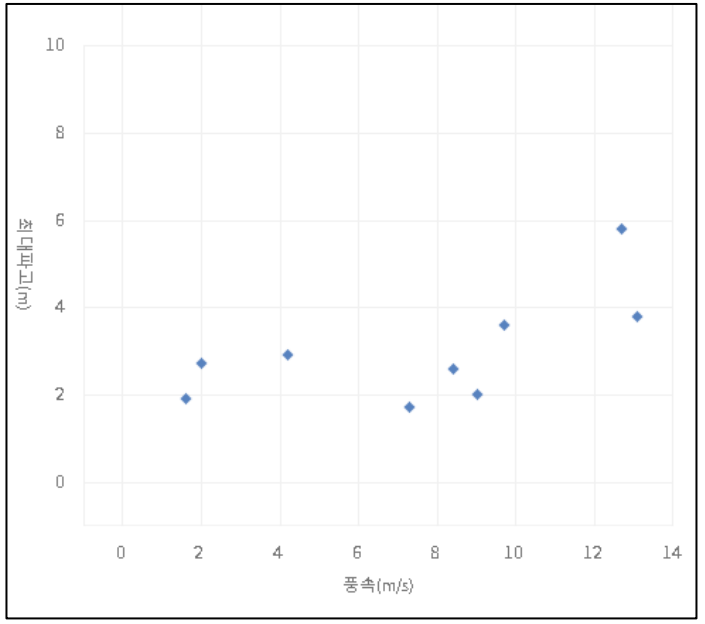


Fail Fast, Fail Often

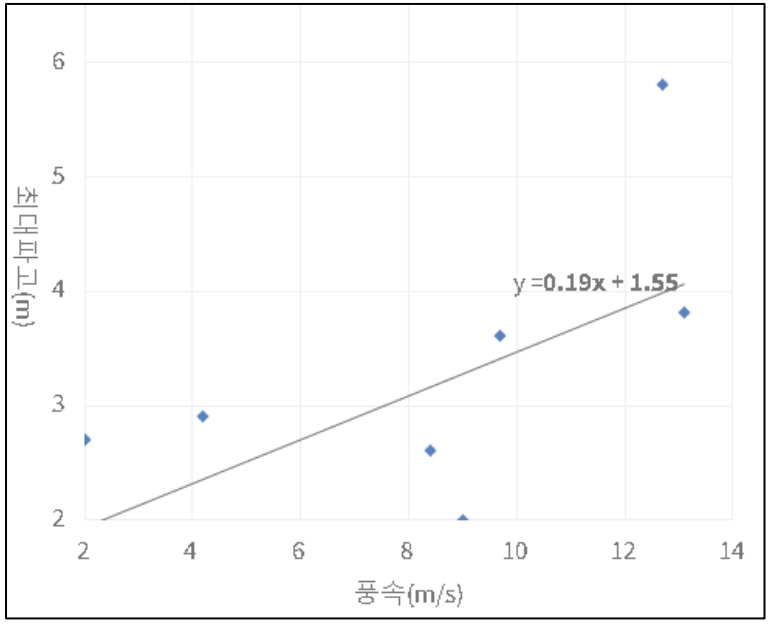
자료의 경향성과 예측

지점	풍속(m/s)	최대 파고(m)
울릉도	12.7	5.8
덕적도	1.6	1.9
포항	13.1	3.8
외연도	2.0	2.7
거제도	7.3	1.7
서귀포	8.4	2.6
통영	9.0	2.0
인천	4.2	2.9
울산	9.7	3.6

(출처: 기상자료개방포털, 2021)



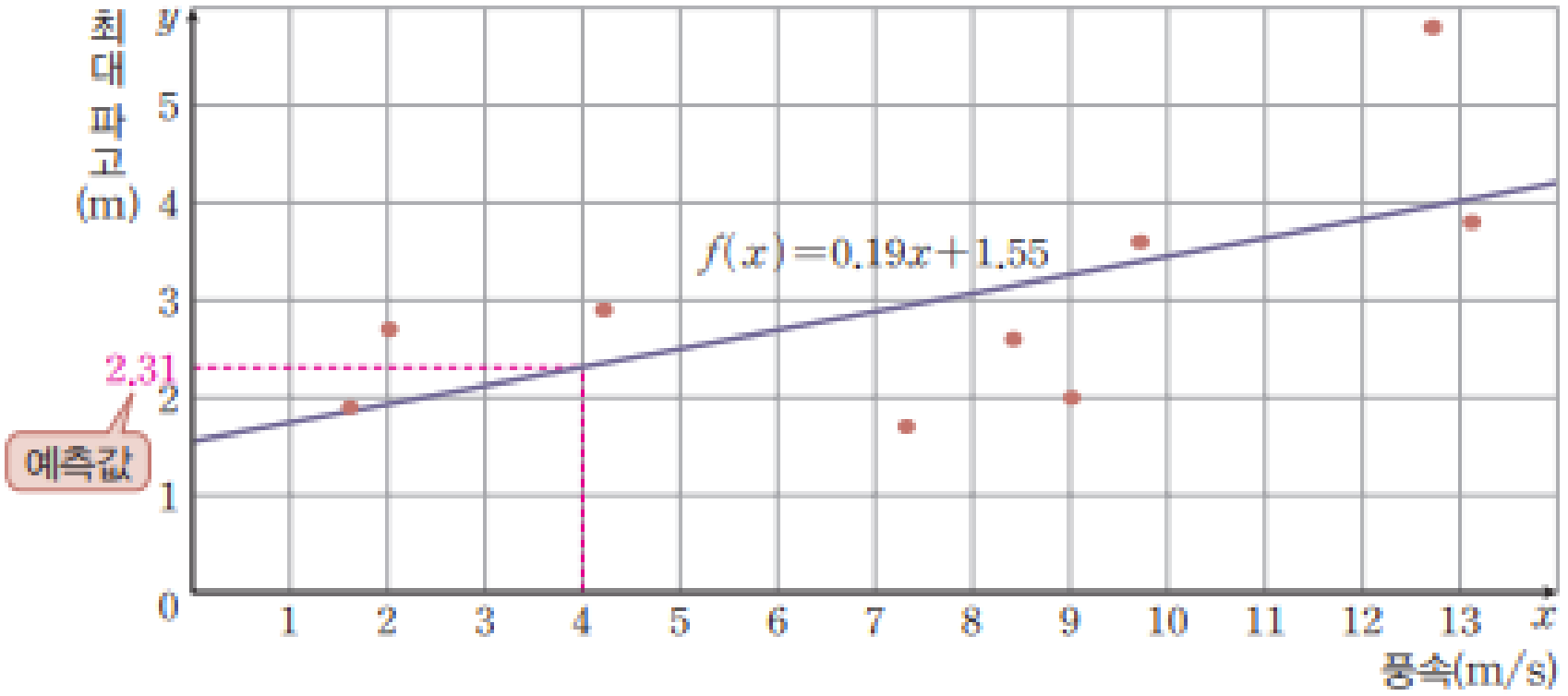
풍속과 최대파고의 산점도



추세선과 관계식

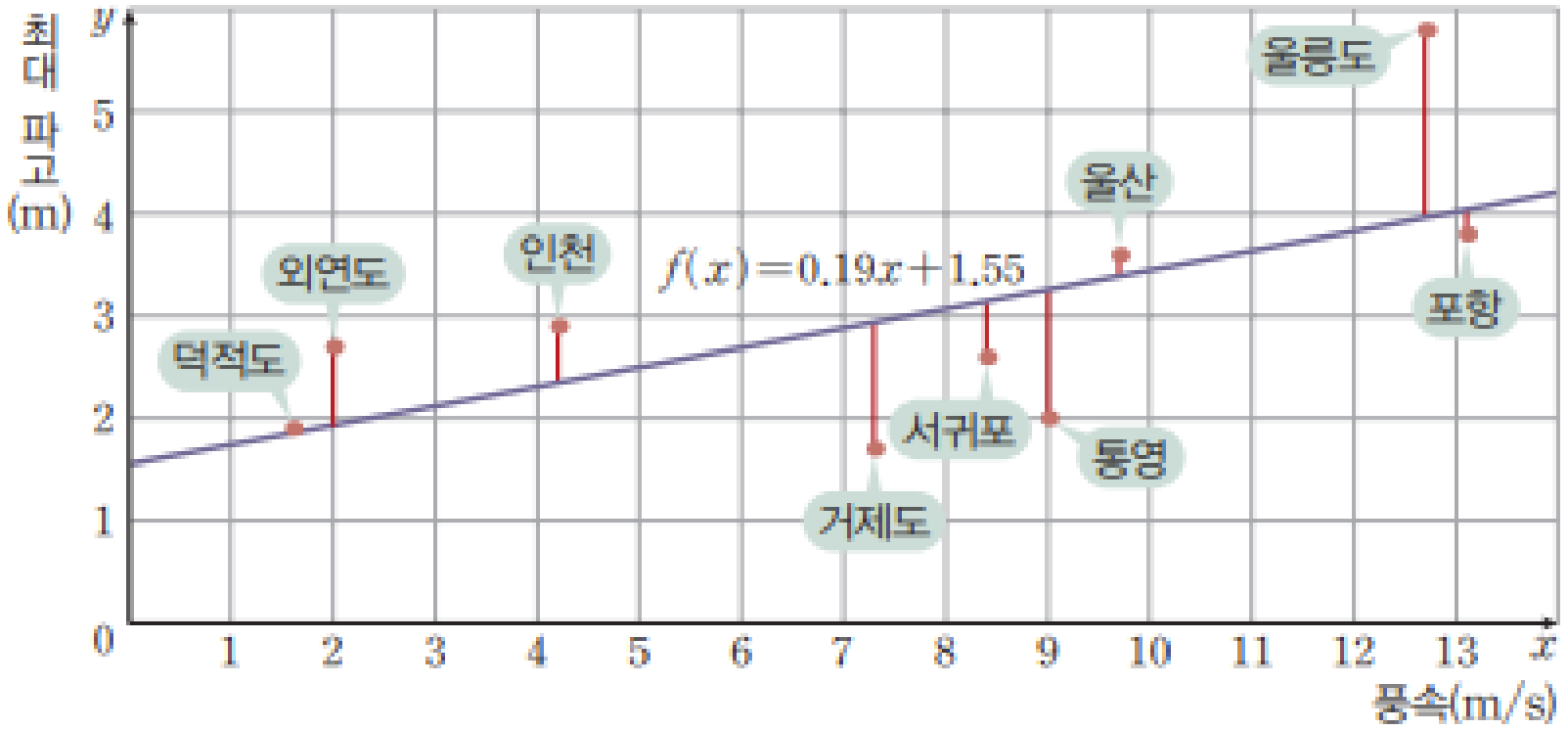
자료의 경향성과 예측

- 우리는, 이 추세선으로 풍속이 4m/s이면 최대파고가 2.31m일 것이라고 "예측"한다.



예측과 오차

- 추세선에 따른 각 지점의 오차



예측과 오차

- 예측값과 오차

지점	풍속(m/s)	최대 파고(m)	예측값(m)	오차(m)
울릉도	12.7	5.8	3.963	1.837
덕적도	1.6	1.9	1.854	0.046
포항	13.1	3.8	4.039	−0.239
외연도	2.0	2.7	1.930	0.770
거제도	7.3	1.7	2.937	−1.237
서귀포	8.4	2.6	3.146	−0.546
통영	9.0	2.0	3.260	−1.260
인천	4.2	2.9	2.348	0.552
울산	9.7	3.6	3.393	0.207

- 예측이 좋은 추세선이란?

예측과 오차

$$\text{오차의 합} = \sum_{i=1}^{n} (p_i - y_i)^2$$

$$\text{평균 제곱 오차}(MSE) = \frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2$$

$$\text{평균 제곱근 오차}(RMSE) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2}$$

선형 회귀(Linear Regression)

선형 회귀(linear regression)란,

임의의 직선($y = ax + b$)을 그어서 이에 대한 평균 제곱근 오차를 구하고,

이 값을 가장 작게 만들어 주는 기울기 a와 절편 b를 찾아가는 작업이다.

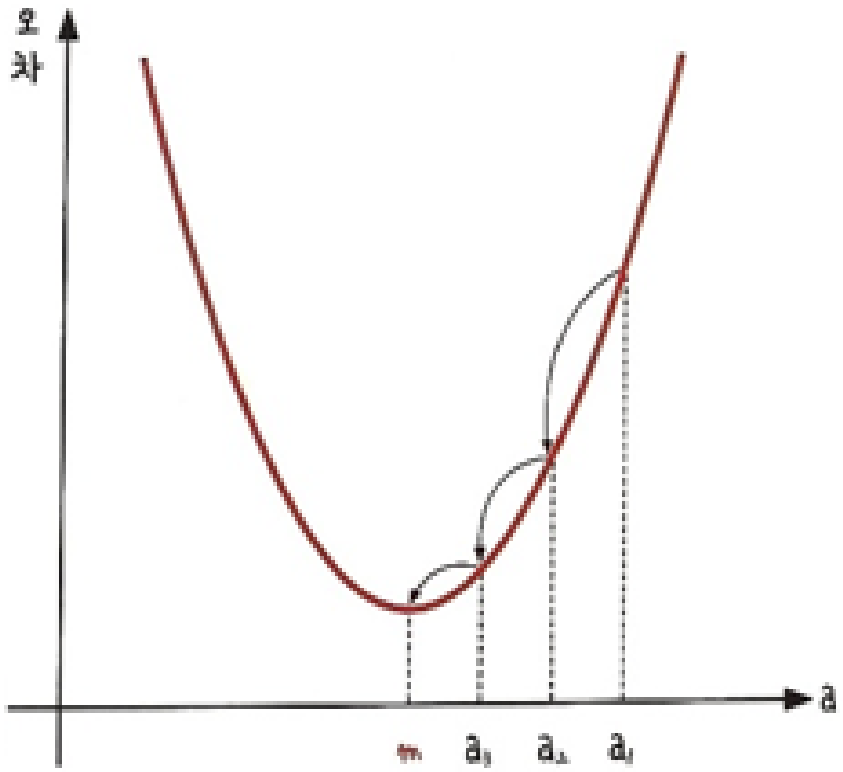
선형 회귀(Linear Regression) : 예측과 오차

- 오차 수정하기

  기울기a와 오차의 상관관계는?

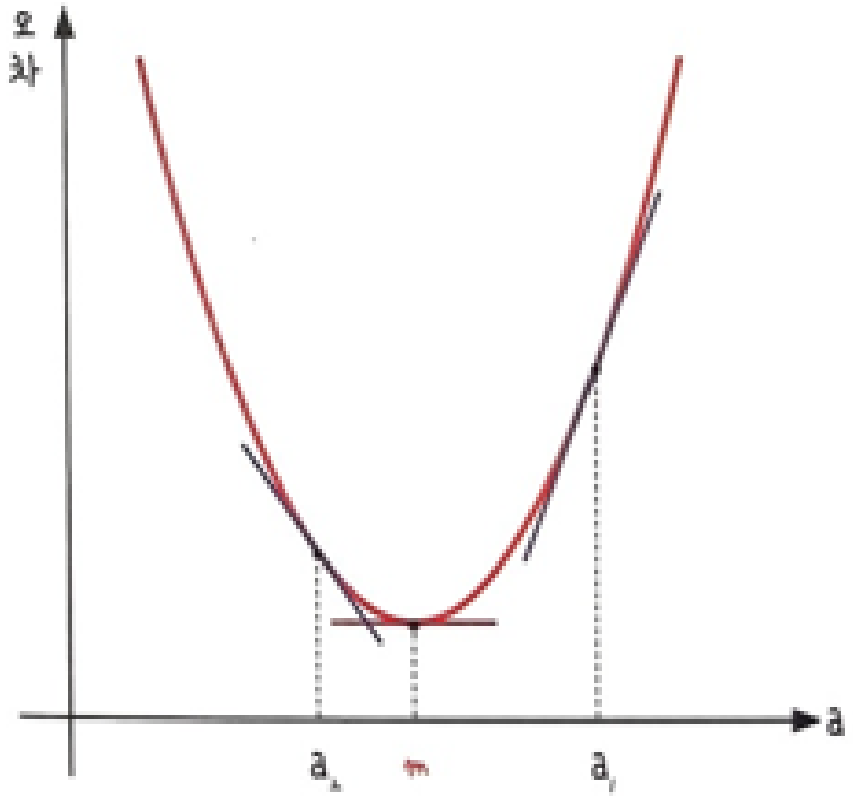
선형 회귀(Linear Regression) : 예측과 오차

- 오차가 가장 작은 점은 어디일까요?



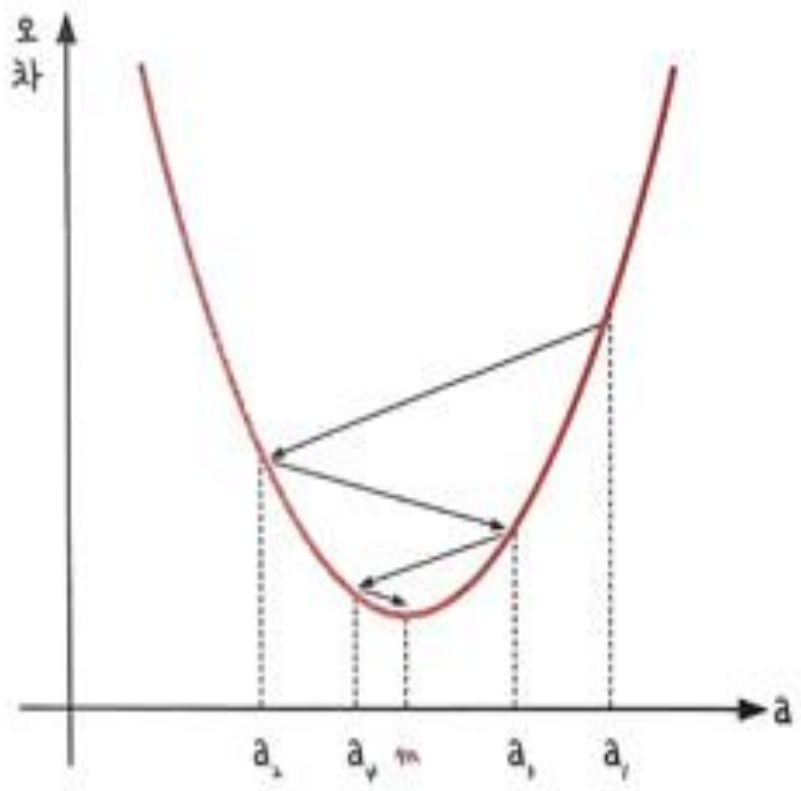
선형 회귀(Linear Regression) : 예측과 오차

- 각 점에서 순간의 기울기를 구할 때, 기울기가 0이 되는 지점



선형 회귀(Linear Regression) : 예측과 오차

- 우리가 오차가 가장 작은 점을 구하는 방법

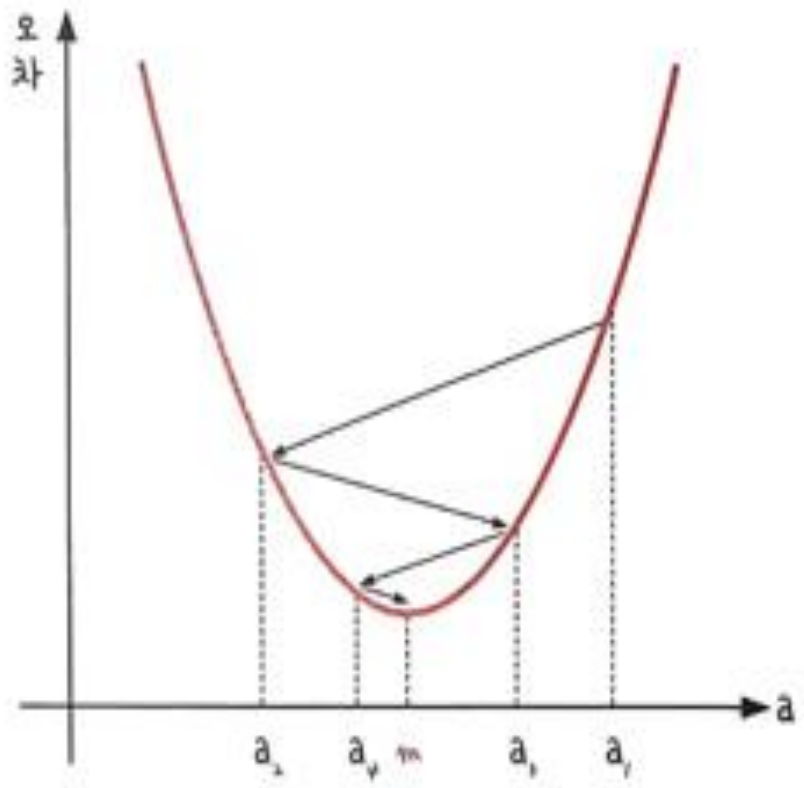


1. a_1에서 미분을 구한다.
2. 구해진 기울기의 반대방향으로 얼마간 이동시킨

   a_2에서 미분을 구한다.
3. a_3에서 미분을 구한다
4. 3의 값이 0이 아니면 위의 과정을 반복한다.

선형 회귀(Linear Regression) : 예측과 오차

- 우리가 오차가 가장 작은 점을 구하는 방법



1. a_1에서 미분을 구한다.

2. 구해진 기울기의 반대방향으로 얼마간 이동시킨

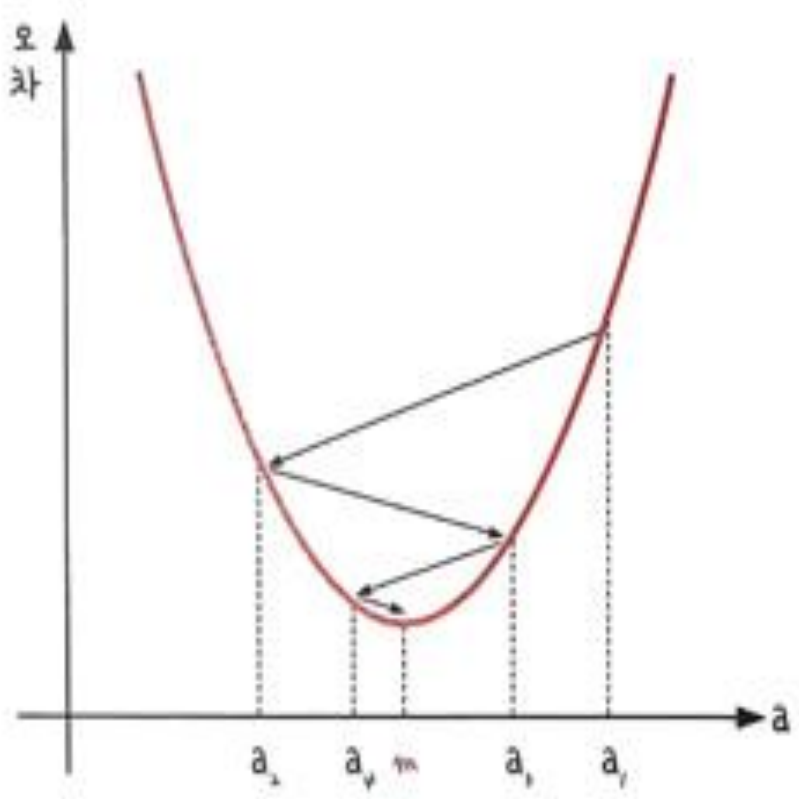
   a_2에서 미분을 구한다.

3. a_3에서 미분을 구한다

4. 3의 값이 0이 아니면 위의 과정을 반복한다.

선형 회귀(Linear Regression) : 예측과 오차

- 우리가 오차가 가장 작은 점을 구하는 방법 : 경사하강법

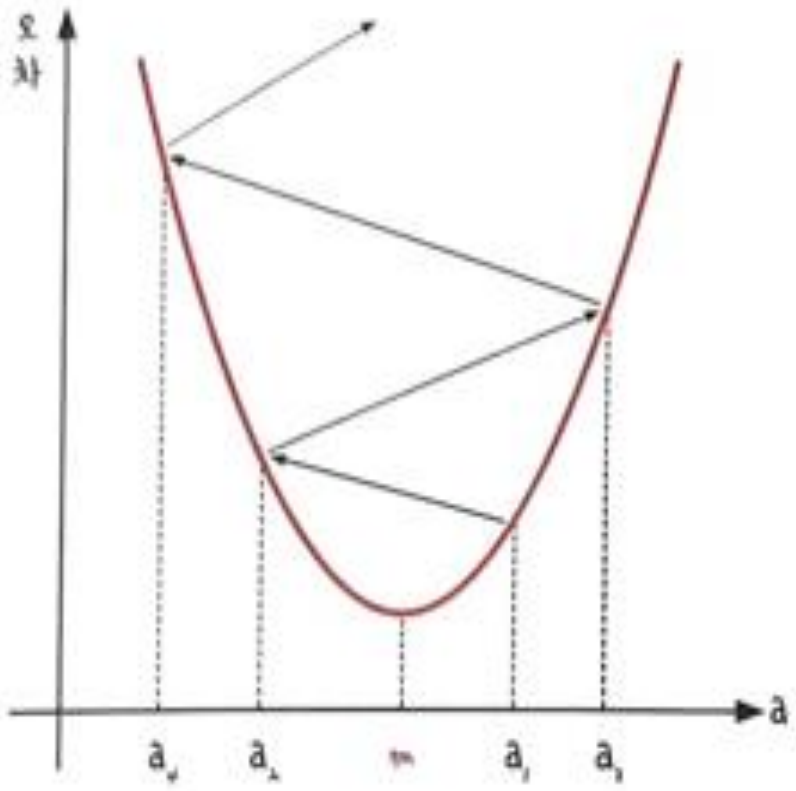


1. a_1에서 미분을 구한다.
2. 구해진 기울기의 반대방향으로 얼마간 이동시킨

   a_2에서 미분을 구한다.
3. a_3에서 미분을 구한다
4. 3의 값이 0이 아니면 위의 과정을 반복한다.

선형 회귀(Linear Regression) : 예측과 오차

- 우리가 오차가 가장 작은 점을 구하는 방법 : 경사하강법



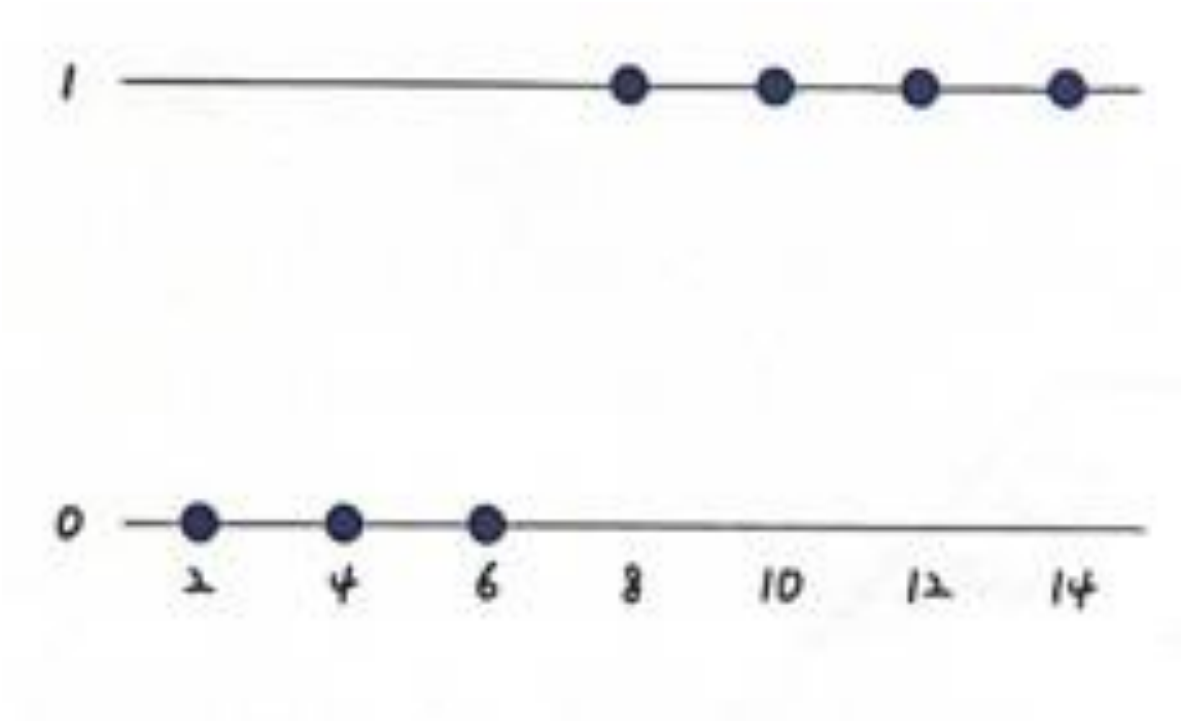
학습률!

1. a_1에서 미분을 구한다.
2. 구해진 기울기의 반대방향으로 얼마간 이동시킨

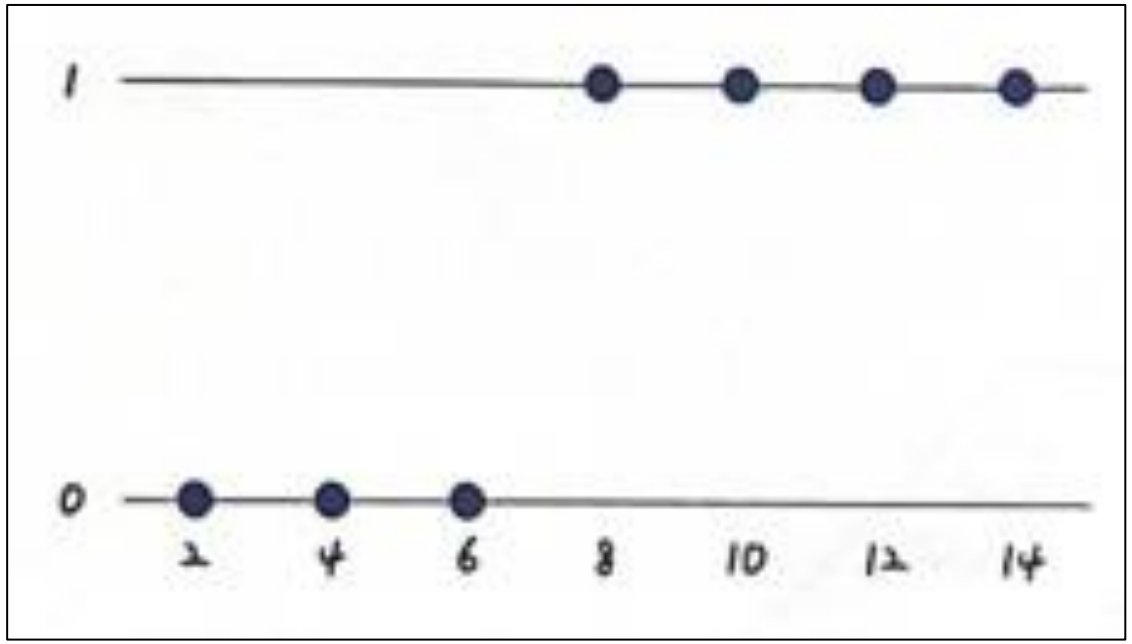
   a_2에서 미분을 구한다.
3. a_3에서 미분을 구한다
4. 3의 값이 0이 아니면 위의 과정을 반복한다.

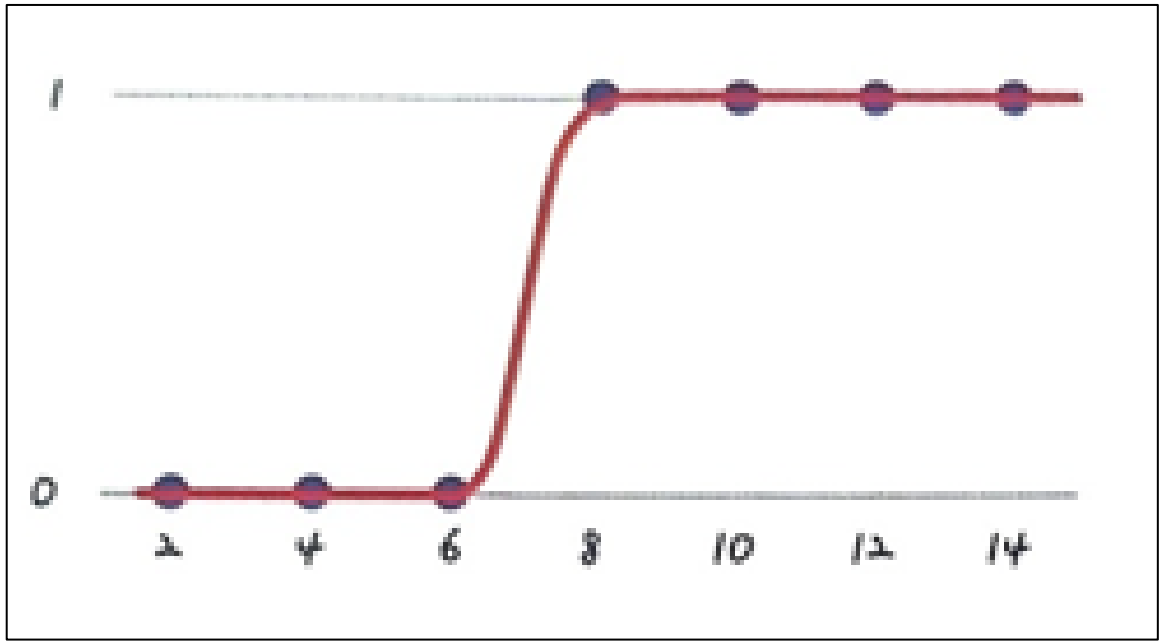
로지스틱 회귀(Logistic Regression)

공부한 시간	합격여부
2	불합격
4	불합격
6	불합격
8	합격
10	합격
12	합격
14	합격

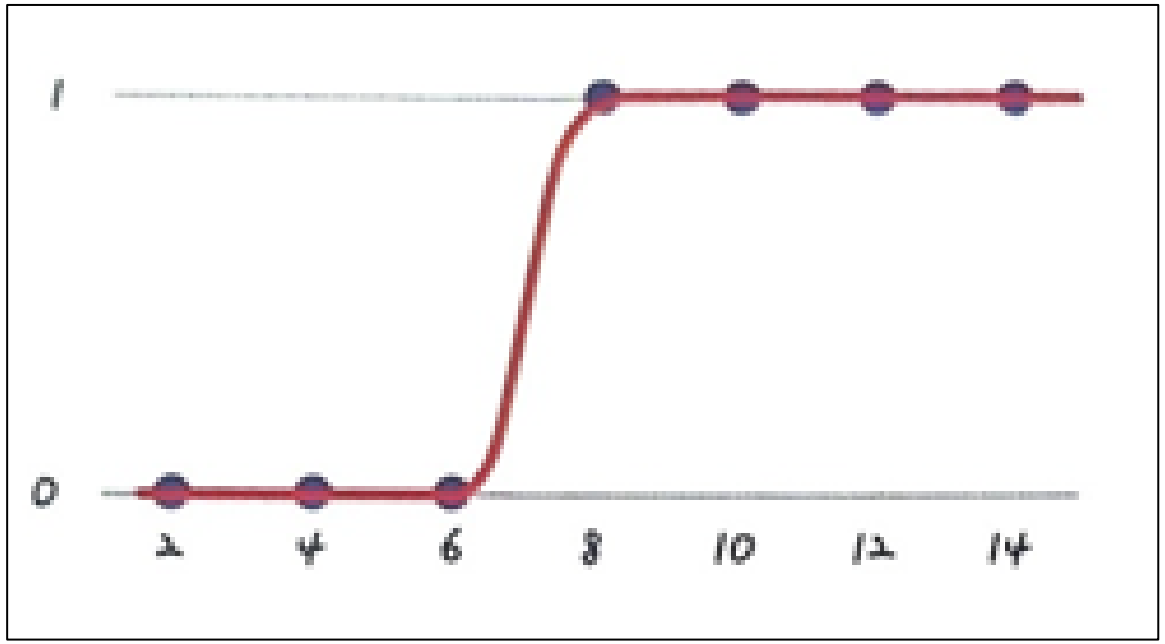


로지스틱 회귀(Logistic Regression) : 선긋기





로지스틱 회귀(Logistic Regression) : 시그모이드 함수



$$y = \frac{1}{1 + e^{(-ax+b)}}$$

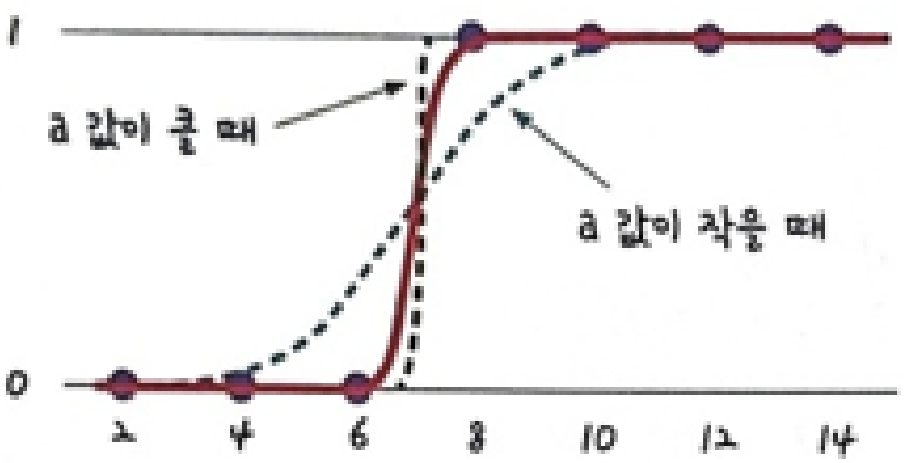
로지스틱 회귀(Logistic Regression) : 시그모이드 함수

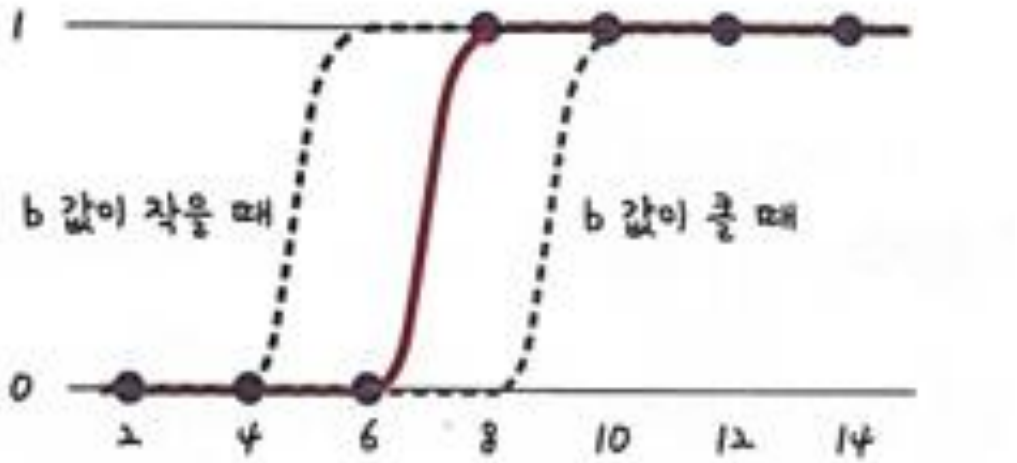
$$y = \frac{1}{1 + e^{(-ax+b)}}$$

$$ax + b$$

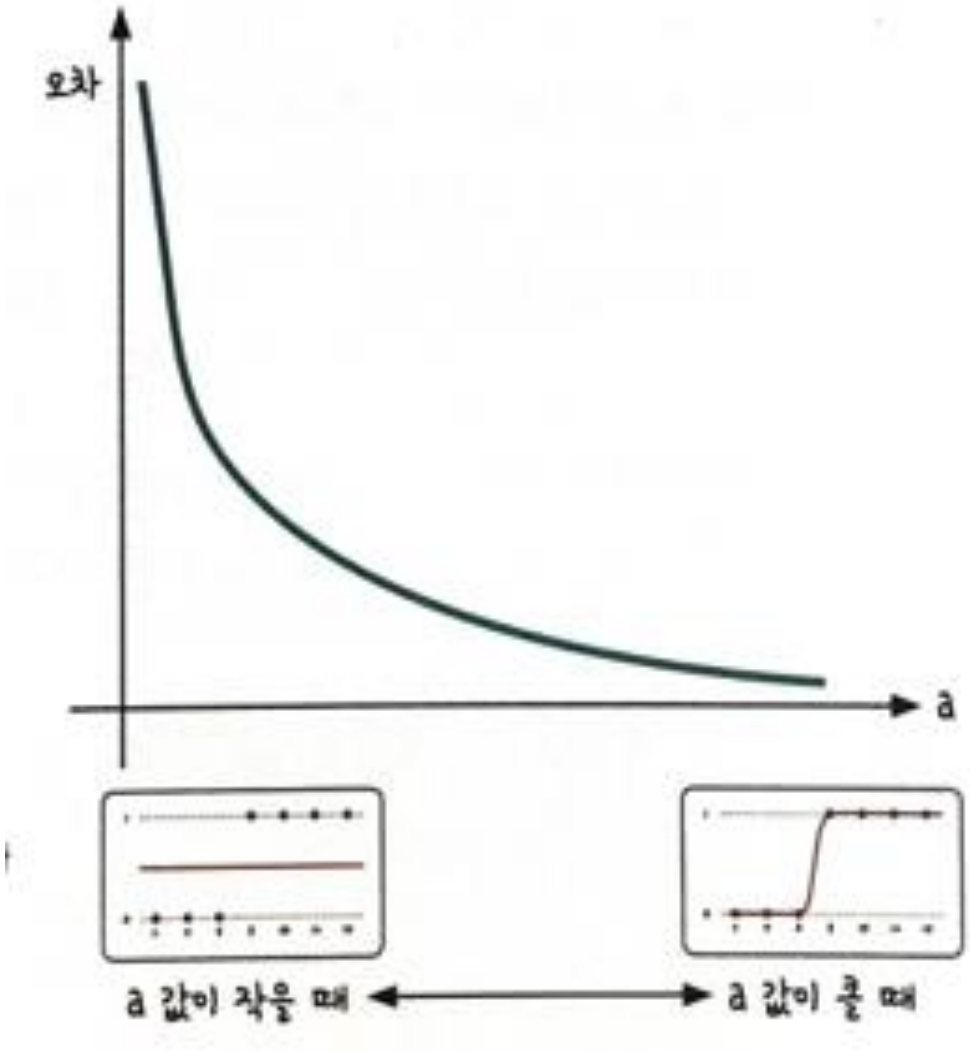
로지스틱 회귀(Logistic Regression) : 시그모이드 함수

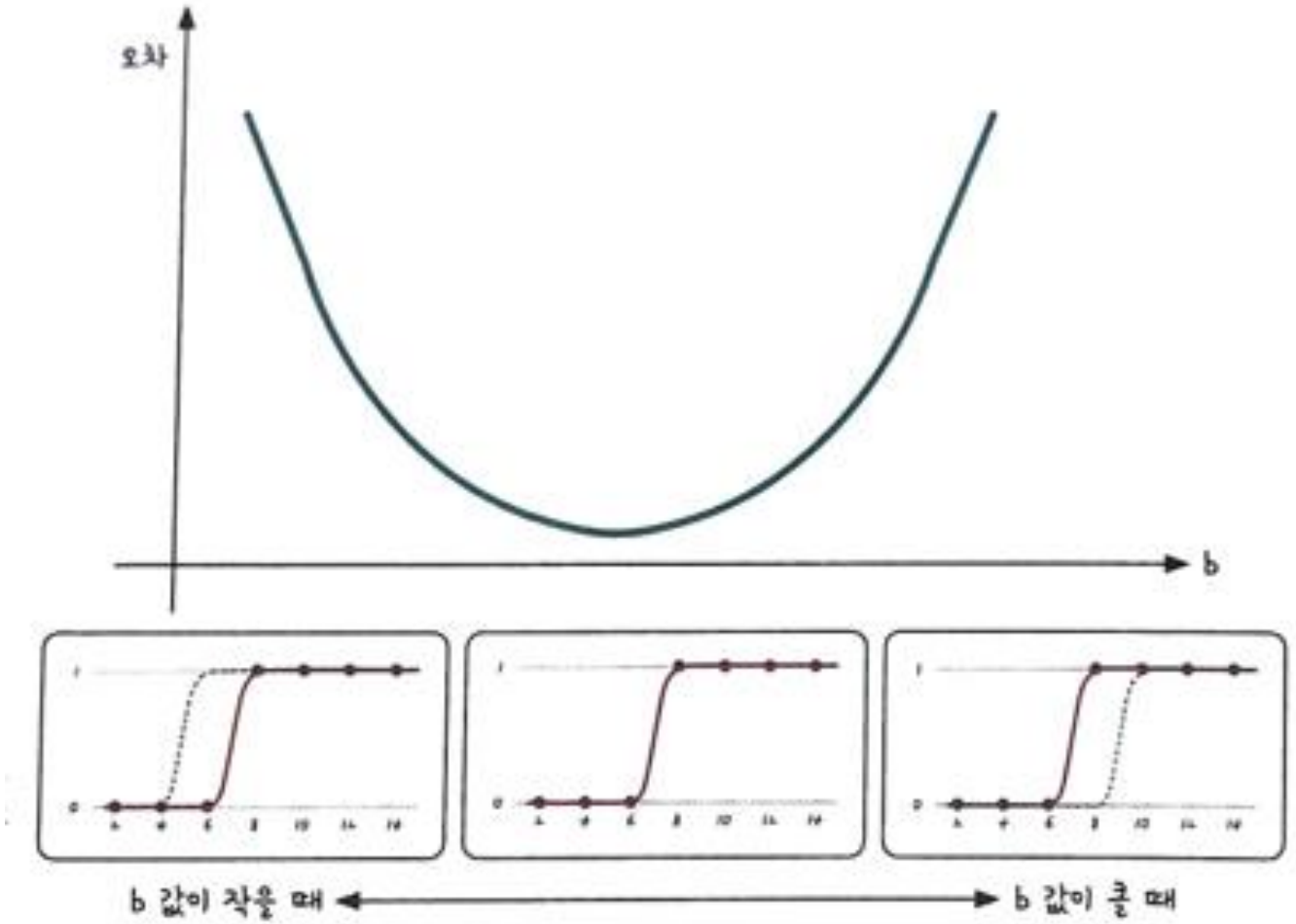
$$ax + b$$



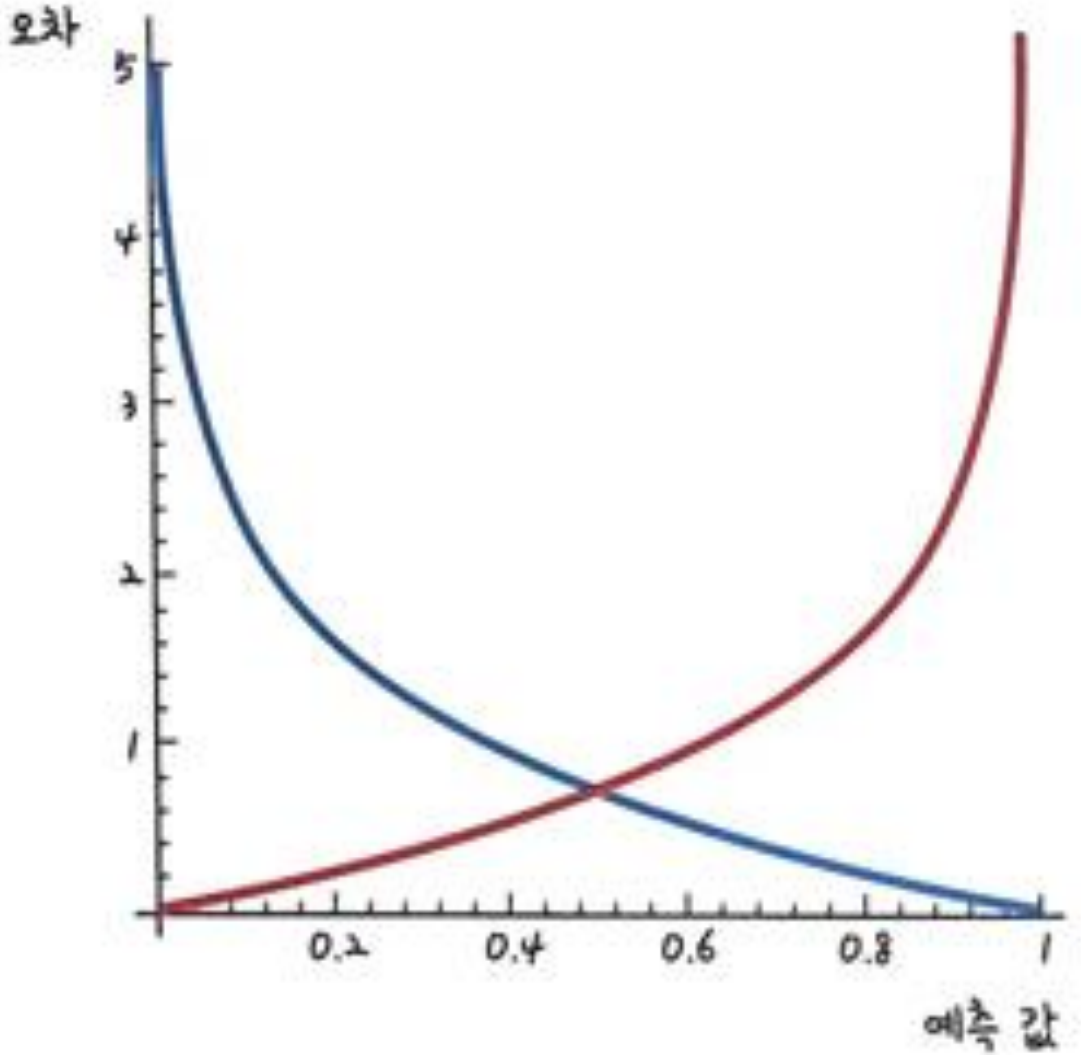


로지스틱 회귀(Logistic Regression) : 시그모이드 함수



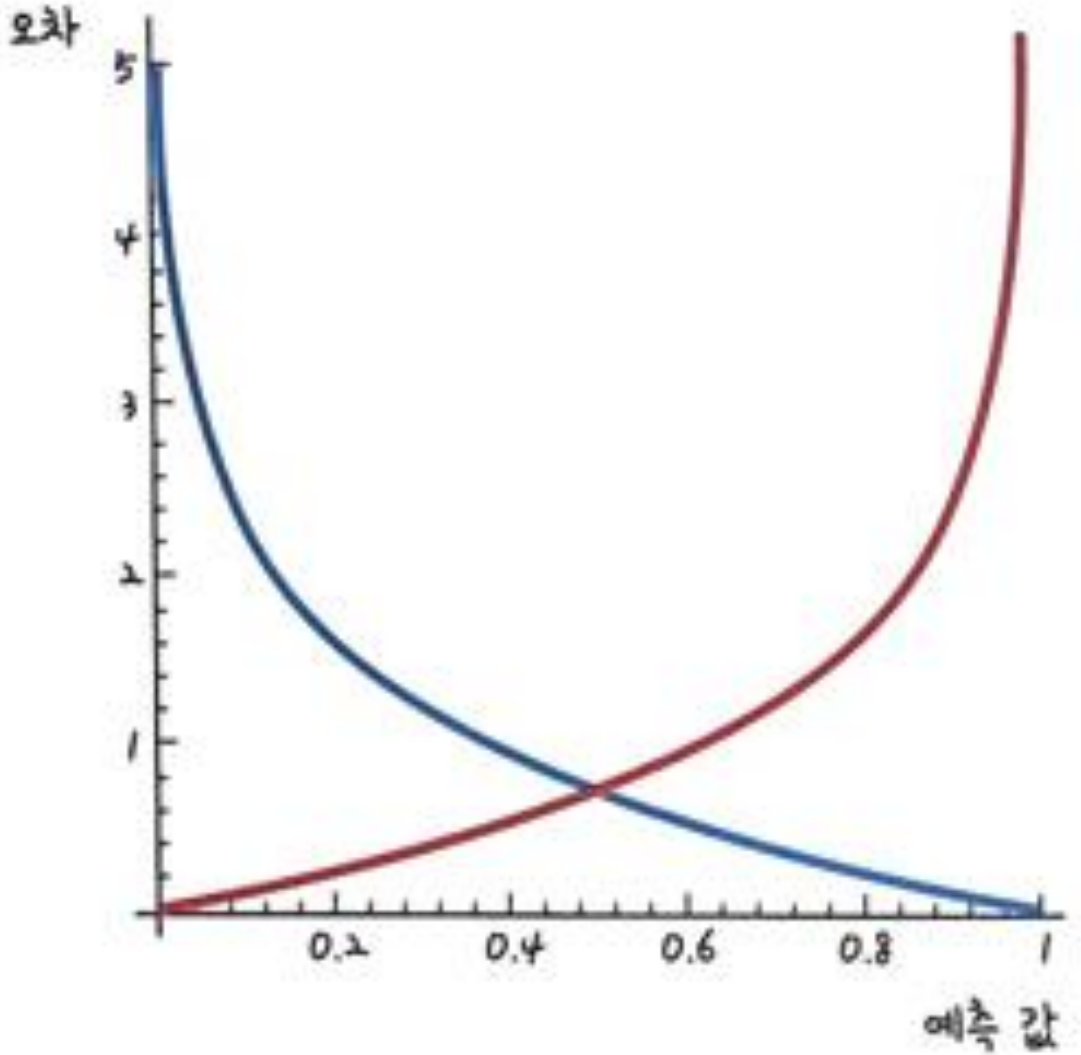


로지스틱 회귀(Logistic Regression) : 오차와 로그함수



$$-\{ylogh + (1-y)\log(1-h)\}$$

로지스틱 회귀(Logistic Regression) : 오차와 로그함수

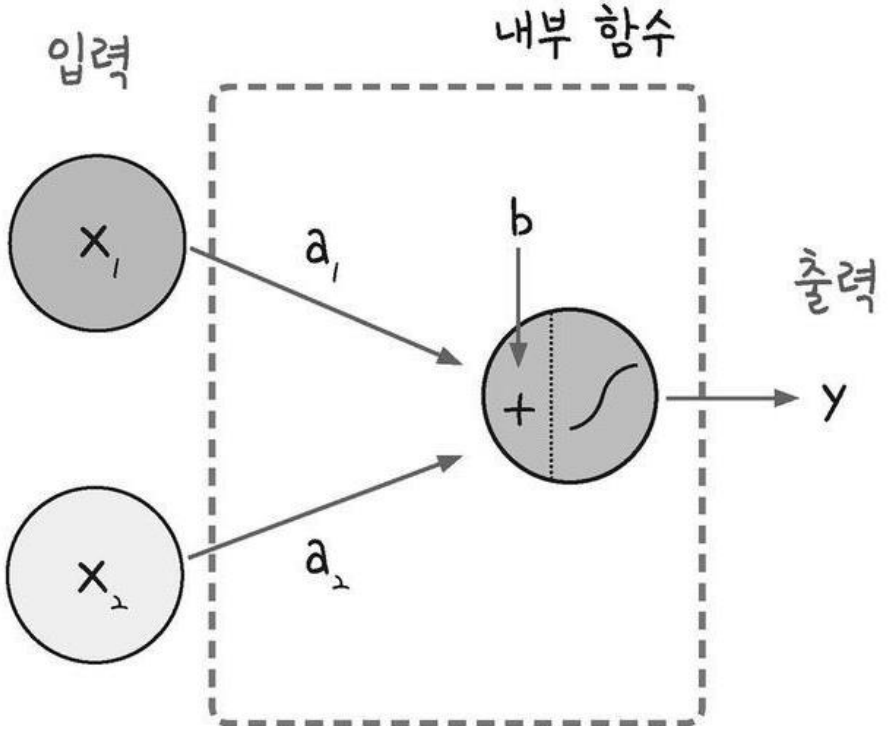


$$-\{ylogh + (1-y)\log(1-h)\}$$

A B

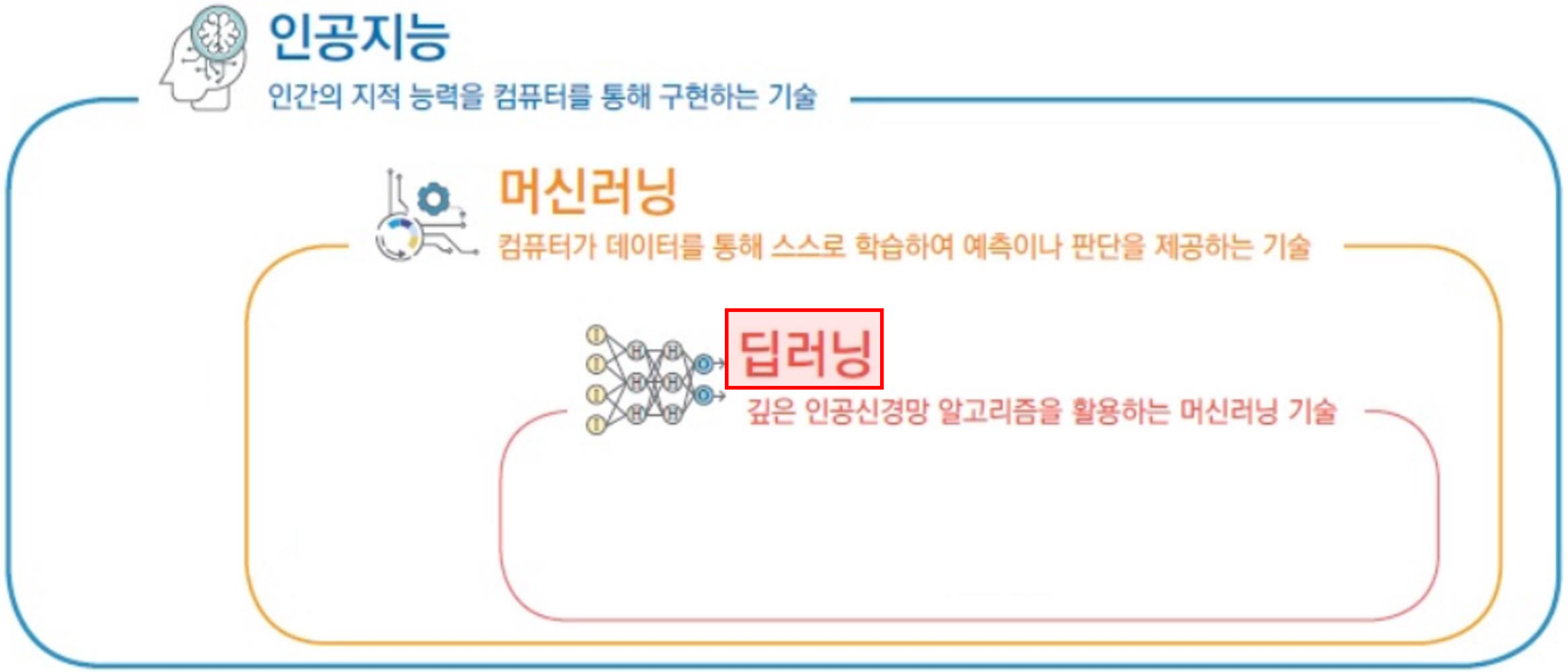
로지스틱 회귀(Logistic Regression) : 다중 입력

$$y = a_1 x_1 + a_2 x_2 + b$$



퍼셉트론

딥러닝

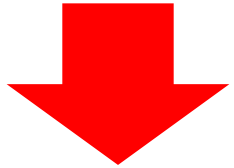


이미지 출처 : https://samstory.coolschool.co.kr/zone/story/modi/streams/76601

퍼셉트론

a는 기울기, b는 절편

$$y = ax + b$$

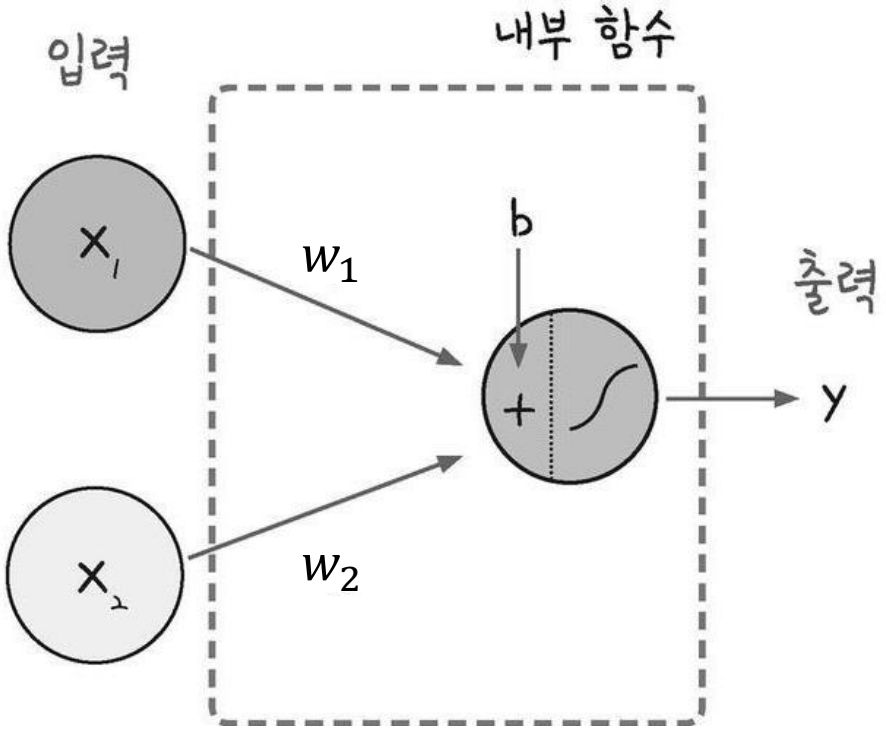


$$y = wx + b$$

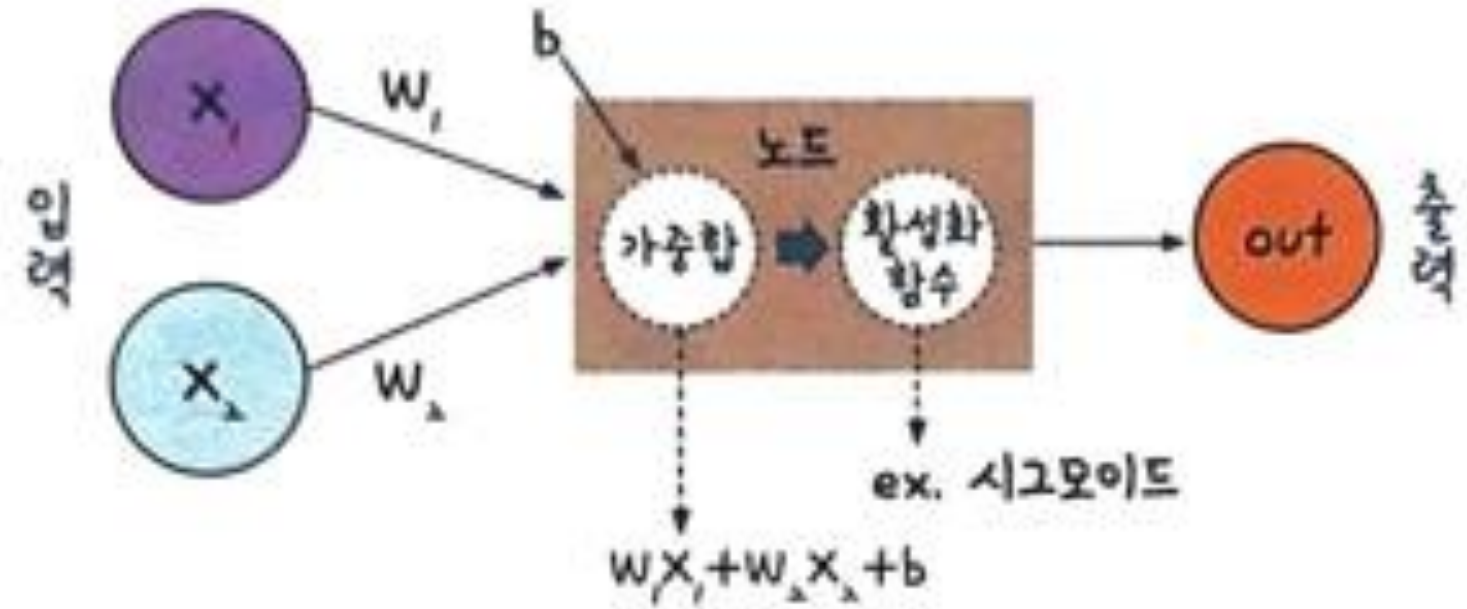
w는 가중치, b는 바이어스

퍼셉트론 , 1956

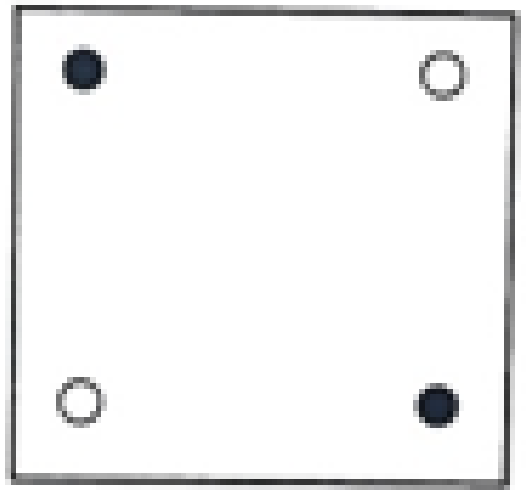
- 가중치, 가중합, 바이어스, 활성화 함수



뉴런, 퍼셉트론, 신경망



퍼셉트론의 문제 : XOR



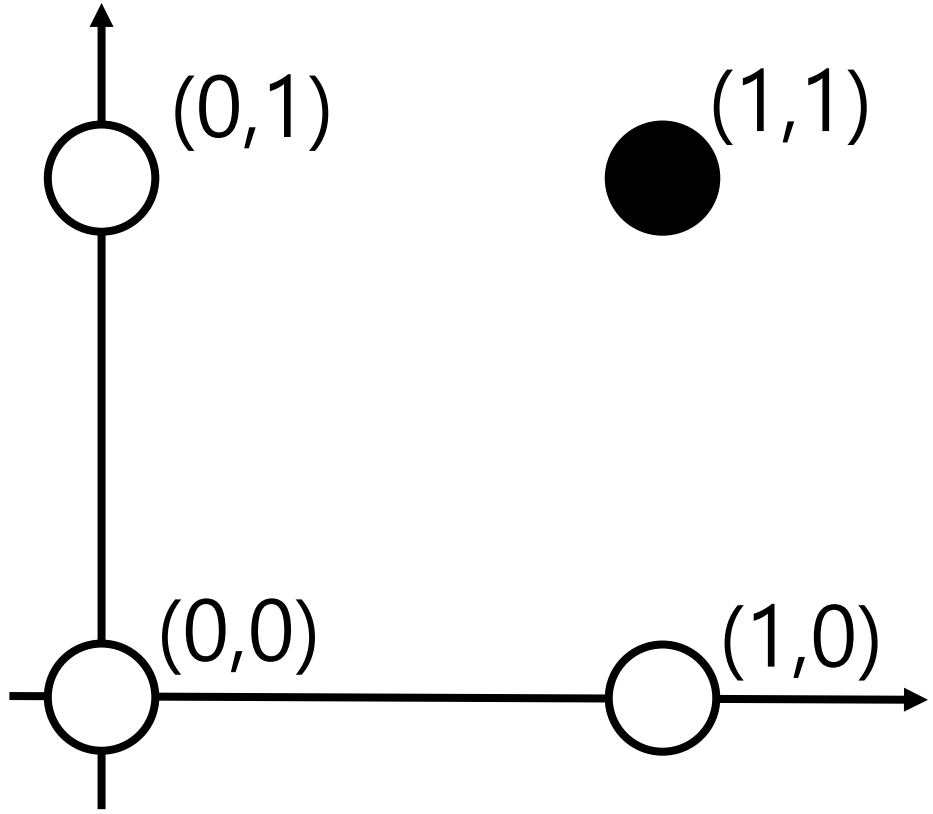
- 네 점 사이에 하나의 직선을 그을 거예요.
- 직선의 한쪽 편에는 검은 점만 있고

  다른 한쪽 편에는 흰 점만 있게 선을 그을 수 있을까요?

퍼셉트론의 문제 : XOR

AND 게이트 

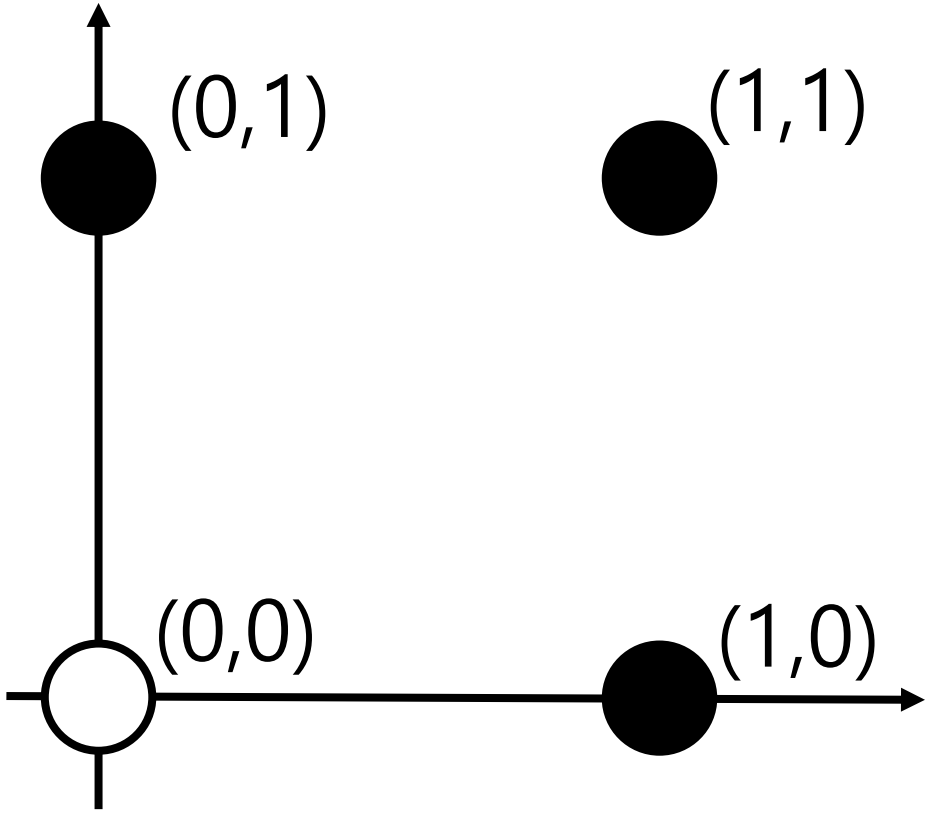
x1	x2	y
0	0	0
0	1	0
1	0	0
1	1	1



퍼셉트론의 문제 : XOR

OR 게이트 

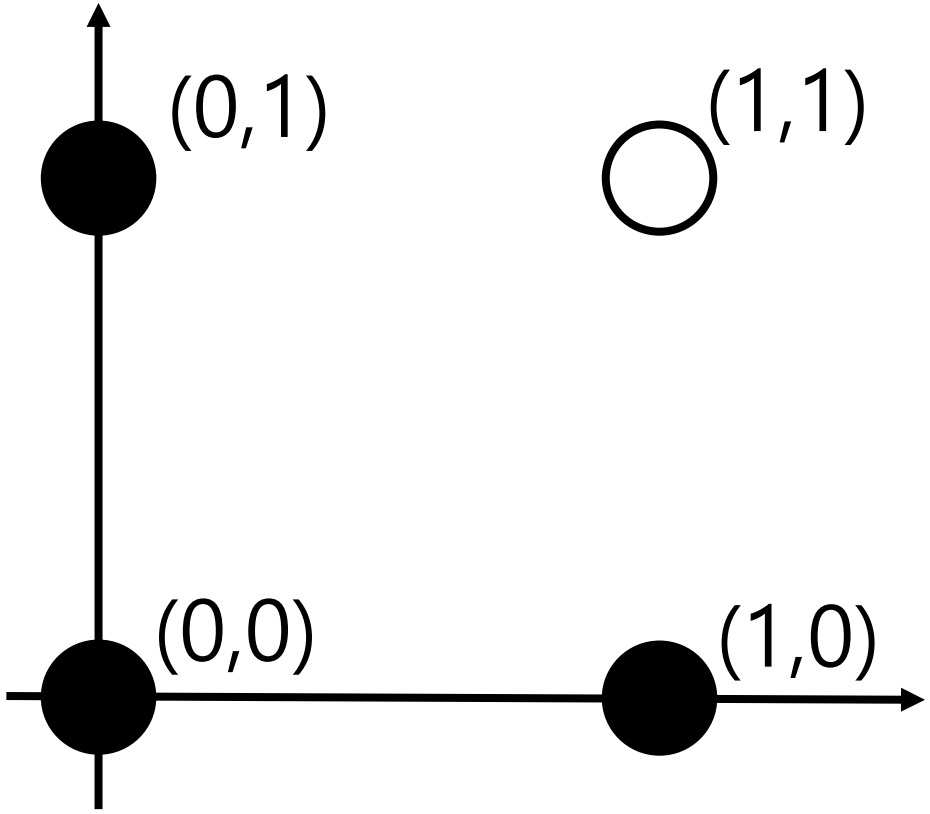
x1	x2	y
0	0	0
0	1	1
1	0	1
1	1	1



퍼셉트론의 문제 : XOR

NAND 게이트 

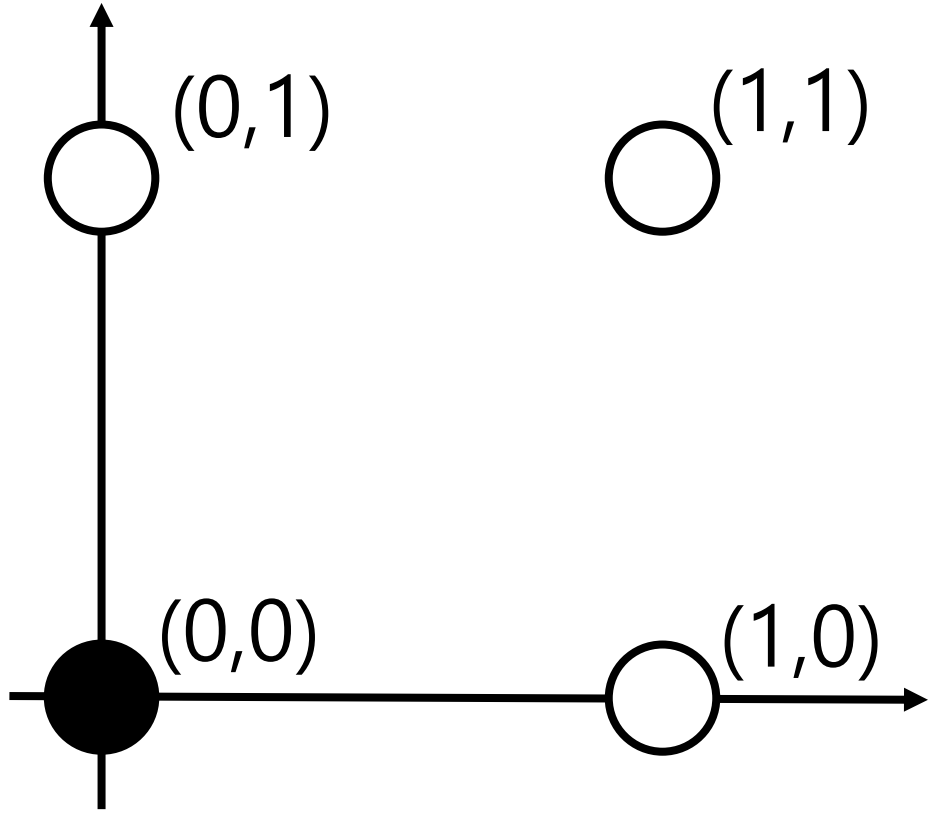
x1	x2	y
0	0	1
0	1	1
1	0	1
1	1	0



퍼셉트론의 문제 : XOR

NOR 게이트 

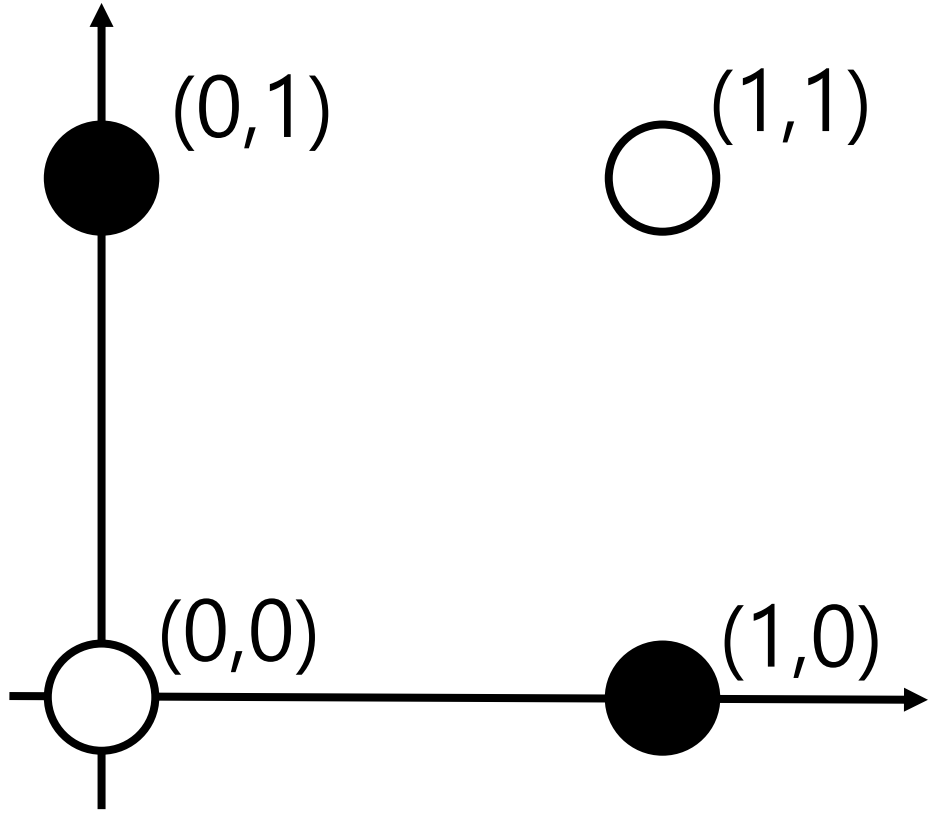
x1	x2	y
0	0	1
0	1	0
1	0	0
1	1	0



퍼셉트론의 문제 : XOR

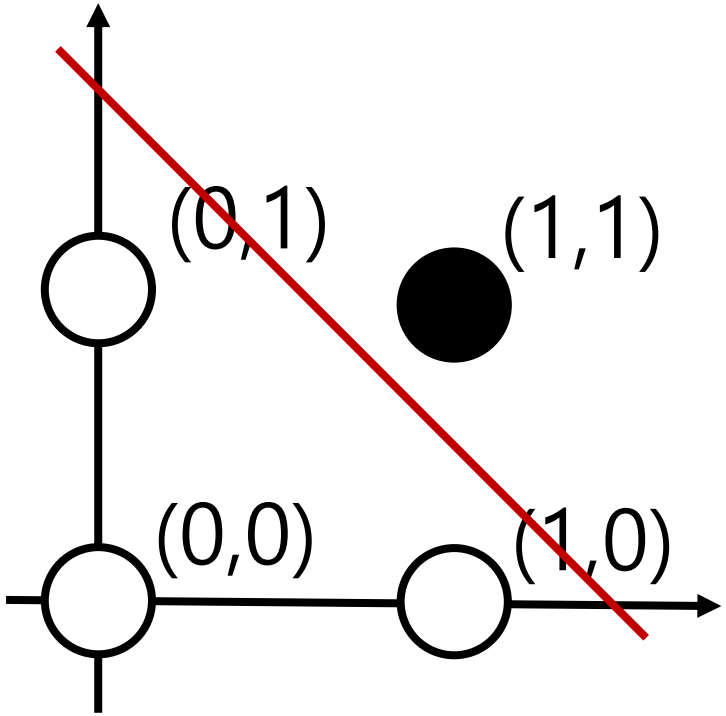
XOR 게이트 

x1	x2	y
0	0	0
0	1	1
1	0	1
1	1	0

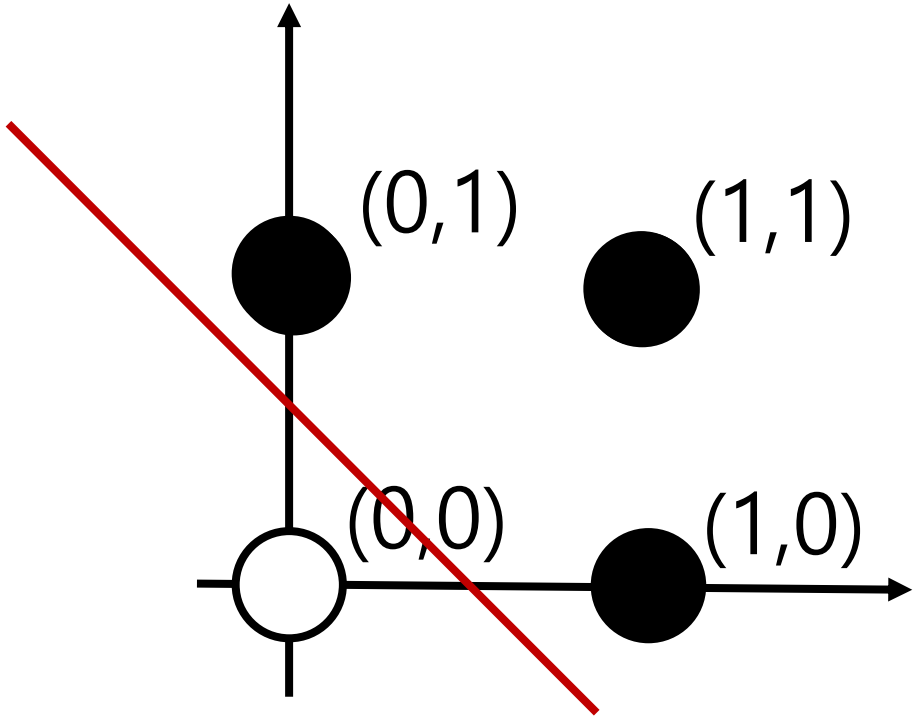


퍼셉트론의 문제 : XOR

AND

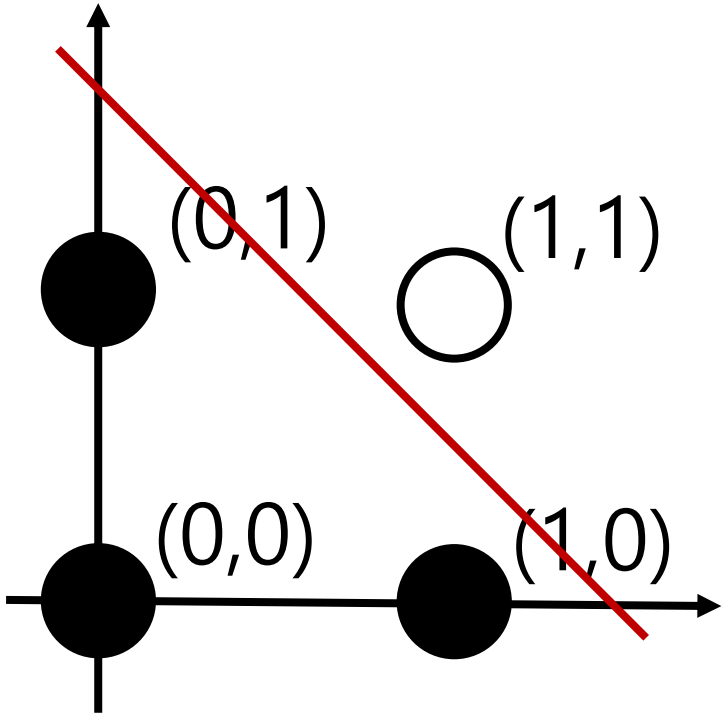


OR

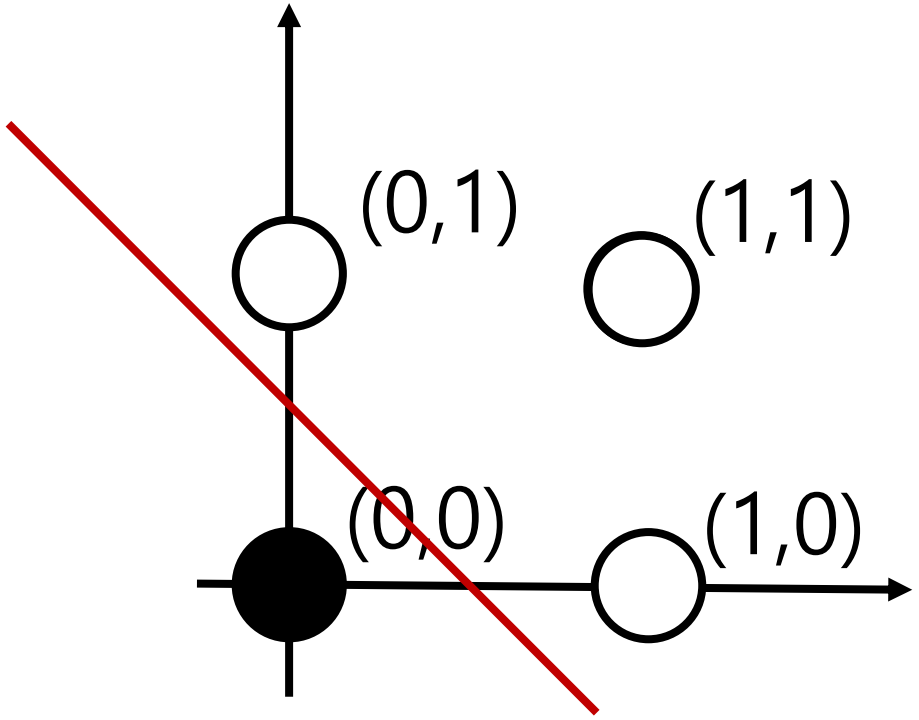


퍼셉트론의 문제 : XOR

NAND



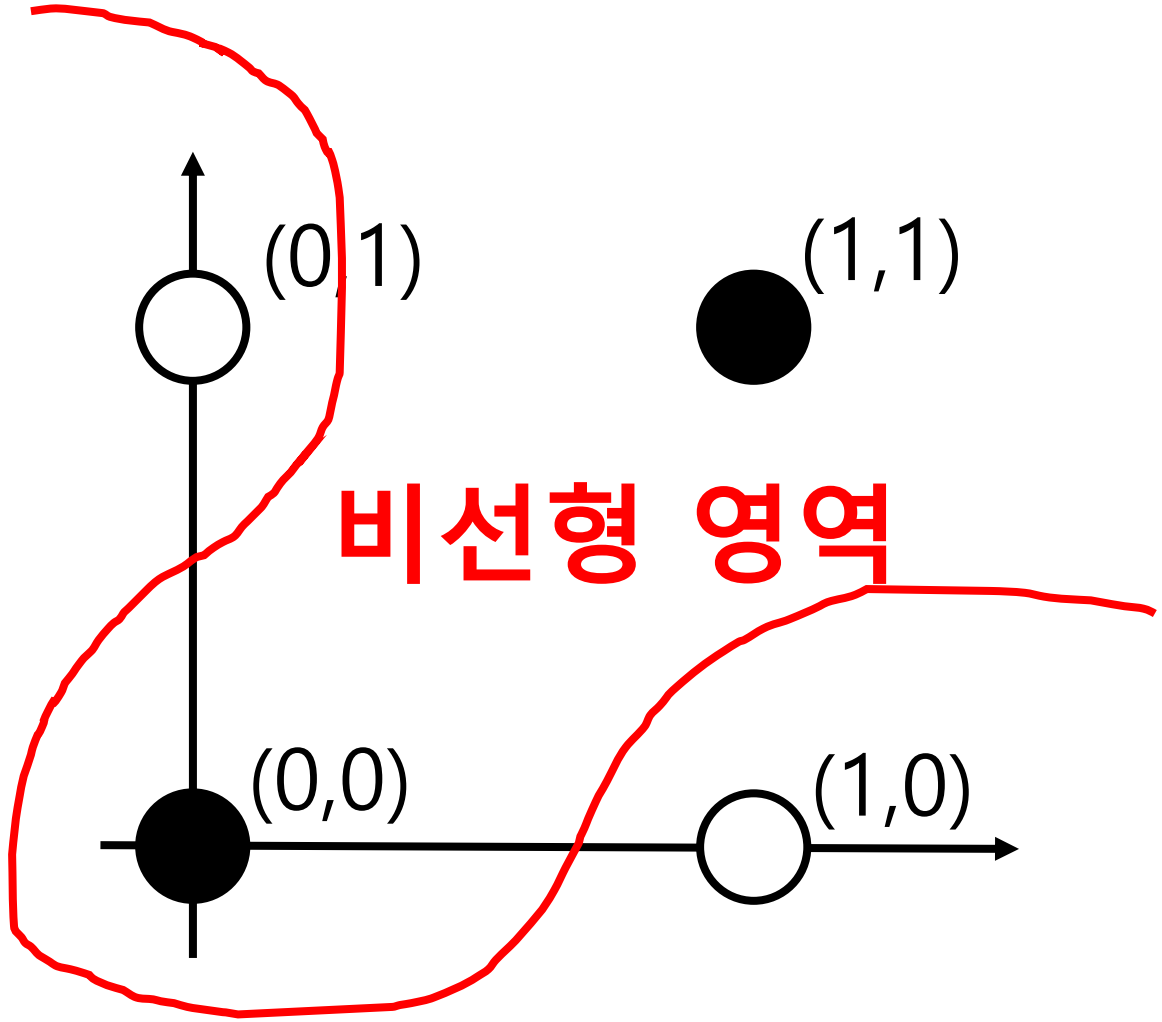
NOR



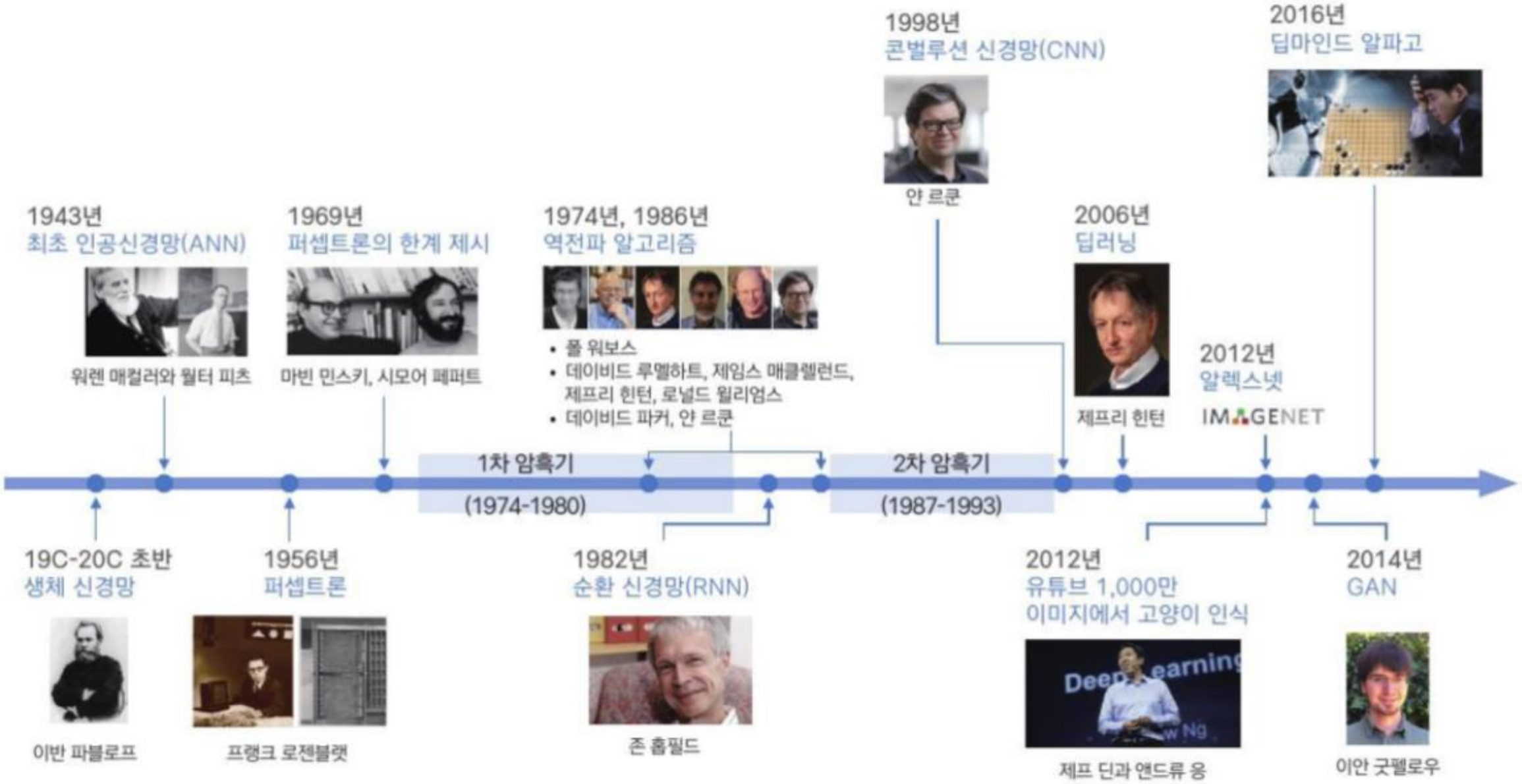
퍼셉트론의 문제 : XOR

XOR 게이트

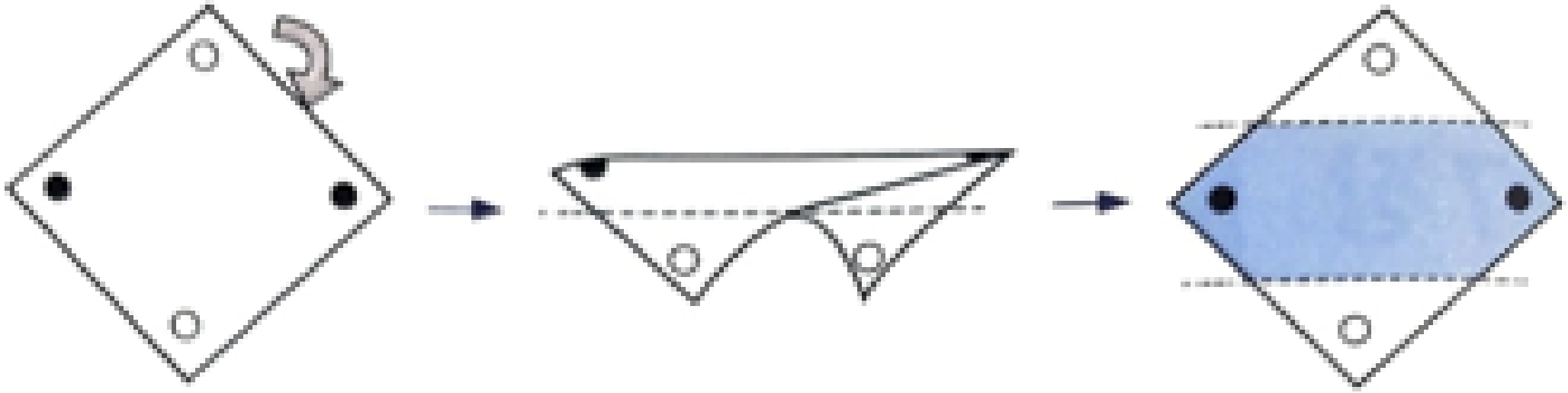
x1	x2	y
0	0	0
1	0	1
0	1	1
1	1	0



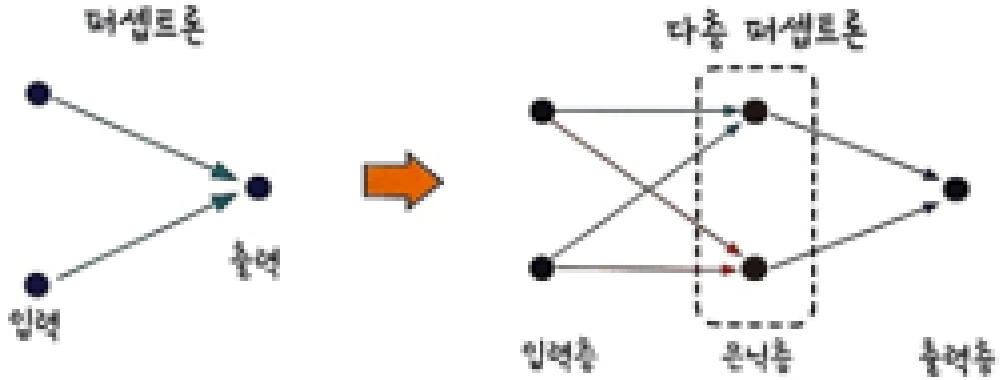
퍼셉트론의 문제 : XOR와 인공지능의 암흑기



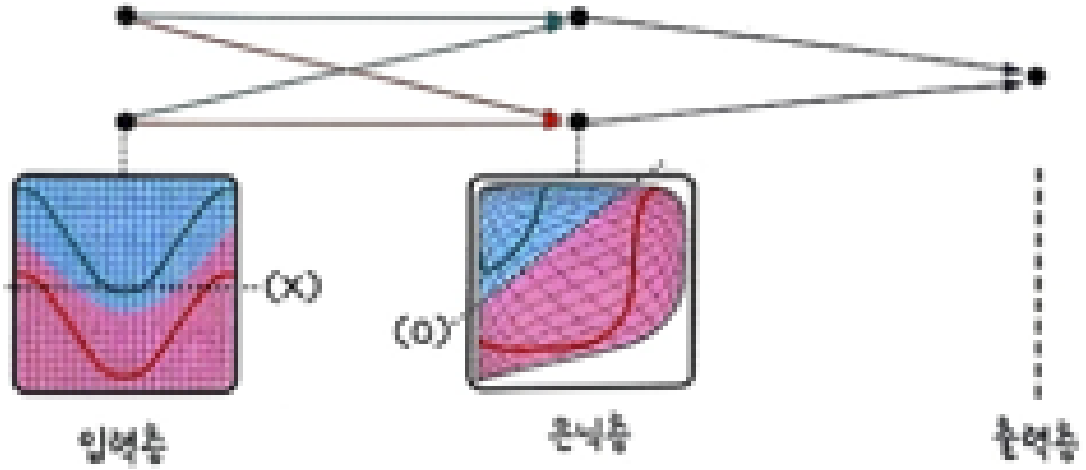
퍼셉트론 XOR 문제의 해결



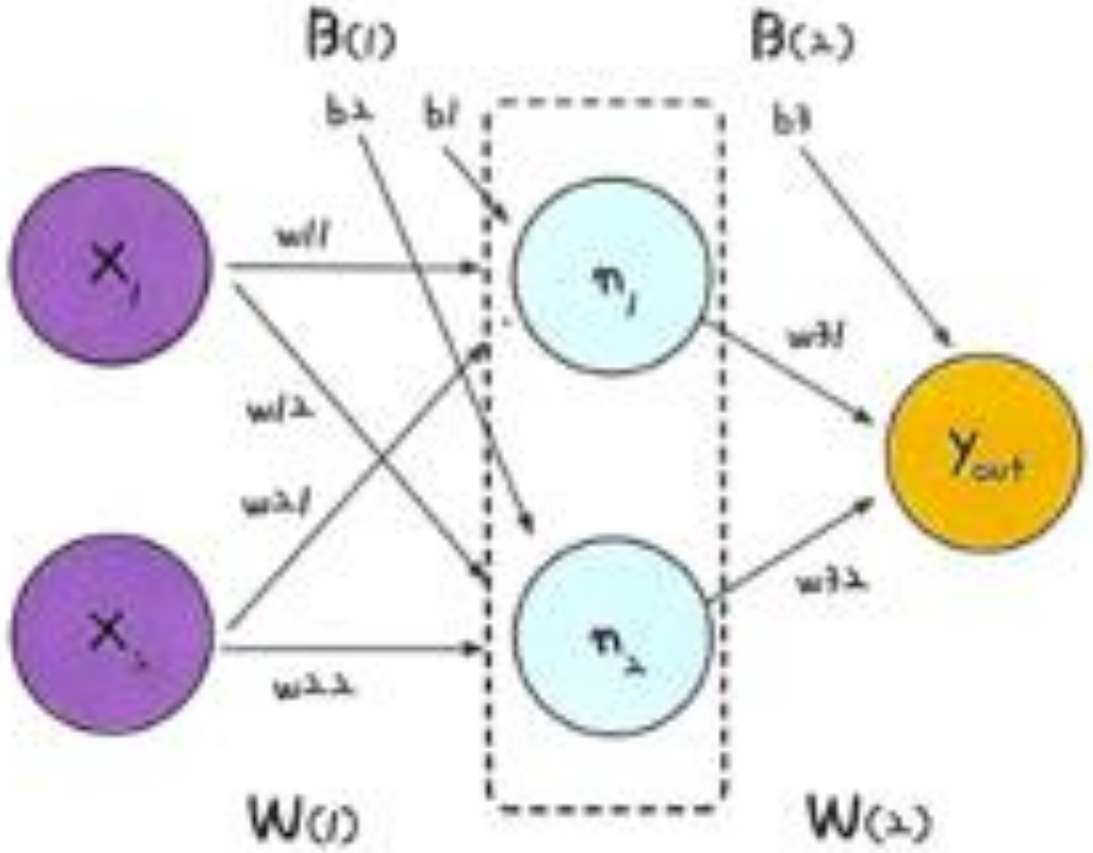
퍼셉트론 XOR 문제의 해결 : 은닉층(Hidden Layer)



우리는,
파란색과 빨간색을 구분하는 선을 그을 거예요!



다층 퍼셉트론

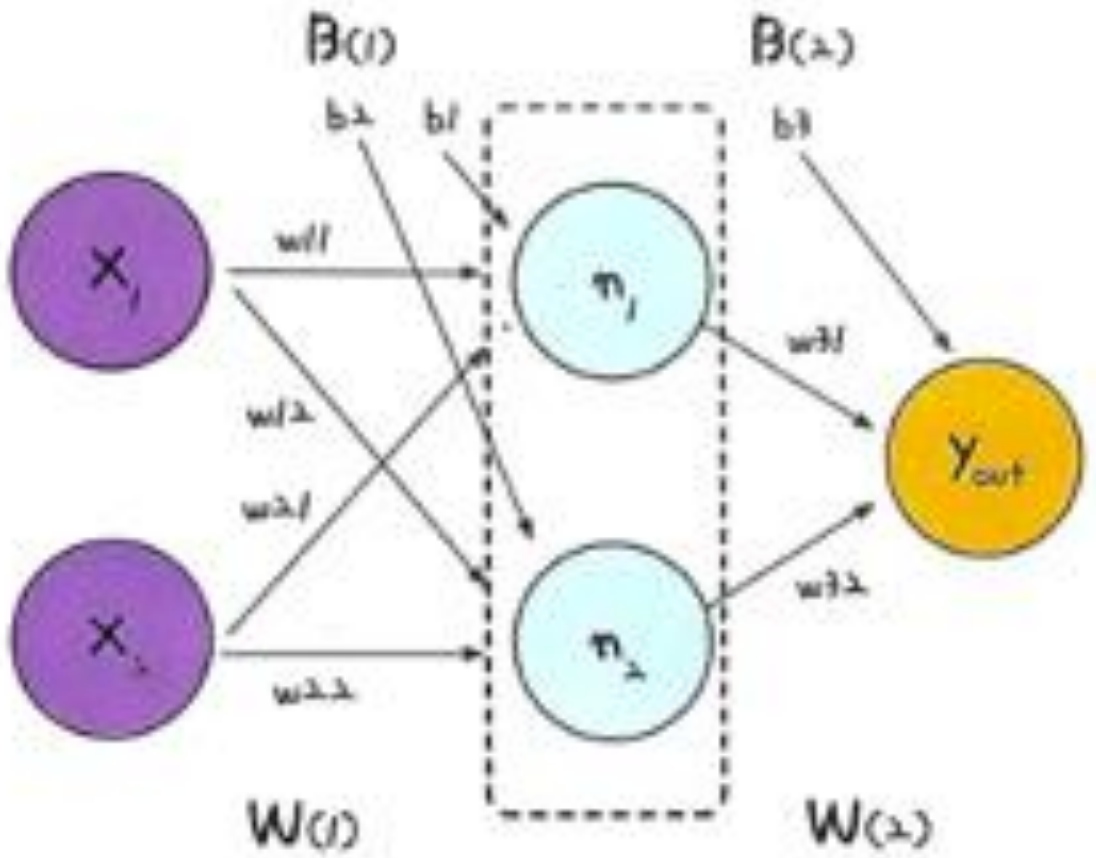


$$n_1 = \partial(x_1 w_{11} + x_2 w_{21} + b_1)$$

$$n_2 = \partial(x_1 w_{12} + x_2 w_{22} + b_2)$$

$$y_{out} = \partial(n_1 w_{31} + n_2 w_{32} + b_3)$$

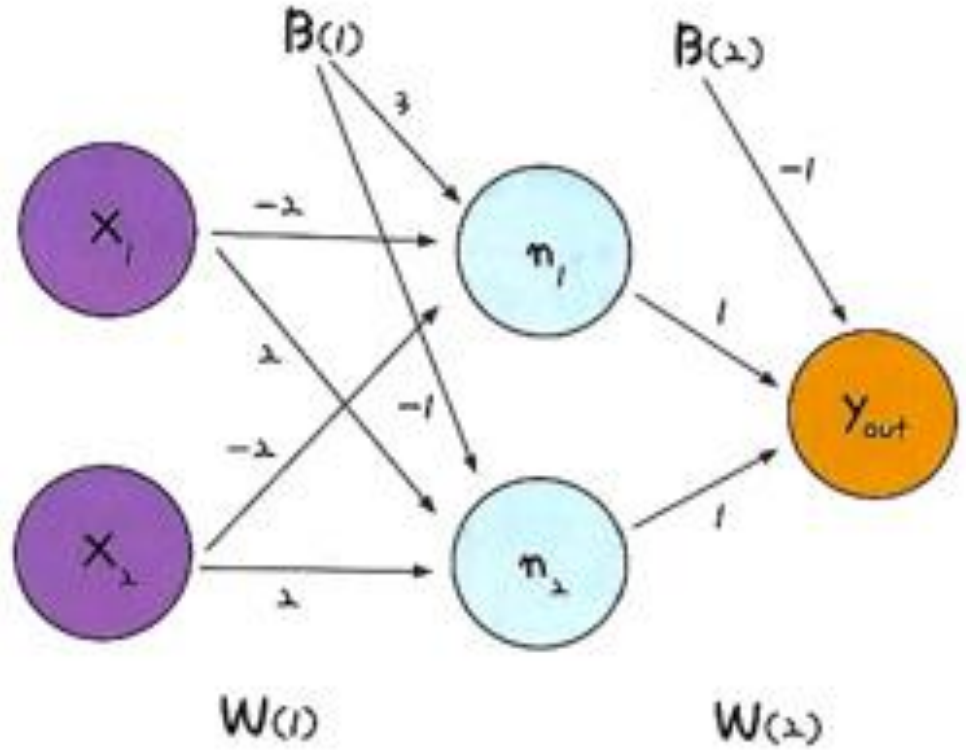
다층 퍼셉트론



$$W(1)=\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \qquad B(1)=\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$W(2)=\begin{bmatrix} w_{31} \\ w_{32} \end{bmatrix} \qquad B(2)=[b_3]$$

다층 퍼셉트론 : XOR 문제의 해결



$$W(1)=\begin{bmatrix}-2 & 2\\ -2 & 2\end{bmatrix} \qquad B(1)=\begin{bmatrix}3\\ 1\end{bmatrix}$$

$$W(2)=\begin{bmatrix}1\\ 1\end{bmatrix} \qquad B(2)=[-1]$$

다층 퍼셉트론 : XOR 문제의 해결

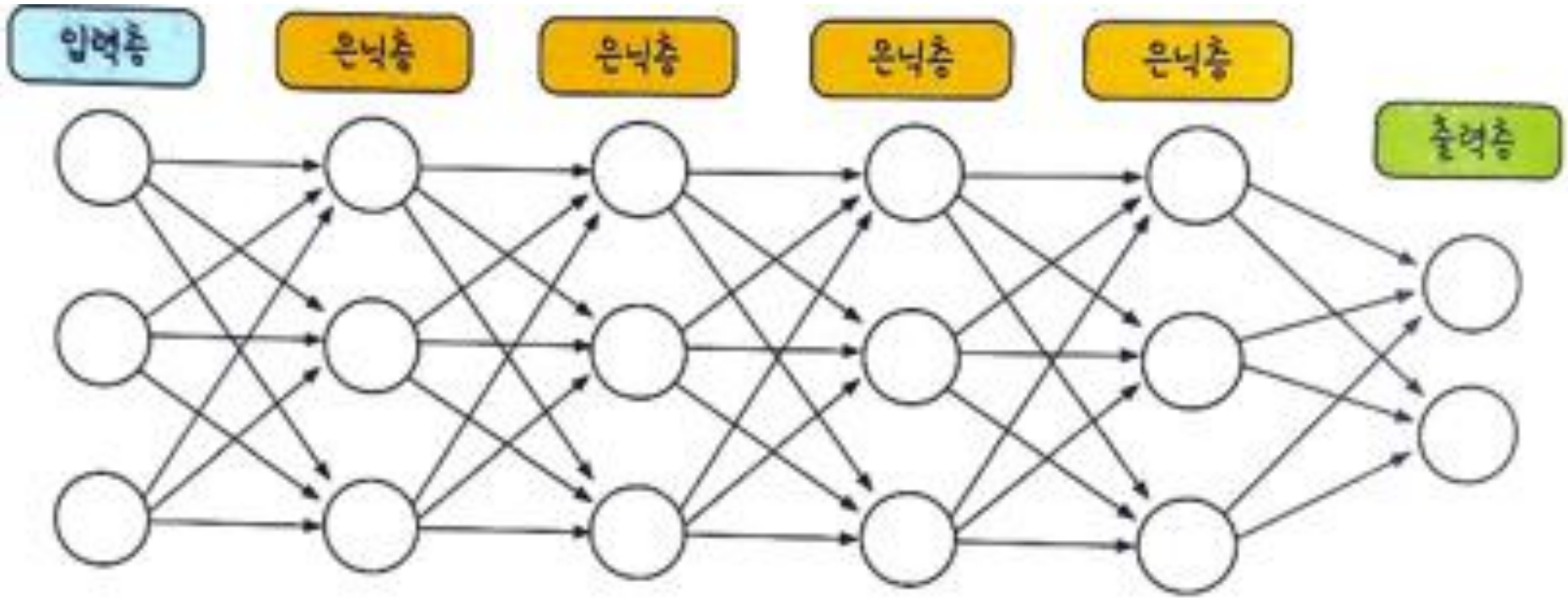
$$n_1 = \partial(x_1 w_{11} + x_2 w_{21} + b_1)$$

$$n_2 = \partial(x_1 w_{12} + x_2 w_{22} + b_2)$$

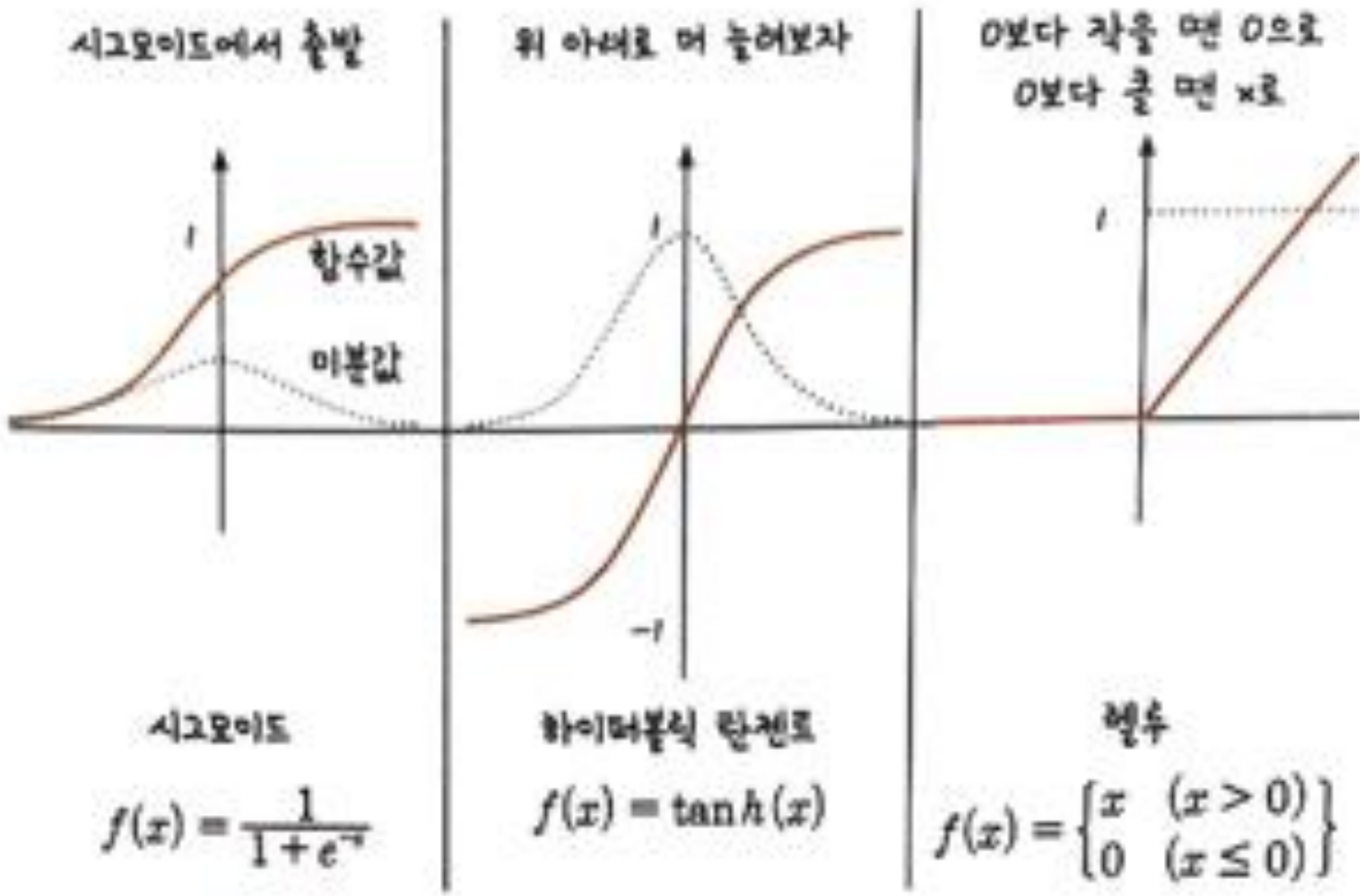
$$y_{out} = \partial(n_1 w_{31} + n_2 w_{32} + b_3)$$

x_1	x_2	n_1	n_2	y_{out}	우리가 원하는 값
0	0	$\sigma(0*(-2)+0*(-2)+3)=1$	$\sigma(0*2+0*2-1)=0$	$\sigma(1*1+0*1-1)=0$	0
0	1	$\sigma(0*(-2)+1*(-2)+3)=1$	$\sigma(0*2+1*2-1)=1$	$\sigma(1*1+1*1-1)=1$	1
1	0	$\sigma(1*(-2)+0*(-2)+3)=1$	$\sigma(1*2+0*2-1)=1$	$\sigma(1*1+1*1-1)=1$	1
1	1	$\sigma(1*(-2)+1*(-2)+3)=0$	$\sigma(1*2+1*2-1)=1$	$\sigma(0*1+1*1-1)=0$	0

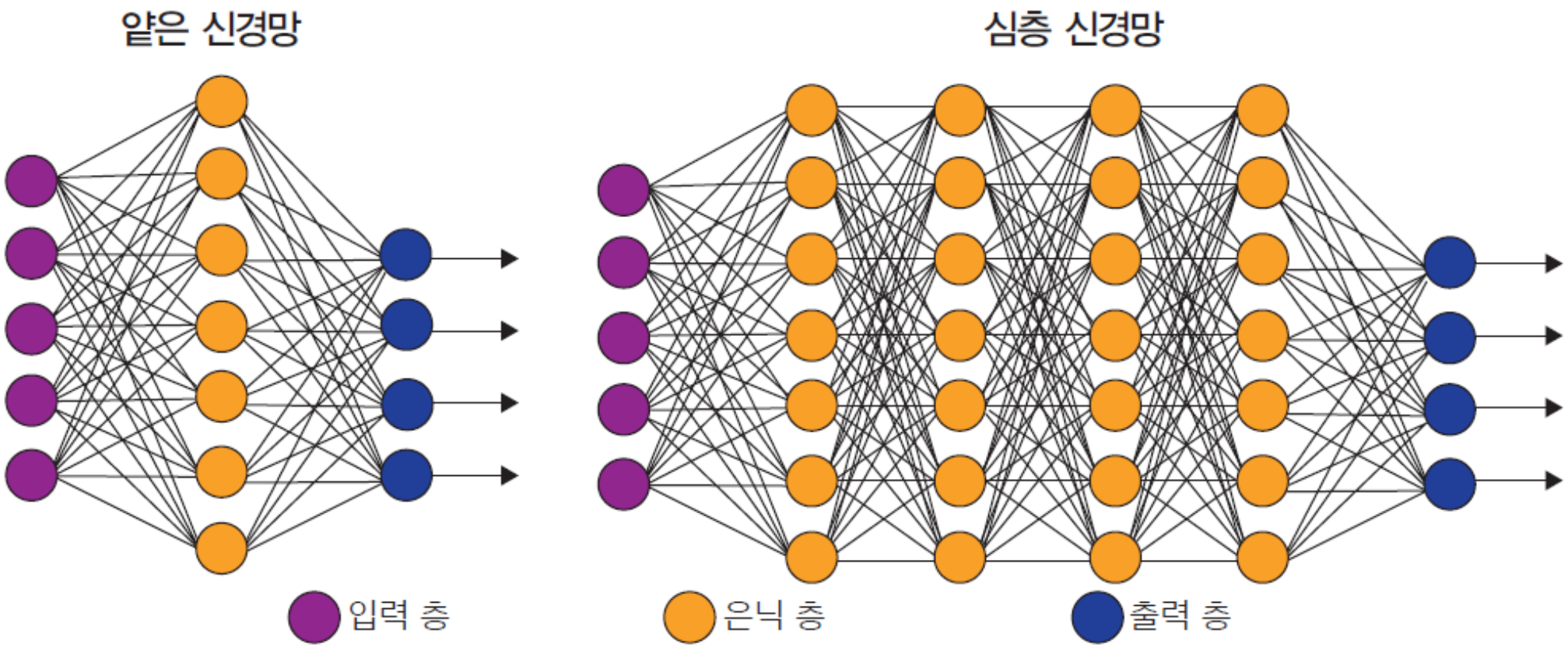
신경망(Neural Network)



신경망(Neural Network) : 기울기 소실 문제와 활성화 함수

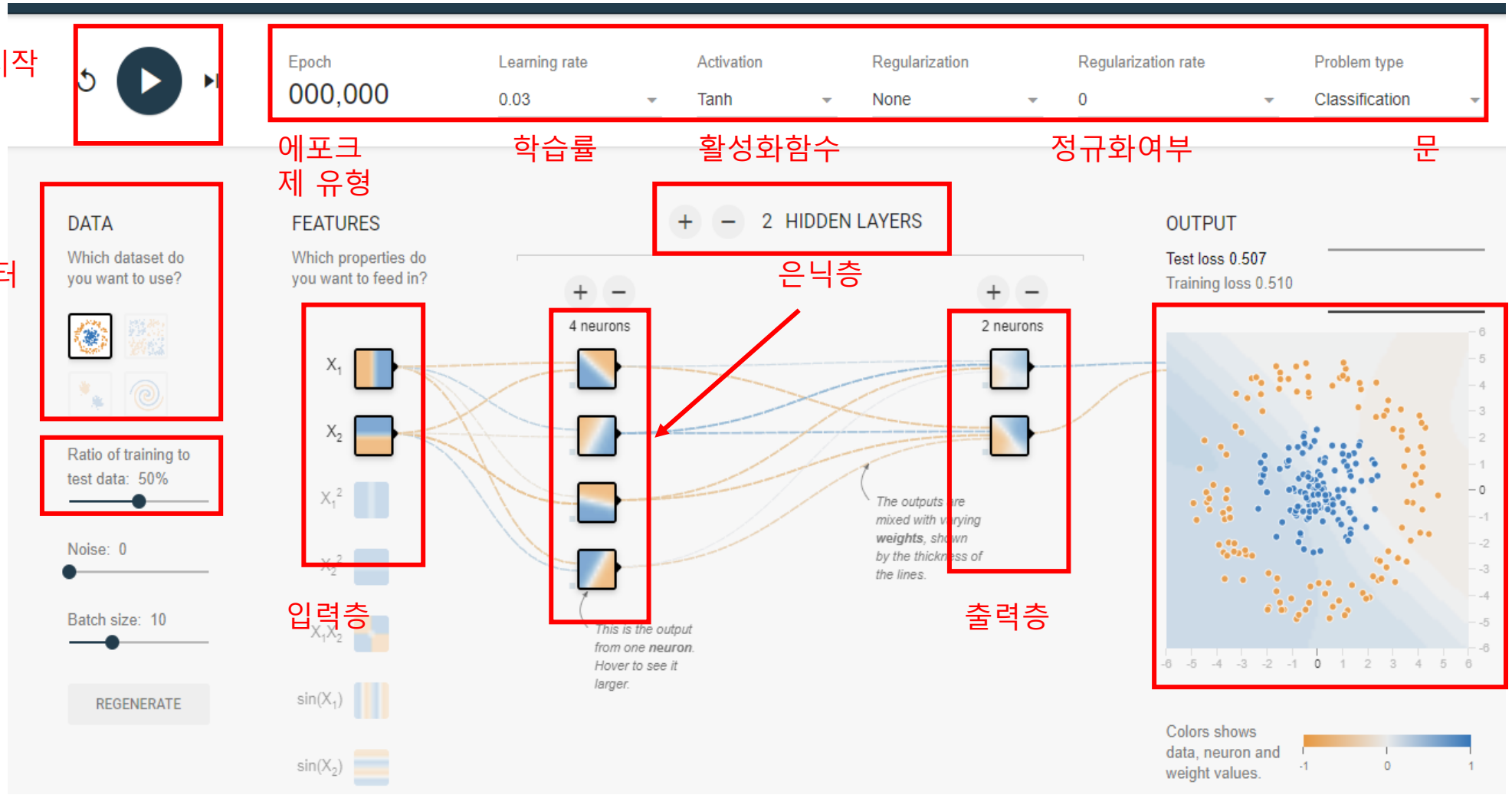


딥러닝(Deep Learning)

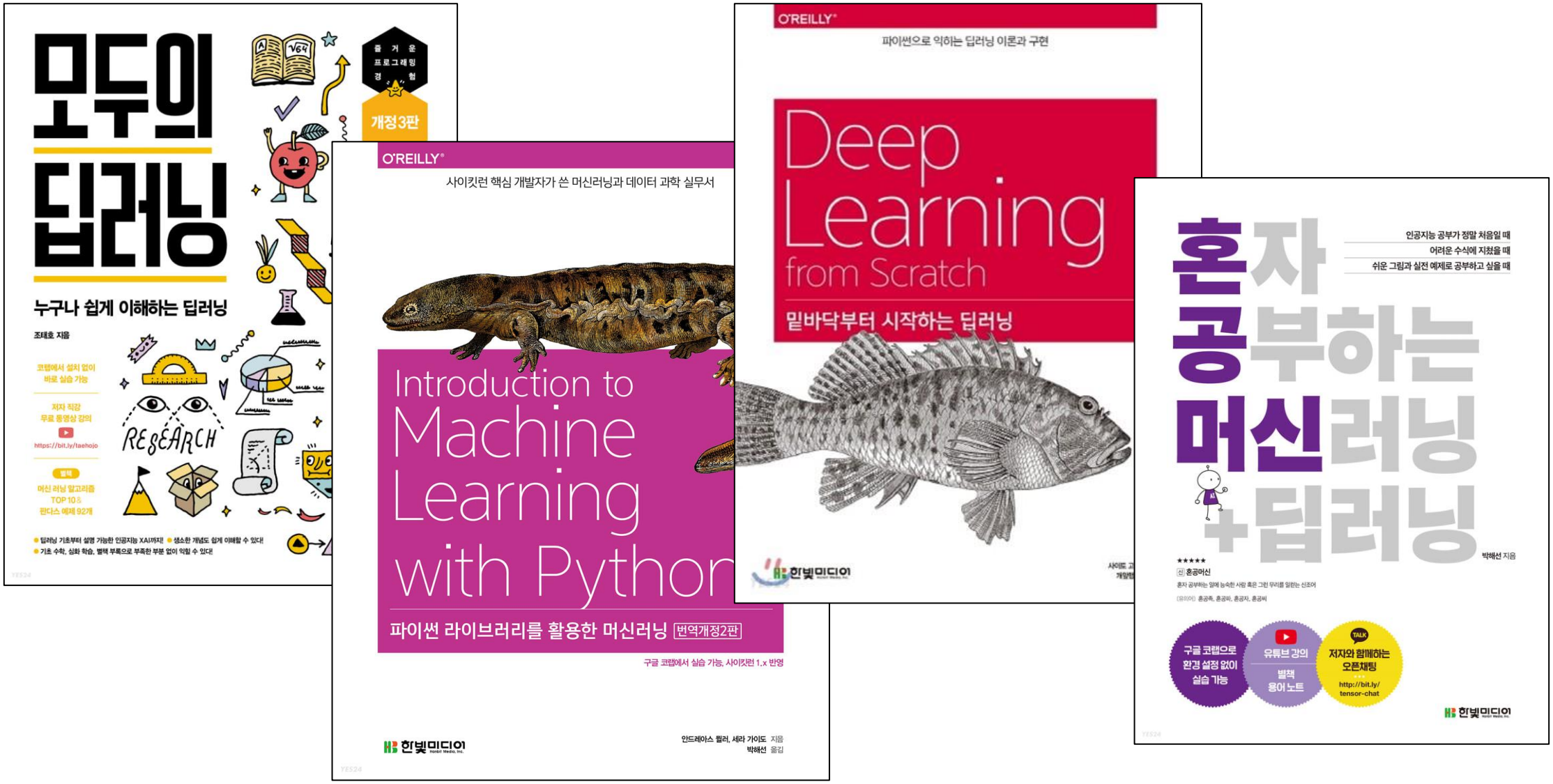


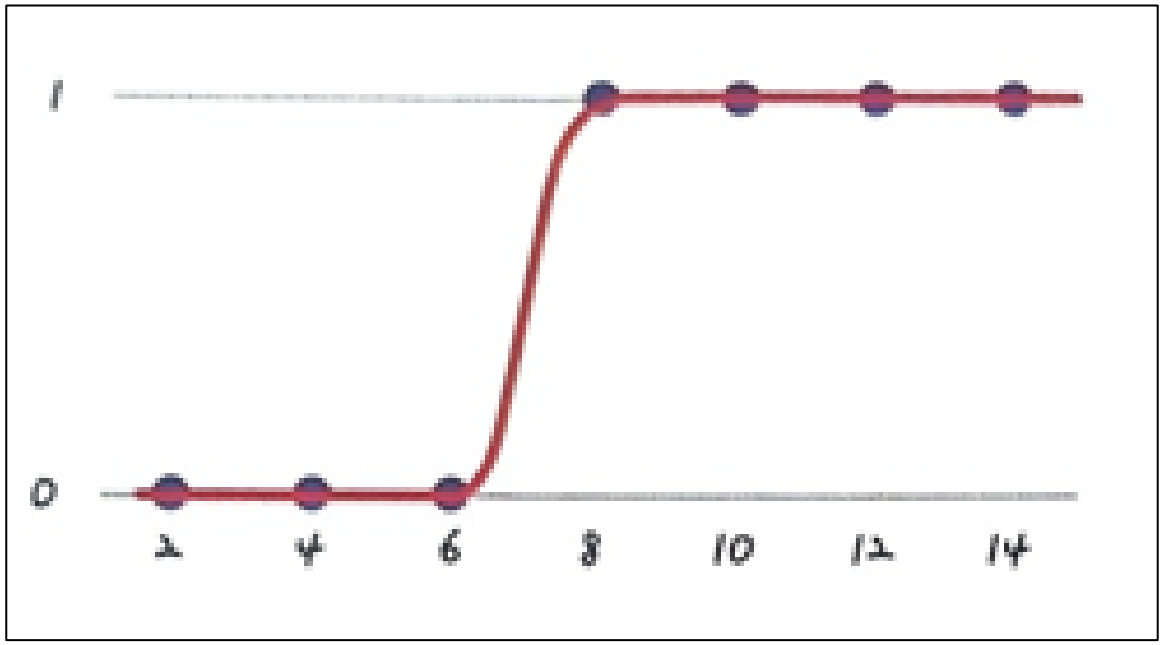
딥러닝(Deep Learning)

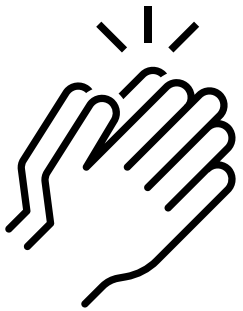
- 구글 플레이그라운드 (https://playground.tensorflow.org/)



어디서부터 어디까지 공부해야 할 지 모를 만큼 방대하지만,







이미, 우리는 1의 상태입니다 :)