

ADVANCED TEXT MINING

by FINGEREDMAN (fingeredman@gmail.com)

APPENDIX 02

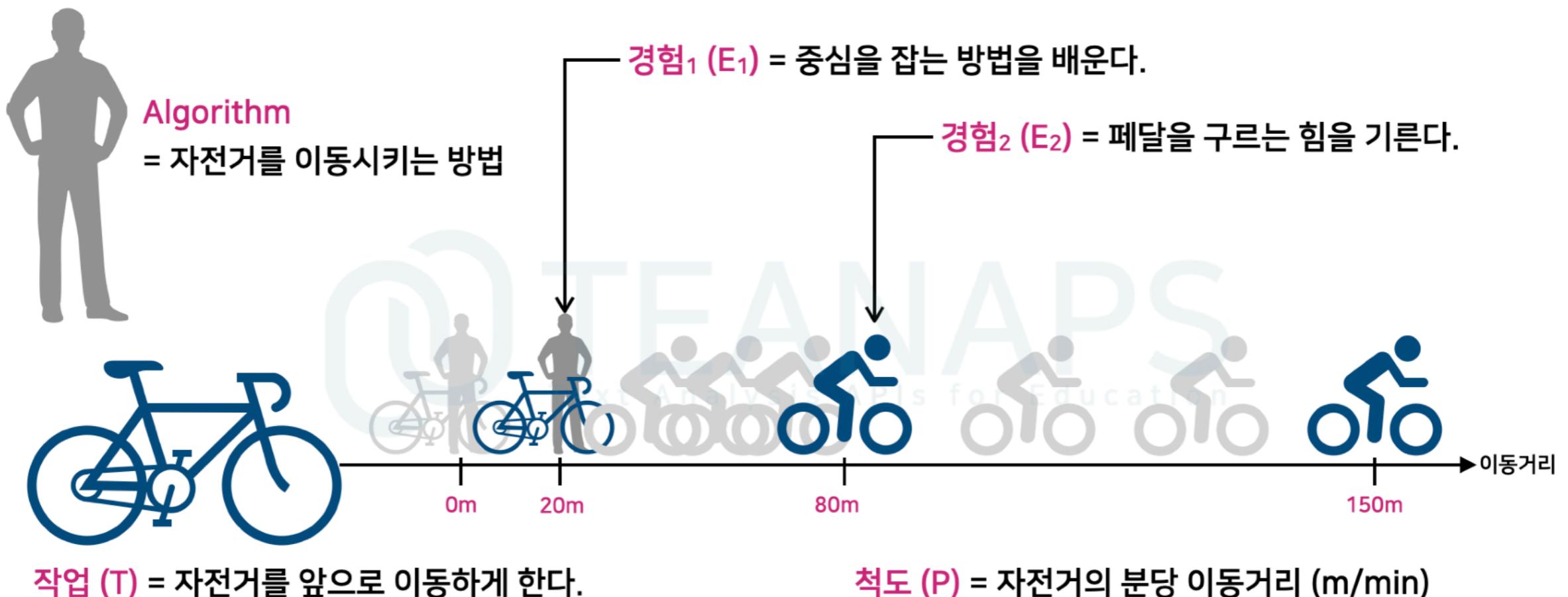
Basic Concepts in Machine Learning

기계학습

(Machine Learning)

Mitchell의 정의

- "A computer **program** is said to learn from **experience E** with respect to some class of **tasks T** and performance **measure P**, if its performance at tasks in T, as measured by P, improves with experience E"
- "컴퓨터 알고리즘(프로그램)이 작업 T를 수행하고 이 알고리즘의 성능을 척도 P_(performance)로 평가할 수 있다면, 경험 E_(experience)를 통해 P가 개선되는 경우 이 알고리즘은 학습이 되었다고 볼 수 있다."





$$y = ax_1 + bx_2$$

기계학습

(Machine Learning)

기계학습의 기본개념

- 외부 환경이 사람을 지도(supervise)하는 것과 같이, 기계가 기존에 할 수 없던 것을 가능하도록 하게하는 과정
- 사람과 기계 모두 외부 지도에 따라 매우는 학습단계와 실제 성능을 평가하는 테스트 과정을 통해 학습함
- 기계학습의 유형

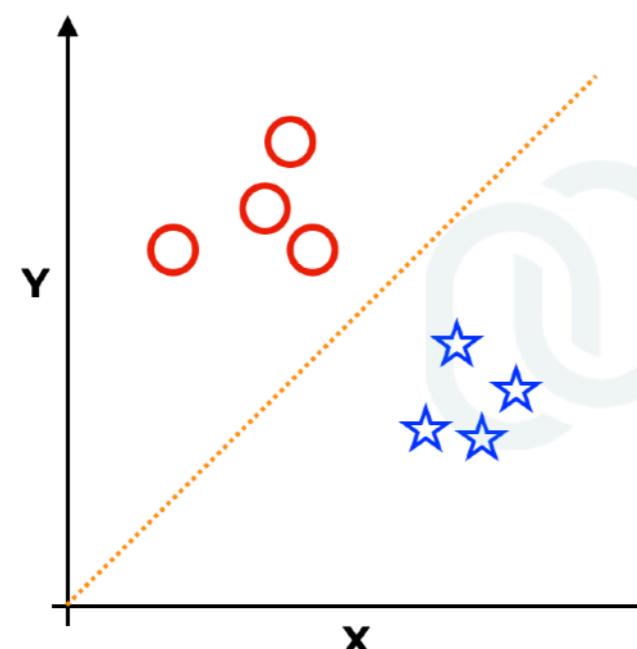
1) 지도학습 (supervised learning) :

입력과 출력을 가지는 데이터로부터 패턴을 추출하여 새로운 입력에 대한 출력을 결정하는 학습방법

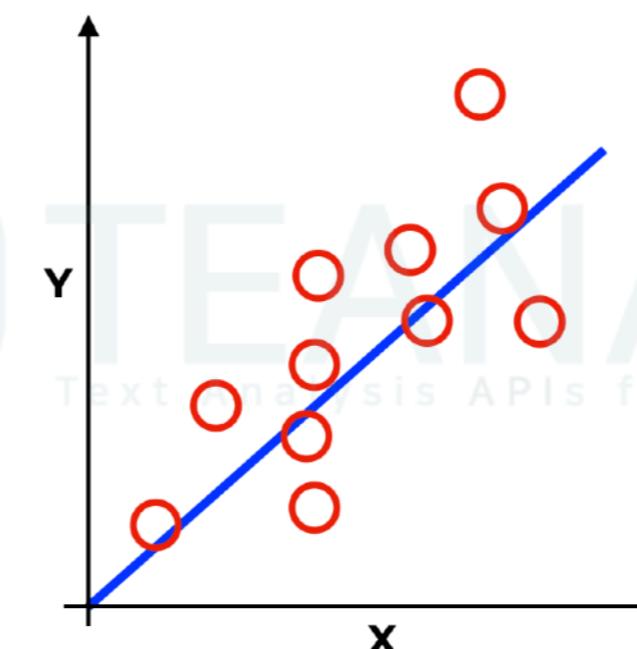
2) 비지도학습 (unsupervised learning) :

출력에 대한 정의가 없는 데이터로부터 의미있는 패턴을 추출하는 학습방법

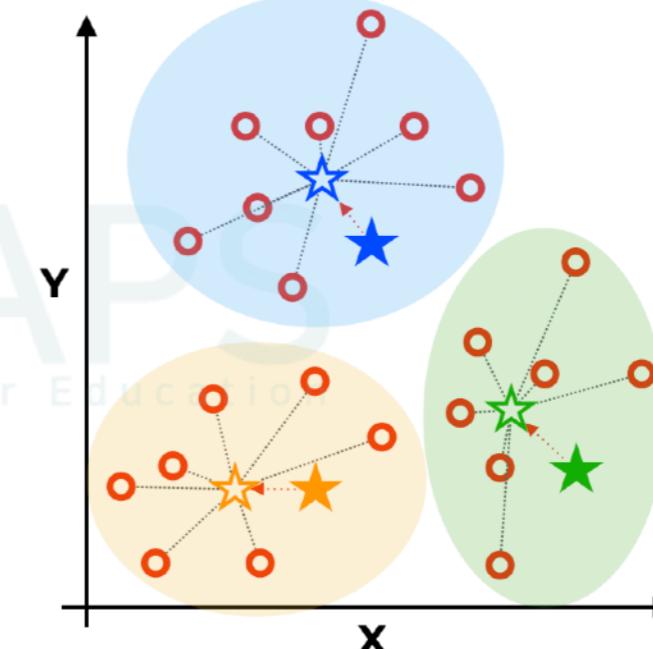
분류 (Classification)



회귀분석 (Regression)



군집화 (Clustering)



기계학습 절차: 데이터 준비

자질추출 (Feature Extraction)

- 기계학습에 필요한 자질(변수, feature)을 추출하고 이를 수치로 표현하는 방법
- 과거 기계학습의 성능은 동일한 알고리즘에 대해 특징을 추출하는 방법에 의해 좌우되었으나, 딥러닝(deep learning)의 등장으로 자질을 추출하는 전처리 과정이 자동화되어 거의 사라짐

| 구분 | 메시지 | 특징 (Feature) | | | | | |
|----|--|--------------|--------|---------|---------|---------|-----------|
| | | 메시지 길이 | URL 여부 | 특수문자 개수 | 해외수신 여부 | 의심단어 개수 | 광고문자 표시여부 |
| 1 | [국제발신] 하루 30만 달 1천만 만원으로 이렇게! http://bit.ly/3f~ | 40 | 1 | 6 | 1 | 1 | 0 |
| 2 | 팀장님 이보람 선임입니다. 출근하시면 결재 부탁드립니다. | 20 | 0 | 1 | 0 | 0 | 0 |
| 3 | (광고)웰컴박하라 vc⑤47③.co 코드 wc1004 무료수신거부 01084510000 | 45 | 1 | 5 | 0 | 1 | 1 |
| 4 | (광고)신한과 함께하는 소중한 미래 따뜻한 금융 [신한]입니다. 2019년에 힘들었던 모든 일들은 다 잊어버리시고, ~ | 80 | 0 | 4 | 0 | 0 | 1 |
| 5 | [WEB발신] 갤럭시 노트20/노트20 울트라 사전예약 오늘이 마지막날입니다-!! 구매를 망설이고 ~ | 75 | 0 | 6 | 0 | 0 | 0 |
| 6 | (광고)등촌역스톤힐 ★더블역세권 9호선 ~ ★선착순으로 동호수 지정분양 가능 ★인근주변 아파트 시세보다 4~5억 저렴 ★~ | 120 | 0 | 8 | 0 | 0 | 1 |
| 7 | [한진택배] 상품 배송 안내 안녕하세요 고객님. ★상품 수령이 편하신 장소를 선택 ~ ①직접수령 ②경비실 ③문앞 ~ | 110 | 0 | 9 | 0 | 0 | 0 |

기계학습 절차: 데이터 준비

레이블링 (Labeling)

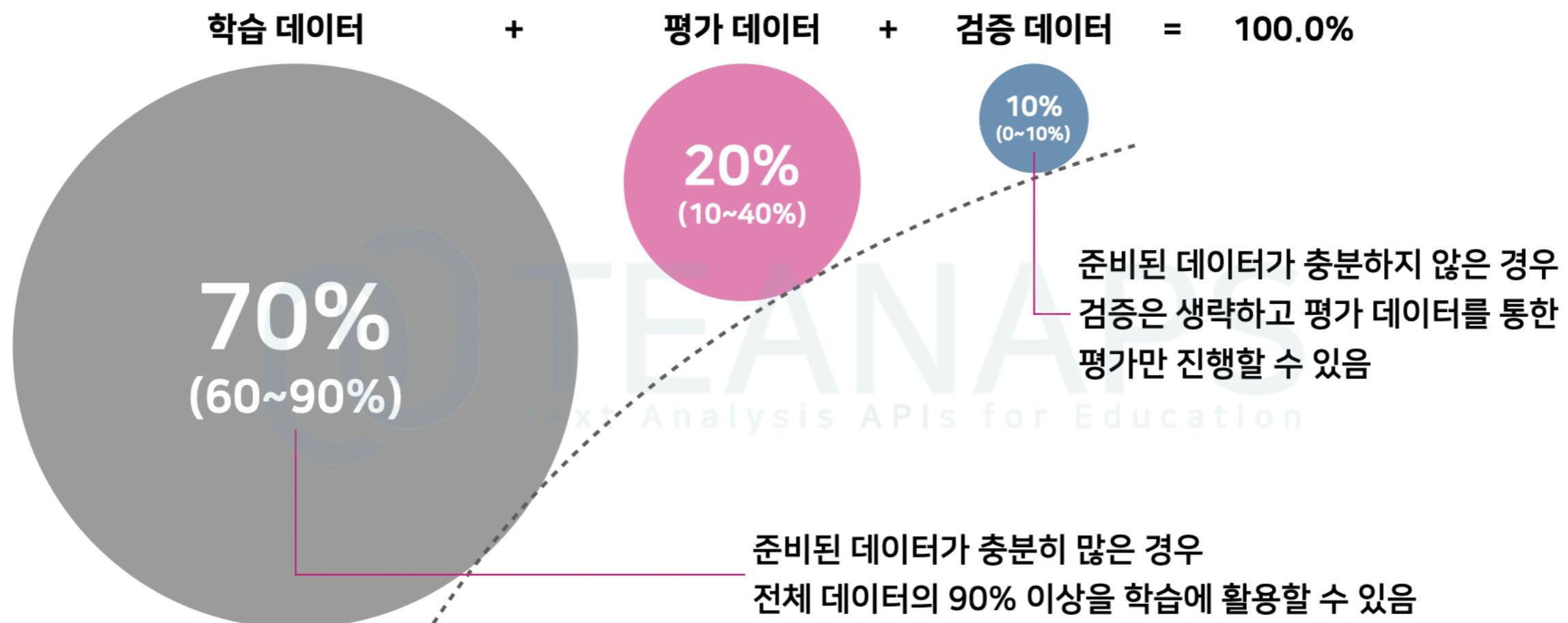
- 준비된 학습데이터에 지도학습을 위한 라벨(label)을 부착하는 과정
- 지도학습을 활용하는 경우, 학습데이터의 양과 레이블링의 정확도가 모델의 성능에 큰 영향을 미칠 수 있음
- 효율적으로 라벨을 부착하는 방법을 찾는 것도 데이터 분석 준비 과정 중 매우 중요한 요소로 작용함

| 구분 | 메시지 | 특징 (Feature) | | | | | | 라벨 (Label) |
|----|--|--------------|--------|---------|---------|---------|-----------|------------|
| | | 메시지 길이 | URL 여부 | 특수문자 개수 | 해외수신 여부 | 의심단어 개수 | 광고문자 표시여부 | |
| 1 | [국제발신] 하루 30만 달 1천만 만원으로 이렇게! http://bit.ly/3f~ | 40 | 1 | 6 | 1 | 1 | 0 | TRUE |
| 2 | 팀장님 이보람 선임입니다. 출근하시면 결재 부탁드립니다. | 20 | 0 | 1 | 0 | 0 | 0 | FALSE |
| 3 | (광고)웰컴박하라 vc⑤47③.co 코드 wc1004 무료수신거부 01084510000 | 45 | 1 | 5 | 0 | 1 | 1 | TRUE |
| 4 | (광고)신한과 함께하는 소중한 미래 따뜻한 금융 [신한]입니다. 2019년에 힘들었던 모든 일들은 다 잊어버리시고, ~ | 80 | 0 | 4 | 0 | 0 | 1 | TRUE |
| 5 | [WEB발신] 갤럭시 노트20/노트20 울트라 사전예약 오늘이 마지막날입니다-!! 구매를 망설이고 ~ | 75 | 0 | 6 | 0 | 0 | 0 | TRUE |
| 6 | (광고)등촌역스톤힐 ★더블역세권 9호선 ~ ★선착순으로 동호수 지정분양 가능 ★인근주변 아파트 시세보다 4~5억 저렴 ★~ | 120 | 0 | 8 | 0 | 0 | 1 | TRUE |
| 7 | [한진택배] 상품 배송 안내 안녕하세요 고객님. ★상품 수령이 편하신 장소를 선택 ~ ①직접수령 ②경비실 ③문앞 ~ | 110 | 0 | 9 | 0 | 0 | 0 | FALSE |

기계학습 절차: 데이터 준비

데이터 분리 (Partitioning)

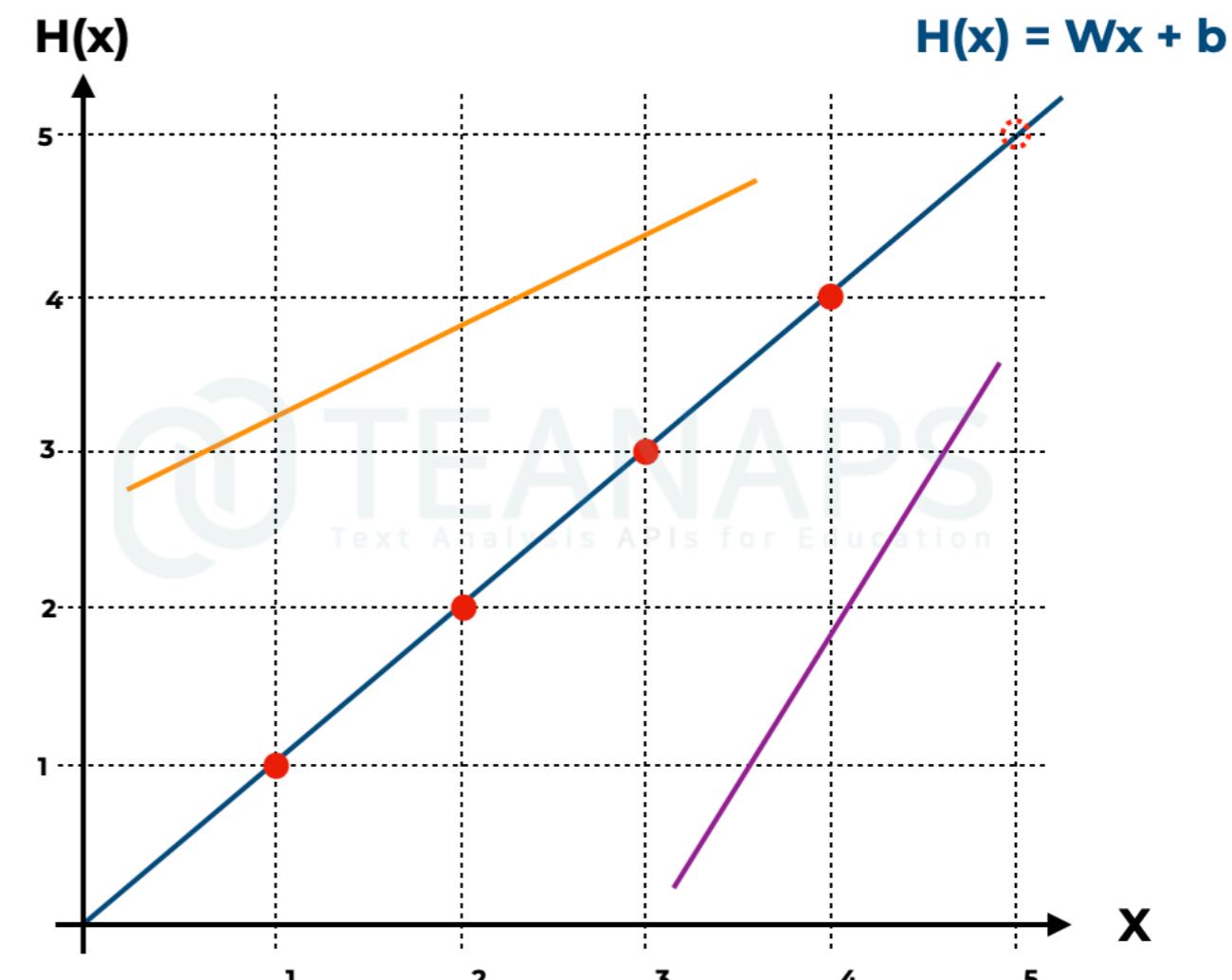
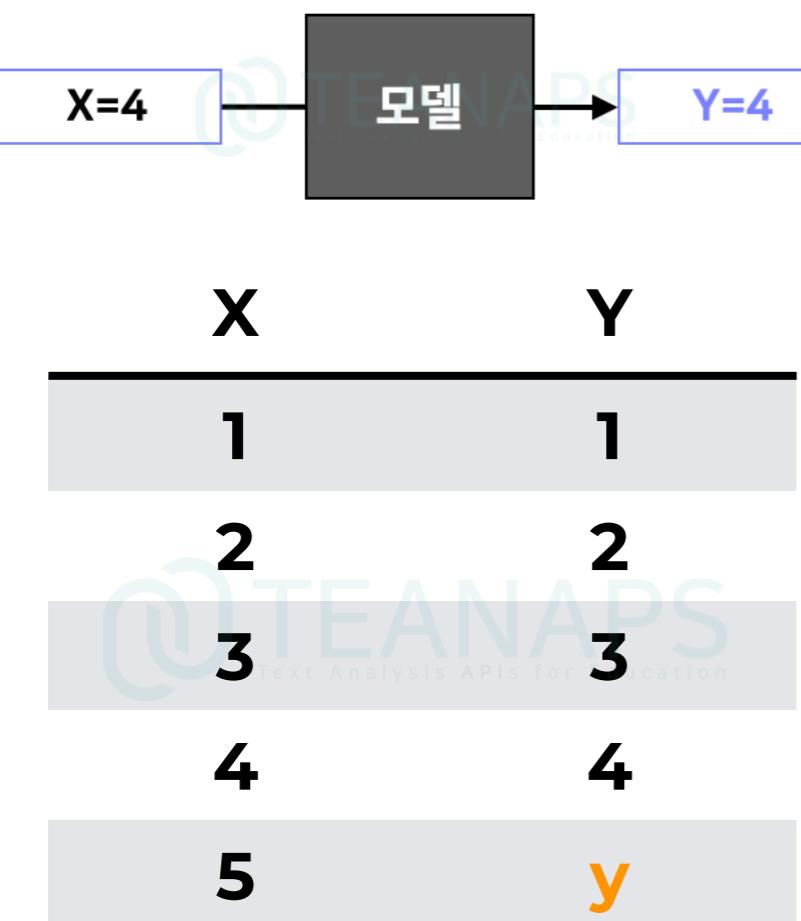
- 효율적인 학습과 평가를 위해, 준비된 데이터를 학습 데이터, 평가 데이터, 검증 데이터로 분리하는 과정
- **학습 데이터** (training data) : 기계학습 모델의 학습을 위한 데이터, 양은 많을수록 좋음
- **평가 데이터** (test data) : 모델의 학습결과 성능을 평가하기 위한 데이터
- **검증 데이터** (validation data) : 평가를 마친 모델에 대해 마지막 검증을 수행하기 위한 데이터



기계학습 절차: 학습 (Training)

| 기계가 데이터를 학습하는 과정 (Machine Training)

- 학습데이터에 정의된 정보나 규칙을 추상적인 형태로 표현하는 모델을 생성하는 과정
- 학습데이터에 포함된 다양한 정보나 규칙을 모델이 얼마나 잘 표현하는가에 따라 머신러닝 모델의 성능이 좌우됨
- **선형가정** (Linear Hypothesis) : 학습데이터의 분포를 선형이라 가정하고 학습데이터를 가장 잘 설명한 직선



기계학습 절차: 학습 (Training)

| 파라미터 접근법 (Parametric Approach)

- 입력변수(x)와 목표변수(y) 사이의 복잡한 관계를 어떠한 파라미터 (w)와의 관계로 표현하는 방식
- 정답을 구하기 위한 적절한 파라미터 (w)를 구하고 예측된 값 (y_n)과 정답 (Y)와의 차이 (error, loss)를 계산하여 그 평균을 최소로하는 적절한 파라미터 (w)를 도출하는 과정

$$y = \mathbf{a}x_1 + \mathbf{b}x_2$$

| a | x_1 | b | x_2 | y_n | Y | $Y - y_n$ |
|-----|-------|---|-------|-------|------------------|-----------|
| 0 | | | 1 | y_1 | 2 | $2 - y_1$ |
| ? | 1 | ? | 2 | y_2 | 6 | $6 - y_2$ |
| 1 | | | 1 | y_3 | 4 | $4 - y_3$ |
| 1.5 | | | 1 | y_4 | 5 | $5 - y_4$ |
| | | | | | Avg($Y - y_n$) | |

기계학습 절차: 학습 (Training)

| 파라미터 접근법 (Parametric Approach)

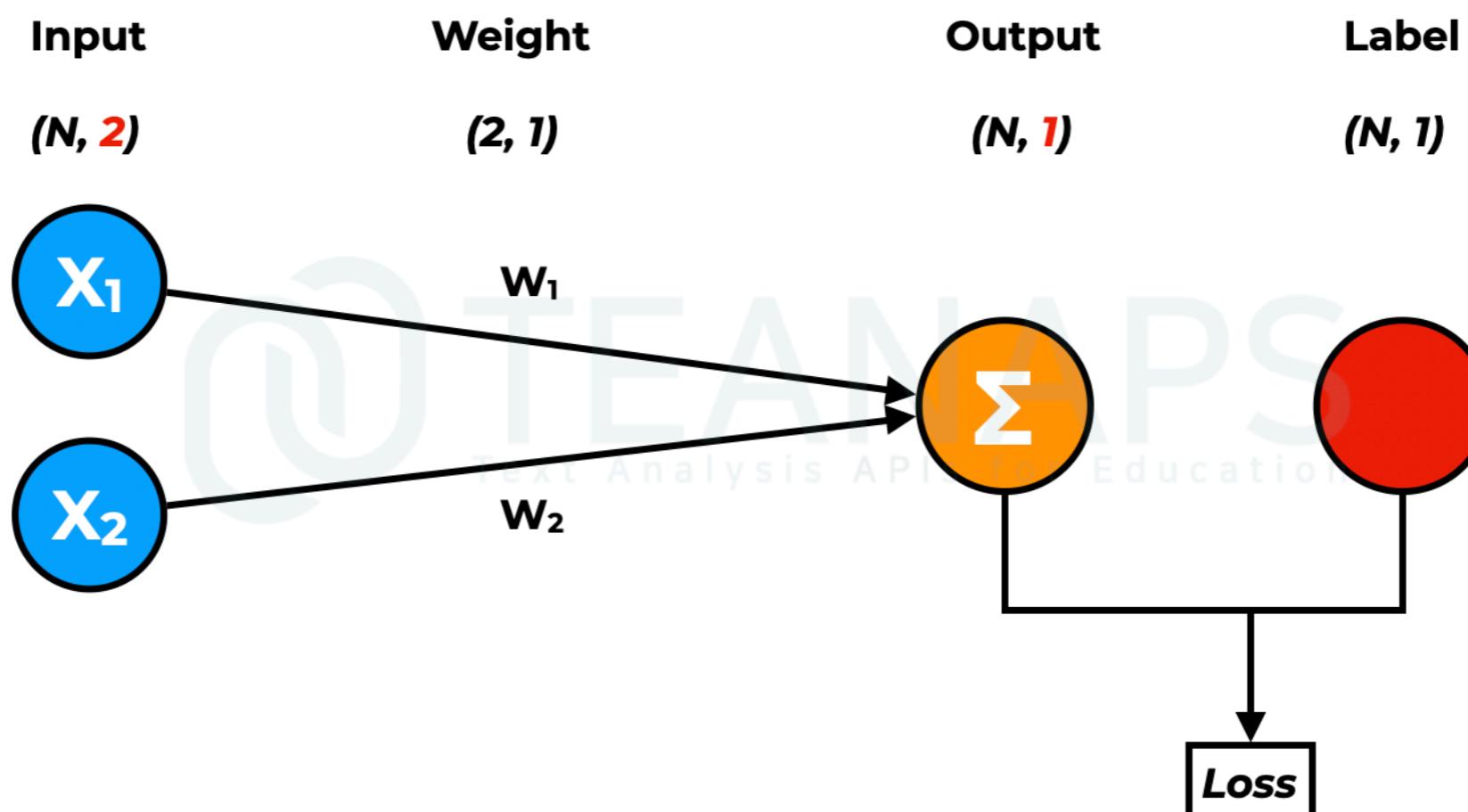
$$Y_n = W_1 \cdot X_1 + W_2 \cdot X_2$$

| W_1 | X_1 | W_2 | X_2 | Y_n | Y | $Y - Y_n$ |
|-------|-------|-------|-------|-------|------------------|-----------|
| 0 | | 1 | | Y_1 | 2 | $2 - Y_1$ |
| ? | 1 | ? | 2 | Y_2 | 6 | $6 - Y_2$ |
| 1 | | 1 | | Y_3 | 4 | $4 - Y_3$ |
| 1.5 | | 1 | | Y_4 | 5 | $5 - Y_4$ |
| | | | | | Avg($Y - Y_n$) | |

기계학습 절차: 학습 (Training)

| 파라미터 접근법 (Parametric Approach)

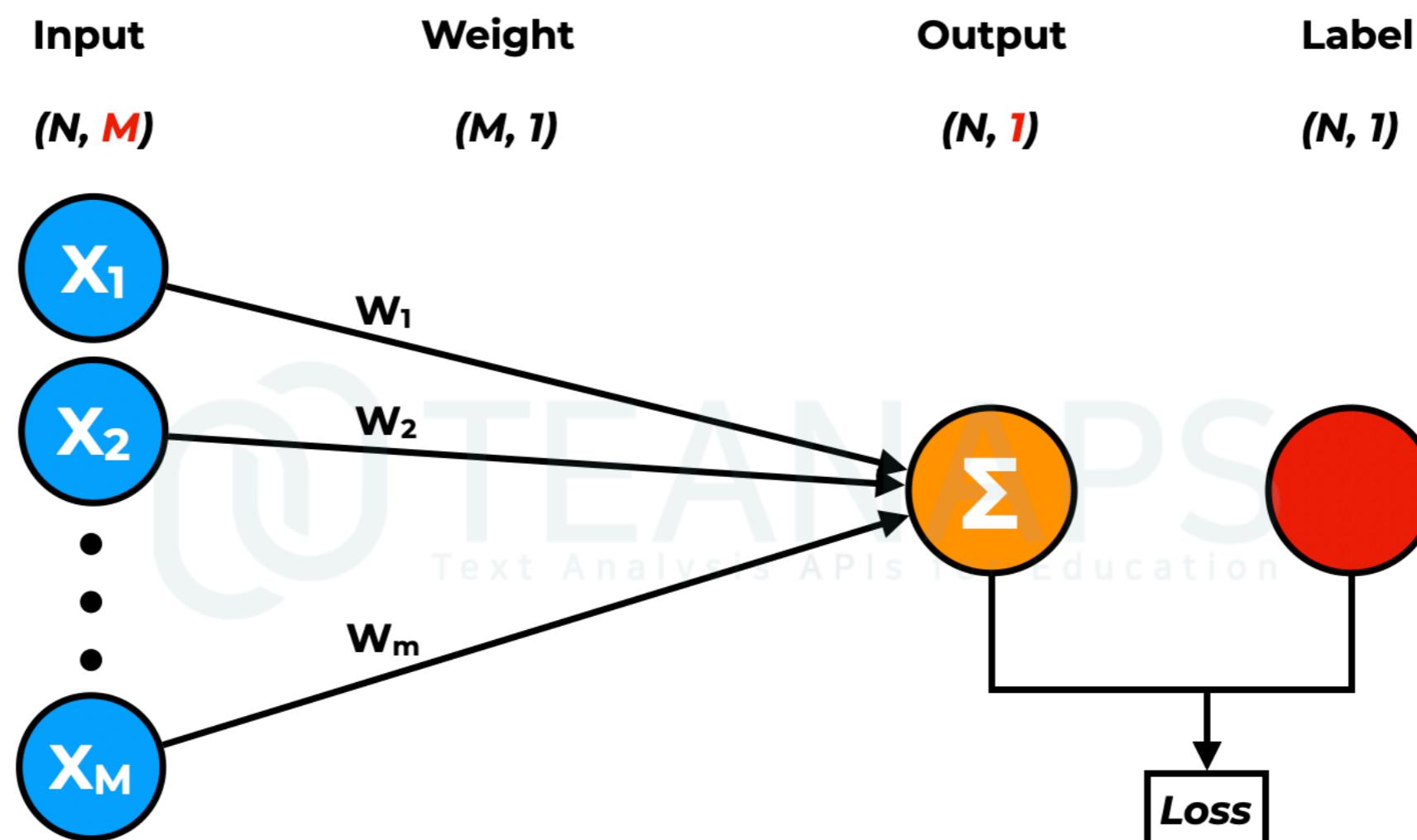
$$Y_n = W_1 \cdot X_1 + W_2 \cdot X_2$$



기계학습 절차: 학습 (Training)

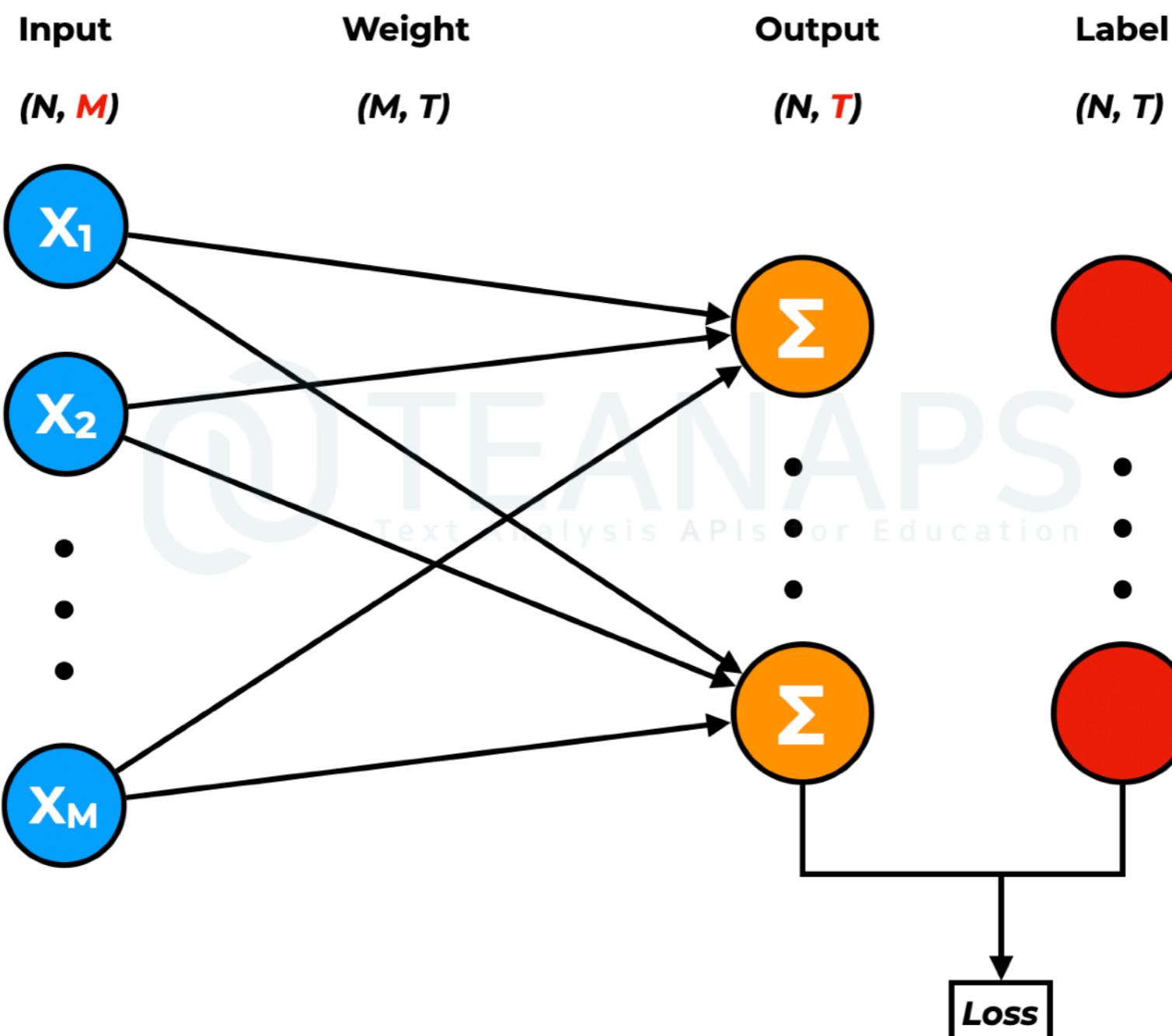
| 파라미터 접근법 (Parametric Approach)

$$Y_n = W_1 \cdot X_1 + W_2 \cdot X_2 + \dots + W_m \cdot X_m$$



기계학습 절차: 학습 (Training)

| 파라미터 접근법 (Parametric Approach)



기계학습 절차: 학습

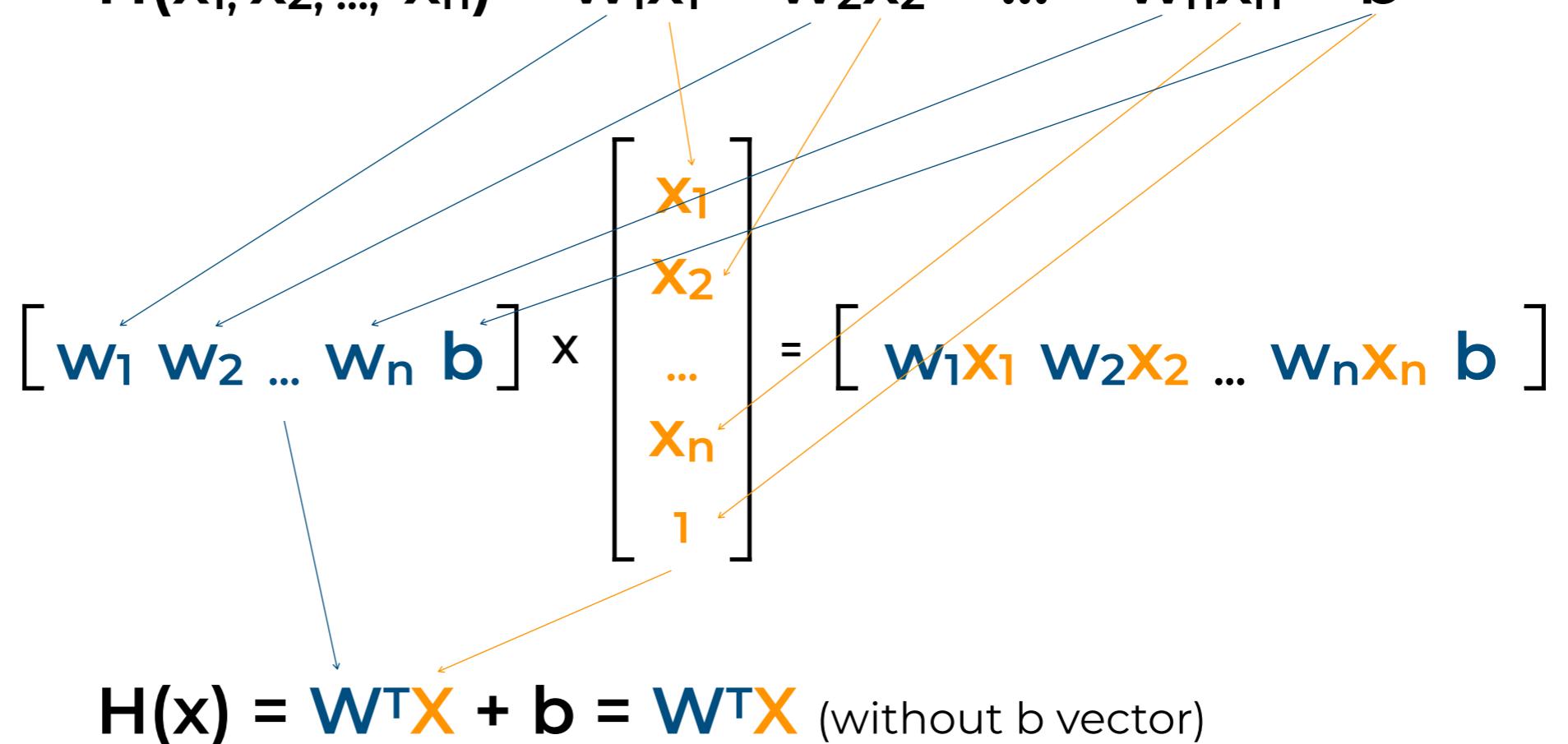
(Training)

| 입력변수가 2개 이상인 경우 (Multi-variable)

Single-variable : $H(x) = wx + b$

Multi-variable : $H(x_1, x_2) = w_1x_1 + w_2x_2 + b$

$$H(x_1, x_2, \dots, x_n) = w_1x_1 + w_2x_2 + \dots + w_nx_n + b$$



기계학습 절차: 학습

(Training)

손실함수: 평균제곱오차 (MSE, Mean Squared Error)

- 선형가정과 학습데이터 간의 차이를 측정하기 위한 방법
- 모델 학습은 최초에 새운 선형가정을 세우고 선형가정과 학습데이터 간의 손실함수가 최소가 되도록 가정을 수정해나가는 과정을 통해 이루어짐 → Cost 함수가 최소가 되는 w와 b를 찾는 과정

Cost(W)

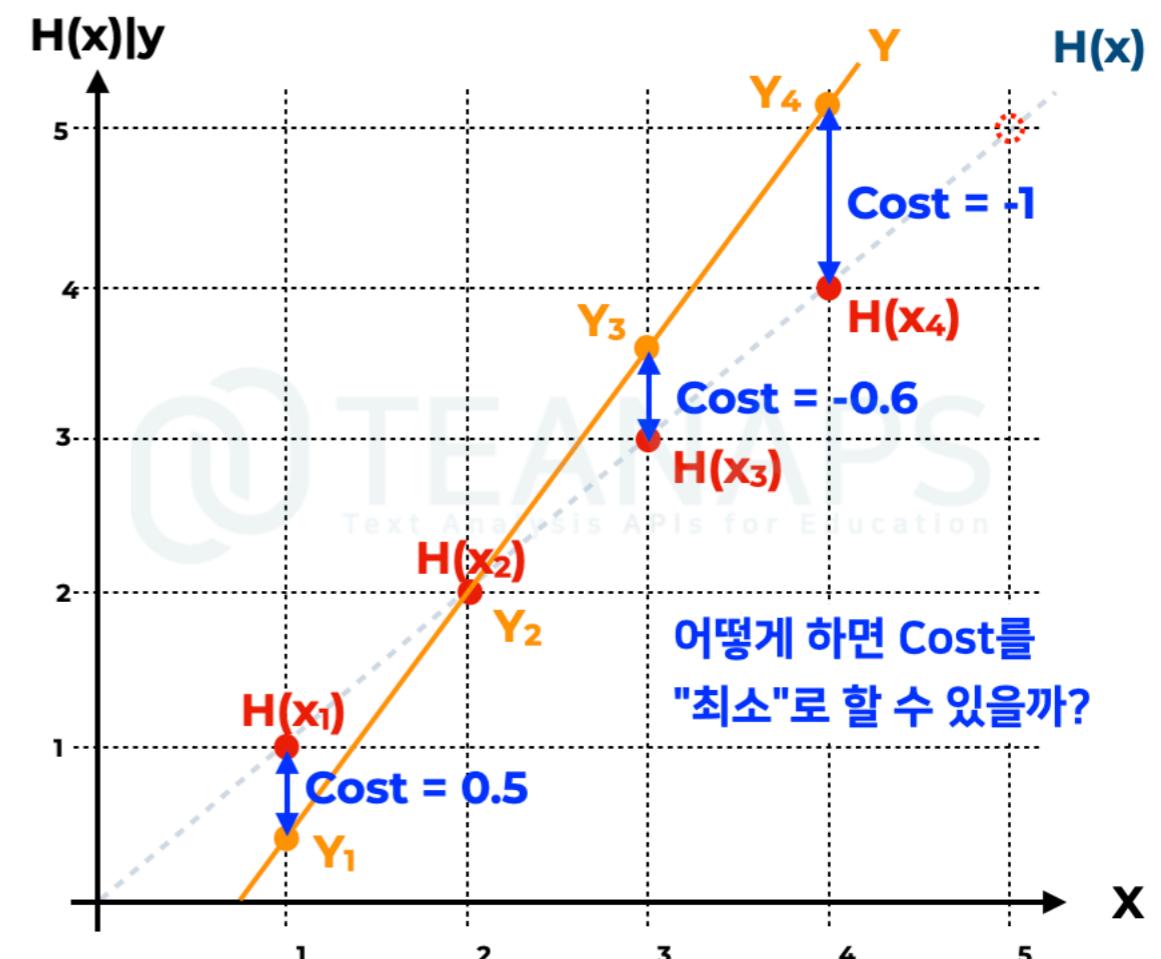
$$= H(x) - Y$$

$$= \{(H(x_1) - Y_1)^2 + (H(x_2) - Y_2)^2 + (H(x_3) - Y_3)^2 + (H(x_n) - Y_n)^2\} / n$$

$$= \frac{1}{n} \sum_{i=1}^n (H(x_i) - Y_i)^2$$

$$= \frac{1}{n} \sum_{i=1}^n (Wx_i - Y_i)^2 \quad (\text{without } b \text{ vector})$$

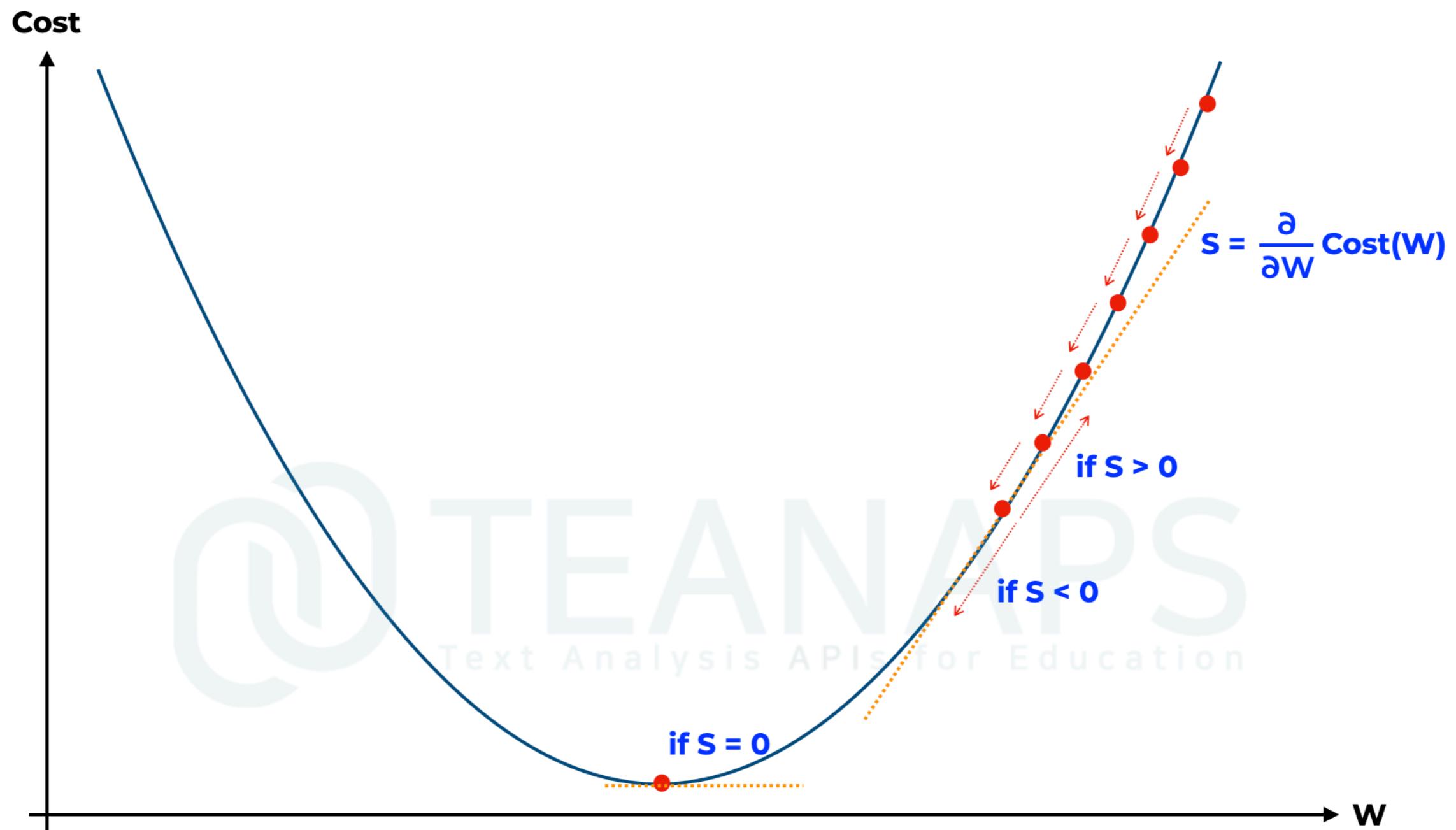
$W(x) = Wx$



어떻게 하면 Cost를
"최소"로 할 수 있을까?

기계학습 절차: 학습 (Training)

$$\text{Cost}(W) = W - \alpha \frac{1}{n} \sum_{i=1}^n (Wx_i - Y_i)x_i \quad (\alpha : \text{learning rate})$$



기계학습 절차: 학습

(Training)

손실이 최소가 되는 w와 b를 찾는 과정

- 손실함수를 설계한 후 손실(Cost)이 적어지는 방향으로 가중치(W)를 더 이상 수정이 불가능할 때 까지 수정하여 결과적으로 손실이 가장 적은 최적의 가중치를 찾아낼 수 있음
- 학습률 (learning rate, α) : 손실함수 상의 가중치를 조정하는 정도를 조절하는 변수

Cost(W)

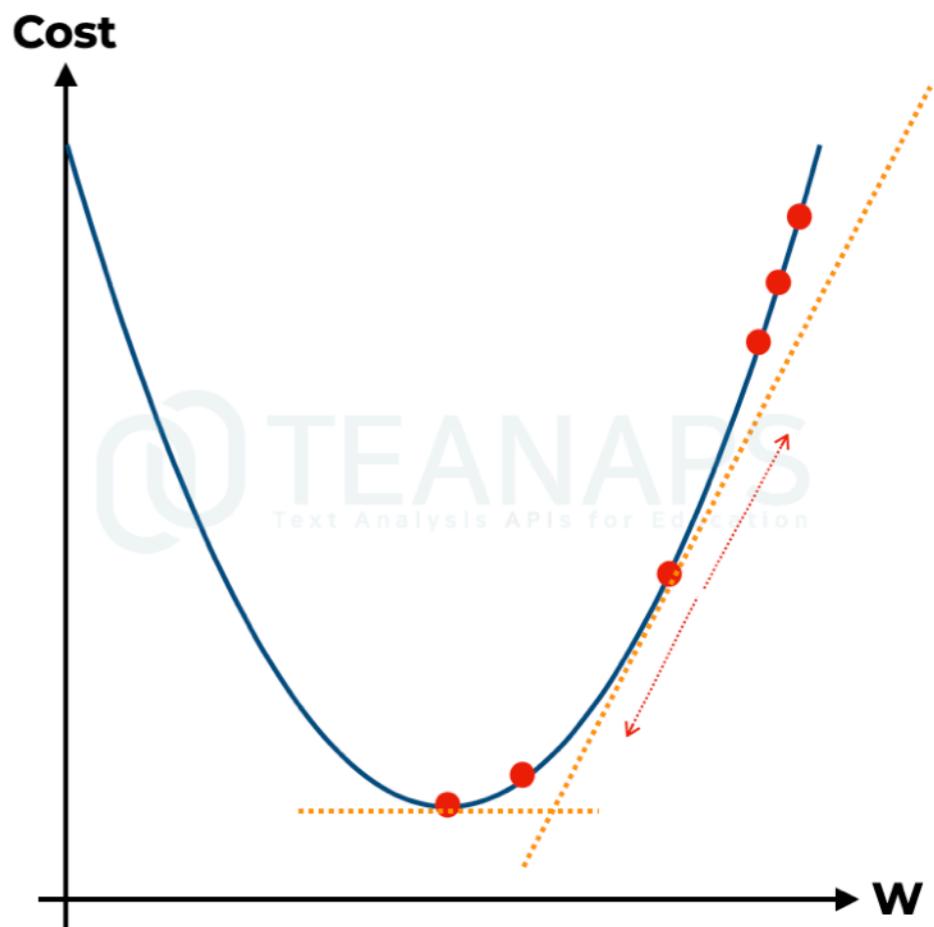
$$= \frac{1}{n} \sum_{i=1}^n (Wx_i - Y_i)^2 \approx \frac{1}{2n} \sum_{i=1}^n (Wx_i - Y_i)^2$$

$$W \rightarrow W - \alpha \frac{\partial}{\partial W} Cost(W)$$

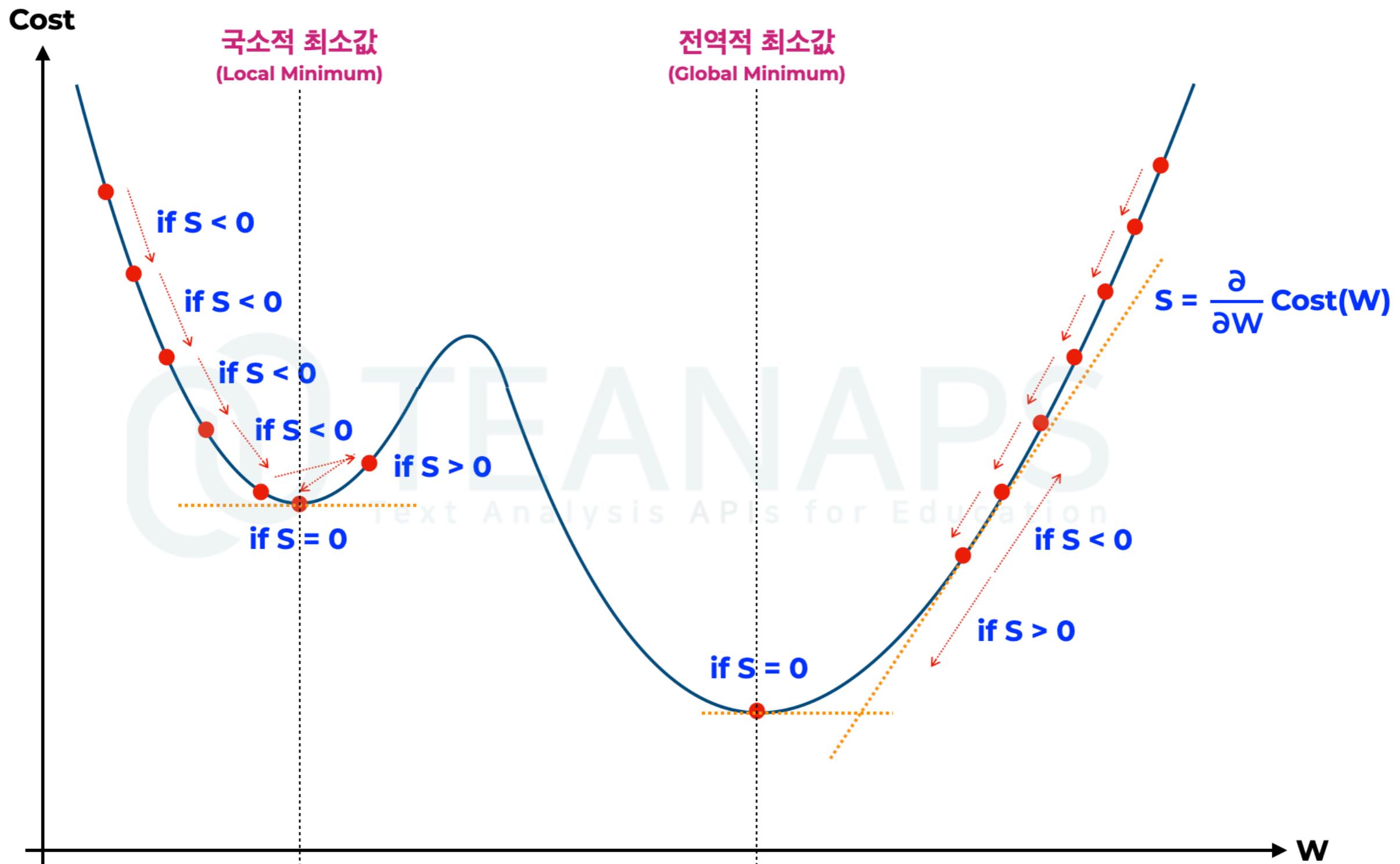
$$= W - \alpha \frac{\partial}{\partial W} \frac{1}{2n} \sum_{i=1}^n (Wx_i - Y_i)^2$$

$$= W - \alpha \frac{\partial}{\partial W} \frac{1}{2n} \sum_{i=1}^n 2(Wx_i - Y_i)x_i$$

$$= W - \alpha \frac{1}{n} \sum_{i=1}^n (Wx_i - Y_i)x_i \quad (\alpha : \text{learning rate})$$



기계학습 절차: 학습 (Training)



기계학습 절차: 학습 (Training)

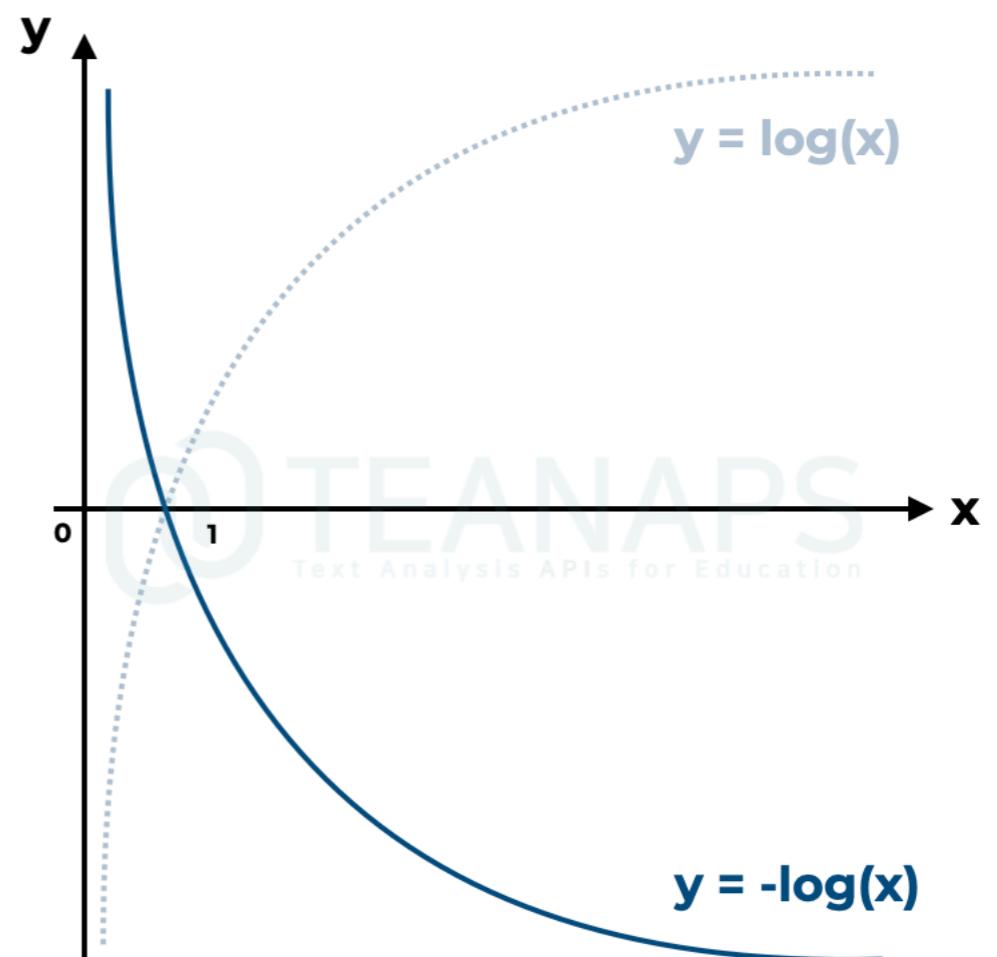
| Local Minimum 제거를 위한 손실함수 재설계

$$H(x) = Wx + b \rightarrow \text{Sigmoid}(H(x)) = \frac{1}{(1 + e^{-H(x)})}$$

$$\text{Cost}(W) = \frac{1}{n} \sum_{i=1}^n (H(x_i) - Y_i)^2 \quad (\text{Local Minimum이 빈번하게 발생함})$$

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

$$\text{Cost}'(W) = \frac{1}{n} \sum c(H(x), y)$$



기계학습 절차: 학습 (Training)

경사하강법 (Gradient Decent Algorithm)

- 미분을 적용하여 곡선의 기울기를 활용한 최적화 문제를 해결하는 알고리즘으로, 손실함수 최적화를 위해 Global Minimum 또는 Local Minimum을 찾는 대표적인 알고리즘

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

$$= -y\log(H(x)) - (1 - y)\log(1 - H(x)) \quad (\text{Cross Entropy})$$

$$\text{Cost}(w) = \frac{1}{n} \sum c(H(x), y)$$

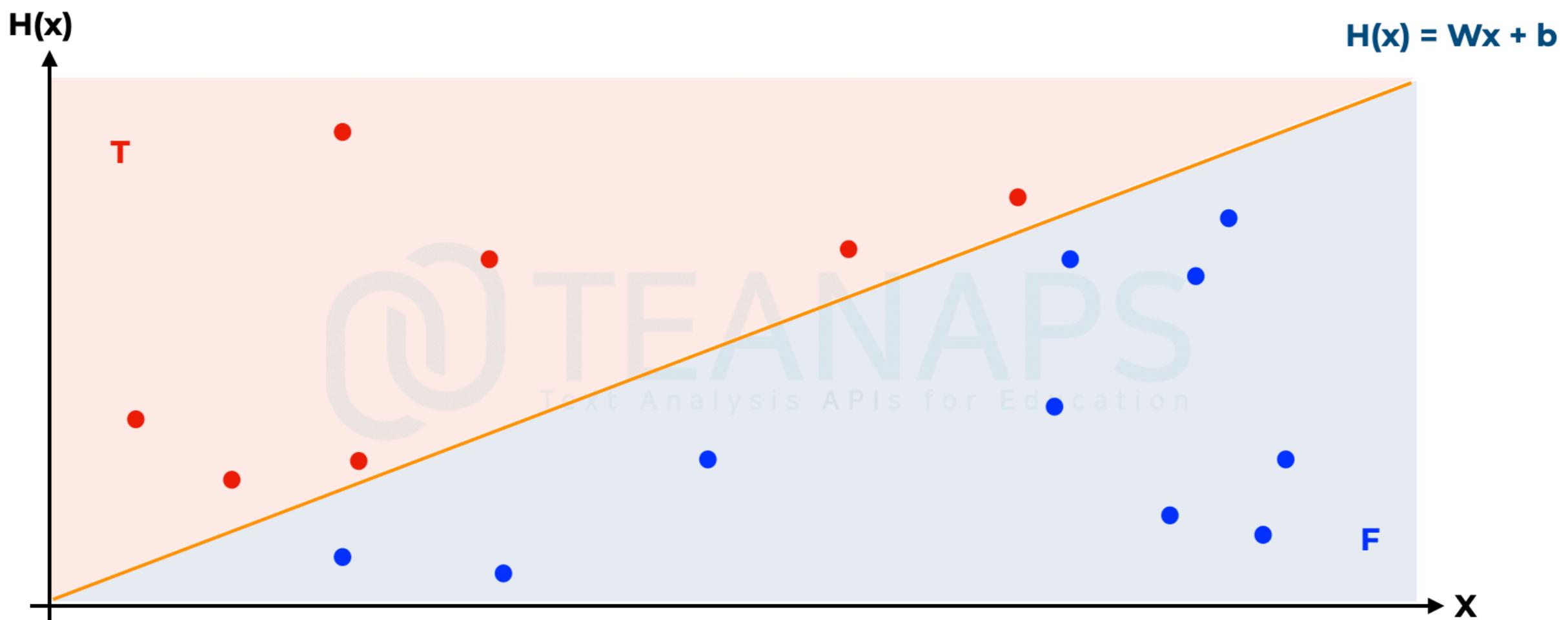
$$\text{Cost}(w) = -\frac{1}{n} \sum y\log(H(x)) - (1 - y)\log(1 - H(x))$$

$$W \rightarrow W - \alpha \frac{\partial}{\partial w} \text{Cost}(W)$$

기계학습 절차: 학습 (Training)

선형회귀로 분류문제를 해결하는 방법

- 선형 함수를 활성화 함수를 통해 로짓 함수로 변환하여 선형가정의 문제를 로짓가정의 문제로 변환할 수 있음
- **로짓가정** (Logistic Hypothesis) : 학습데이터의 분포를 로짓이라 가정하고 학습데이터를 가장 잘 설명한 직선
- **활성화 함수** (activation function) : 선형함수를 입력으로 활성화/비활성화 여부를 결정하여 출력하는 함수
(계단함수(Step Function), 시그모이드(Sigmoid), 하이퍼볼릭 탄젠트(tanh), Relu)

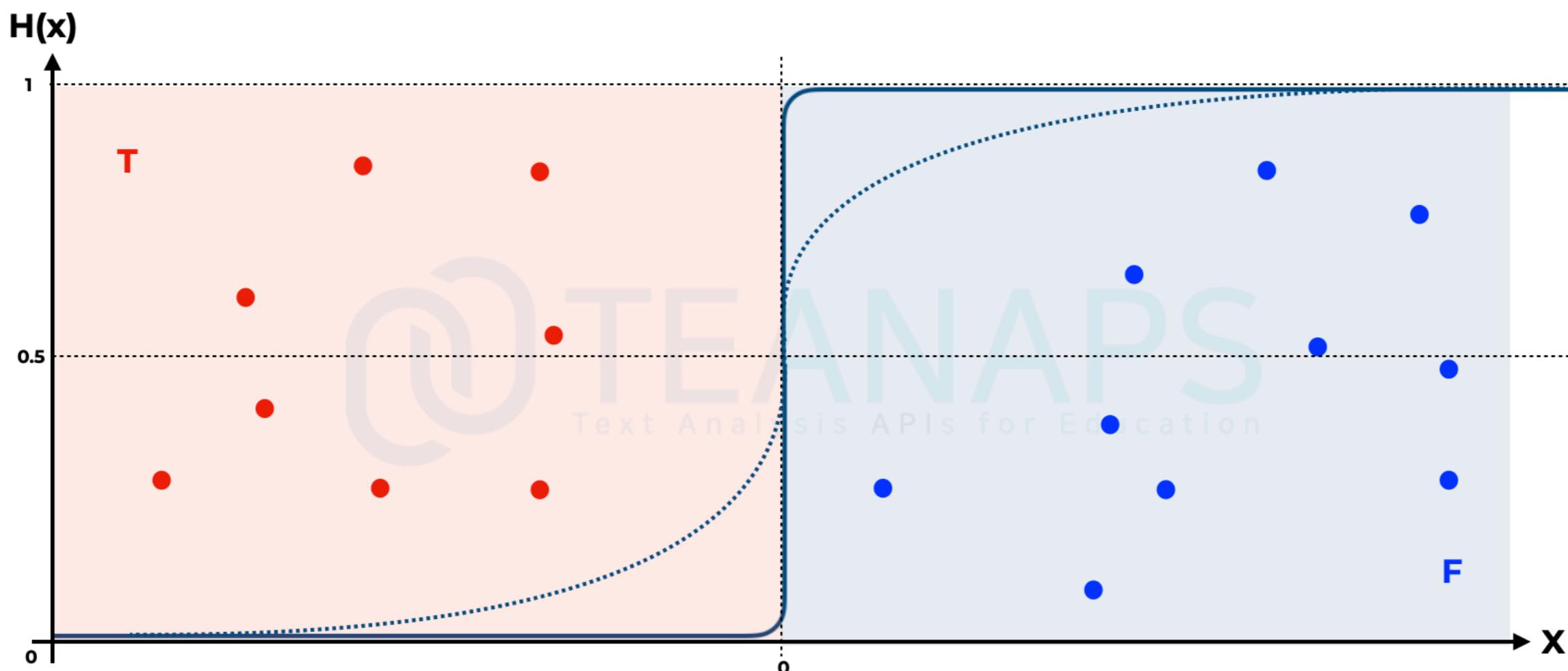


기계학습 절차: 학습 (Training)

시그모이드 함수를 통한 로짓변환

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

$$H(x) = Wx + b \rightarrow \text{Sigmoid}(H(x)) = \frac{1}{(1 + e^{-H(x)})}$$



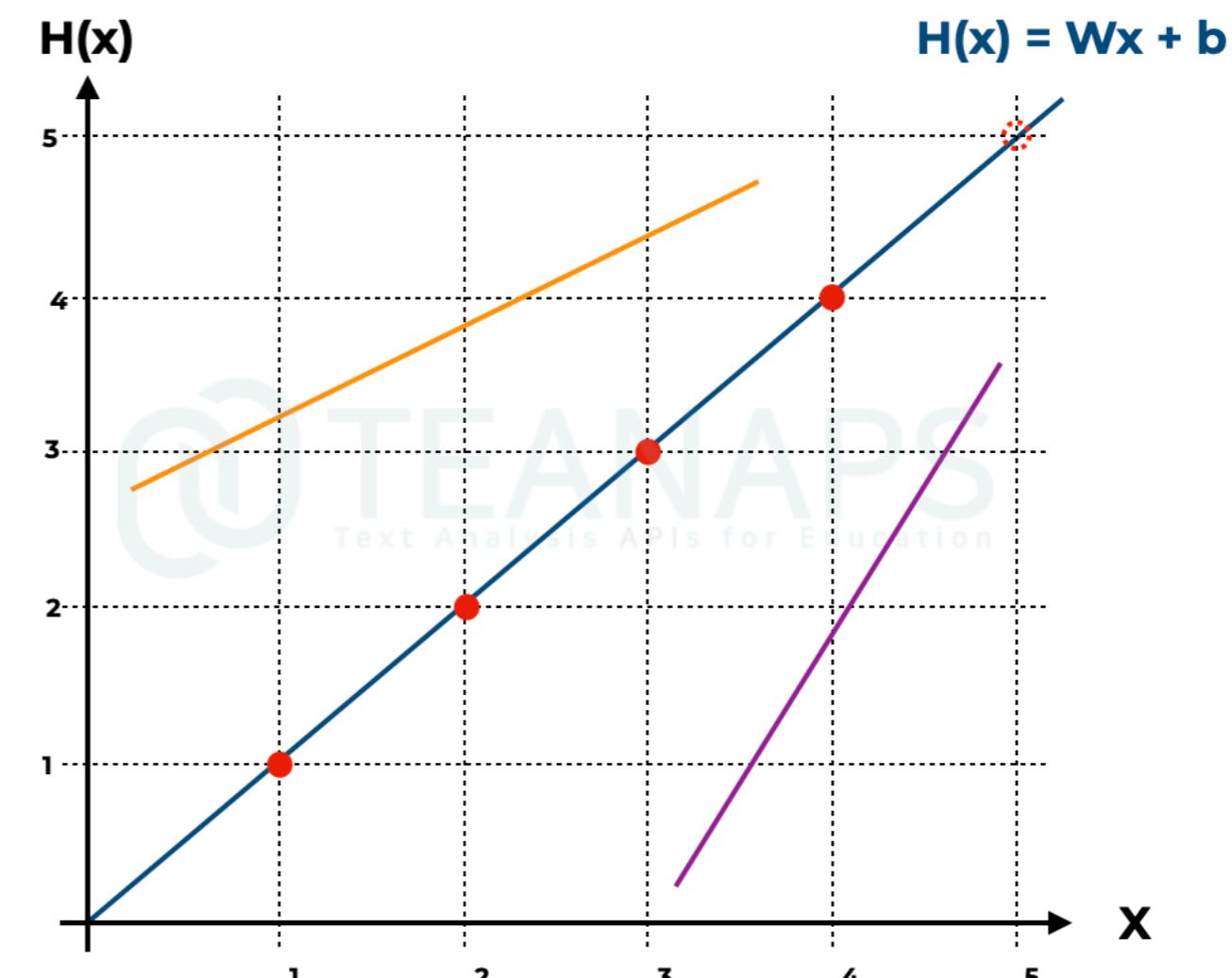
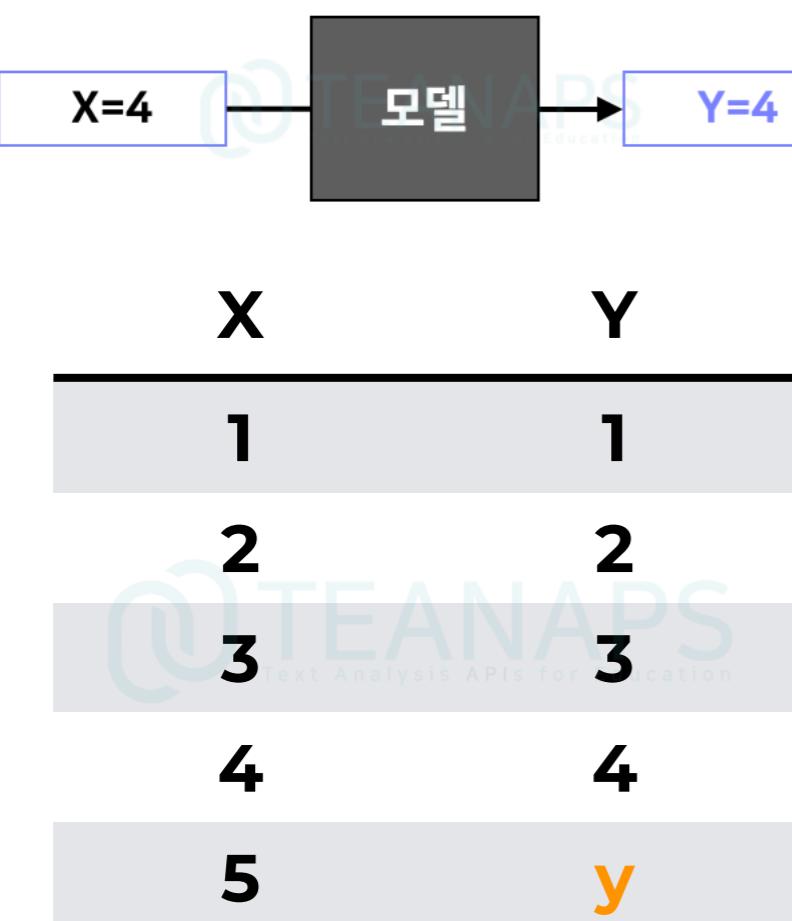
기계학습 절차: 학습

(Training)

Review

기계가 데이터를 학습하는 과정 (Machine Training)

- 학습데이터에 정의된 정보나 규칙을 추상적인 형태로 표현하는 모델을 생성하는 과정
- 학습데이터에 포함된 다양한 정보나 규칙을 모델이 얼마나 잘 표현하는가에 따라 머신러닝 모델의 성능이 좌우됨
- **선형가정** (Linear Hypothesis) : 학습데이터의 분포를 선형이라 가정하고 학습데이터를 가장 잘 설명한 직선

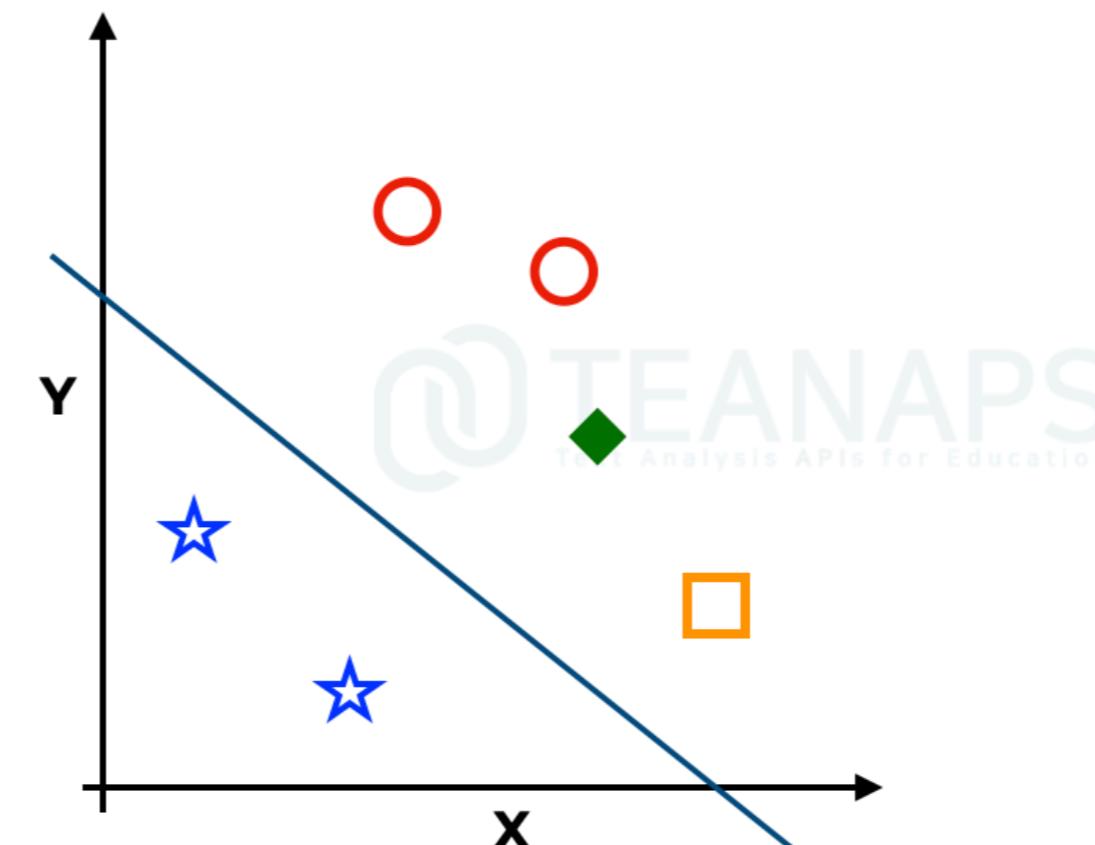


기계학습 절차: 학습 (Training)

다차원 분류 (Multinomial Classification)

- 3개이상의 범주를 가지는 분류문제
- 다차원 분류 문제는 2진분류^(binary classification) 문제를 해결하는 모델을 여러개 활용하여 해결 가능함

| X_1 | X_2 | Y |
|-------|-------|-----|
| 10 | 5 | A |
| 9 | 5 | A |
| 3 | 2 | B |
| 2 | 4 | B |
| 11 | 1 | C |

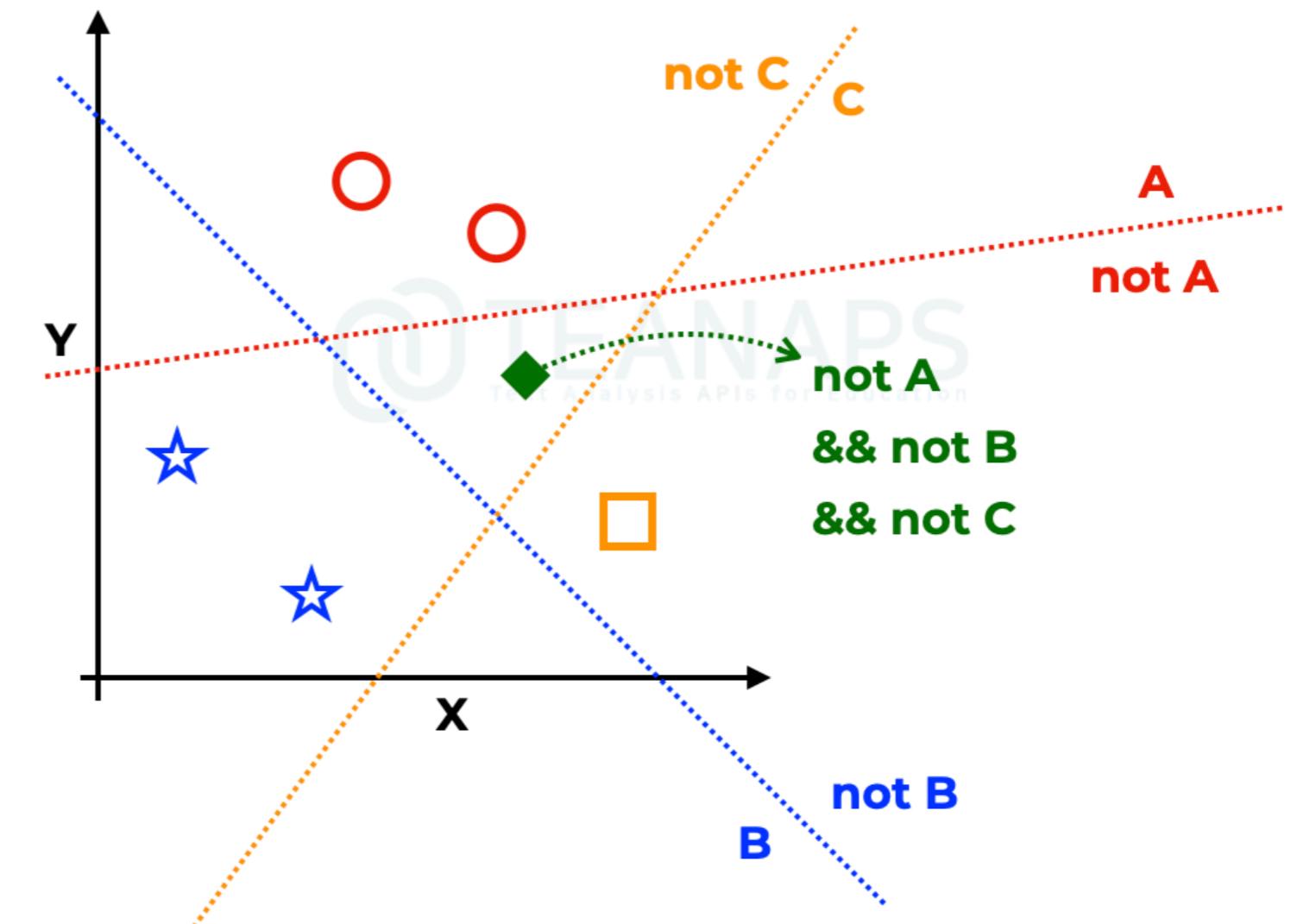


기계학습 절차: 학습 (Training)

다차원 분류 (Multinomial Classification)

- 3개이상의 범주를 가지는 분류문제
- 다차원 분류 문제는 2진분류^(binary classification) 문제를 해결하는 모델을 여러개 활용하여 해결 가능함

| X ₁ | X ₂ | Y |
|----------------|----------------|---|
| 10 | 5 | A |
| 9 | 5 | A |
| 3 | 2 | B |
| 2 | 4 | B |
| 11 | 1 | C |

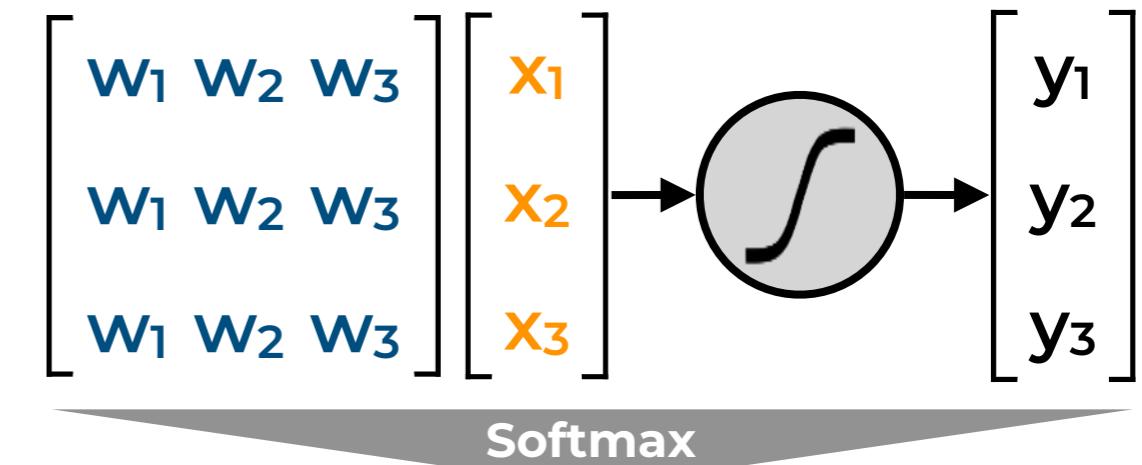


기계학습 절차: 학습

(Training)

다차원 분류 (Multinomial Classification)

$$\begin{bmatrix} W_1 & W_2 & W_3 \end{bmatrix} \times \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} W_1X_1 & W_2X_2 & W_3X_3 \end{bmatrix}$$



$$\begin{bmatrix} W_1 & W_2 & W_3 \end{bmatrix} \times \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} W_1X_1 & W_2X_2 & W_3X_3 \end{bmatrix}$$

$$\begin{bmatrix} W_1 & W_2 & W_3 \end{bmatrix} \times \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} W_1X_1 & W_2X_2 & W_3X_3 \end{bmatrix}$$

Softmax

$$\begin{bmatrix} \text{Softmax}(y_1) \\ \text{Softmax}(y_2) \\ \text{Softmax}(y_3) \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix}$$

확률값
(Probability)

argmax

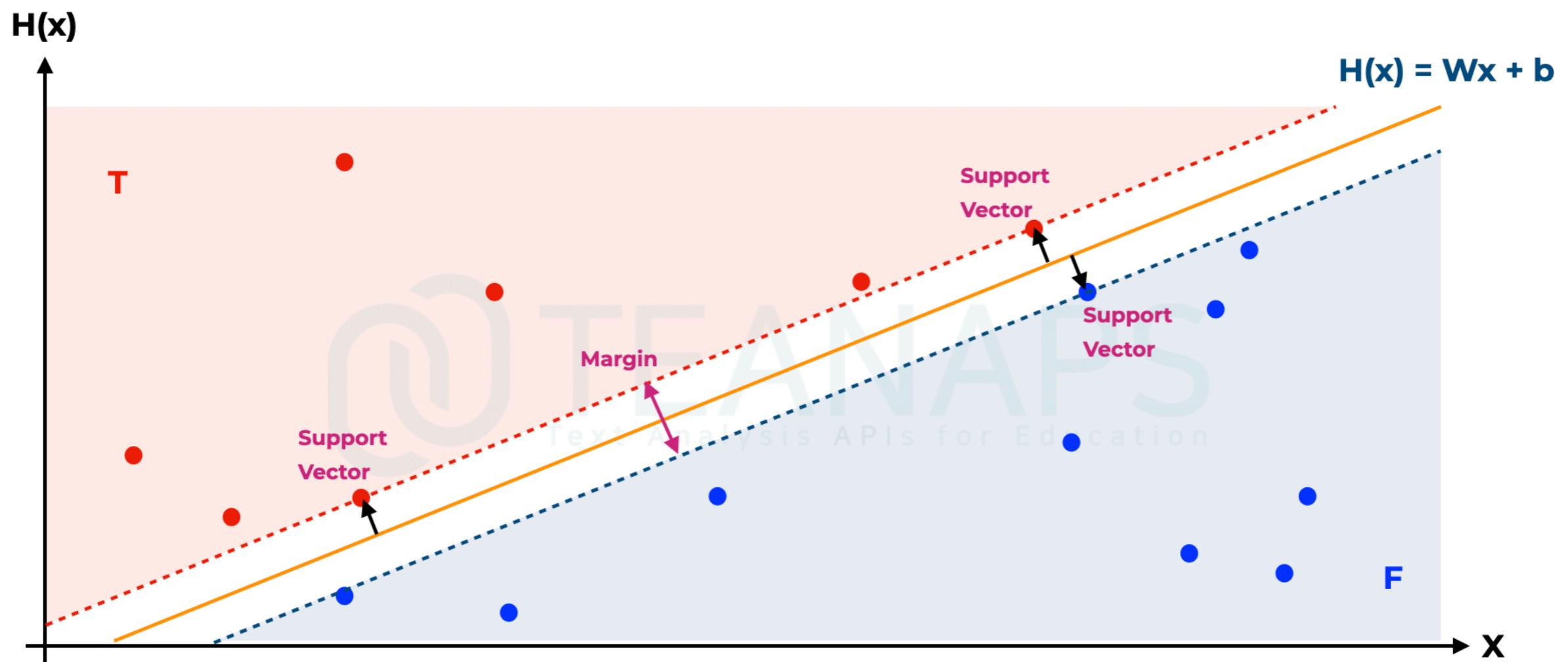
$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

One-hot (Binary)

분류 알고리즘: 서포트 벡터 머신

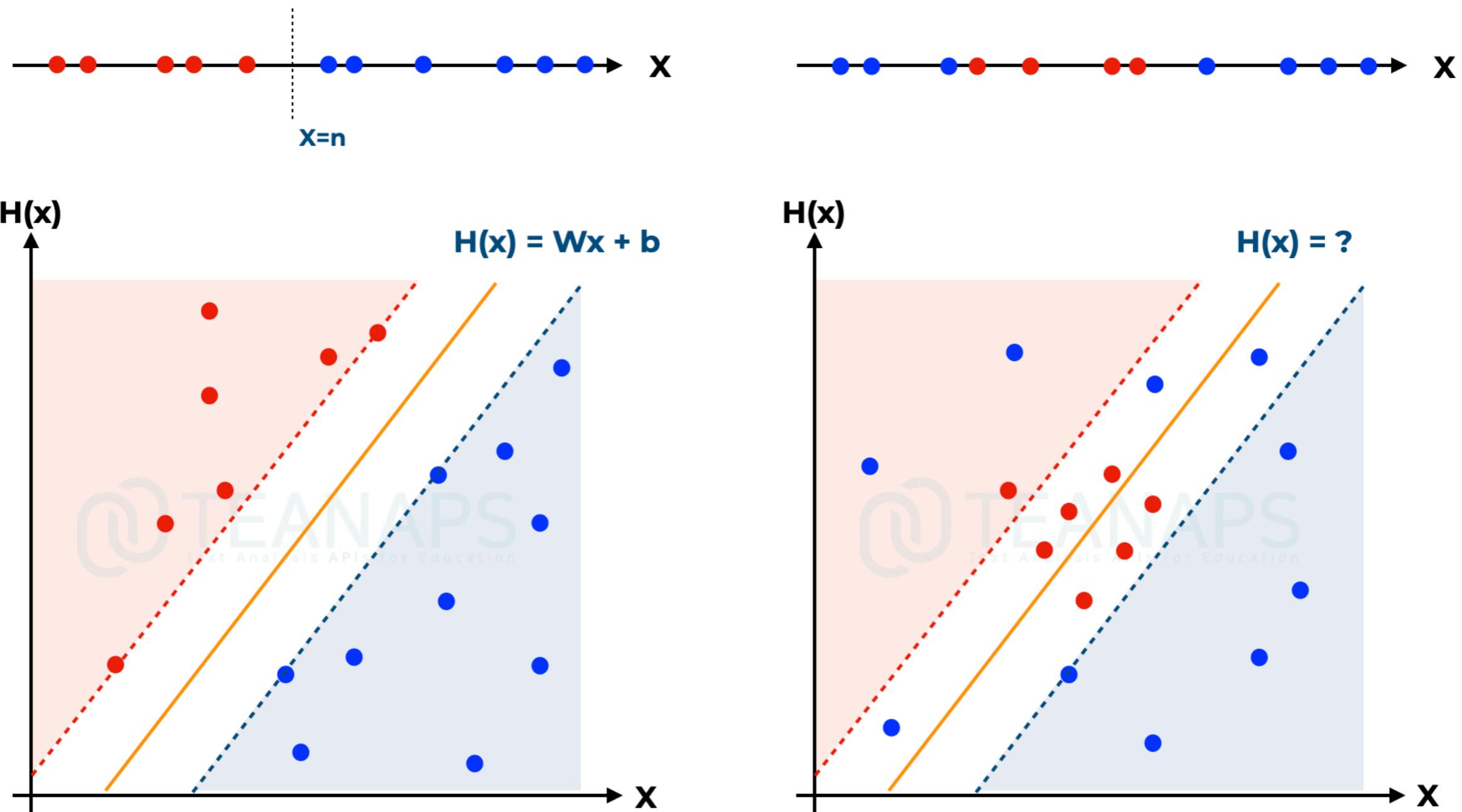
서포트 벡터 머신 (Support Vector Machine, SVM)

- 지지벡터로 이루어진 초평면과 마진을 최대로 하는 직선으로 선형 분류하는 기계학습 알고리즘
- 데이터를 선형함수로 분류할 수 없더라도 커널함수를 활용해 데이터를 고차원 공간으로 이동한 후 분류 가능함
- **지지벡터** (Support Vector) : 선형분류의 경계에 존재하는 데이터
- **커널함수** (kernel function) : 선형분류를 위해 데이터를 다른 차원으로 표현할 수 있도록 하는 함수



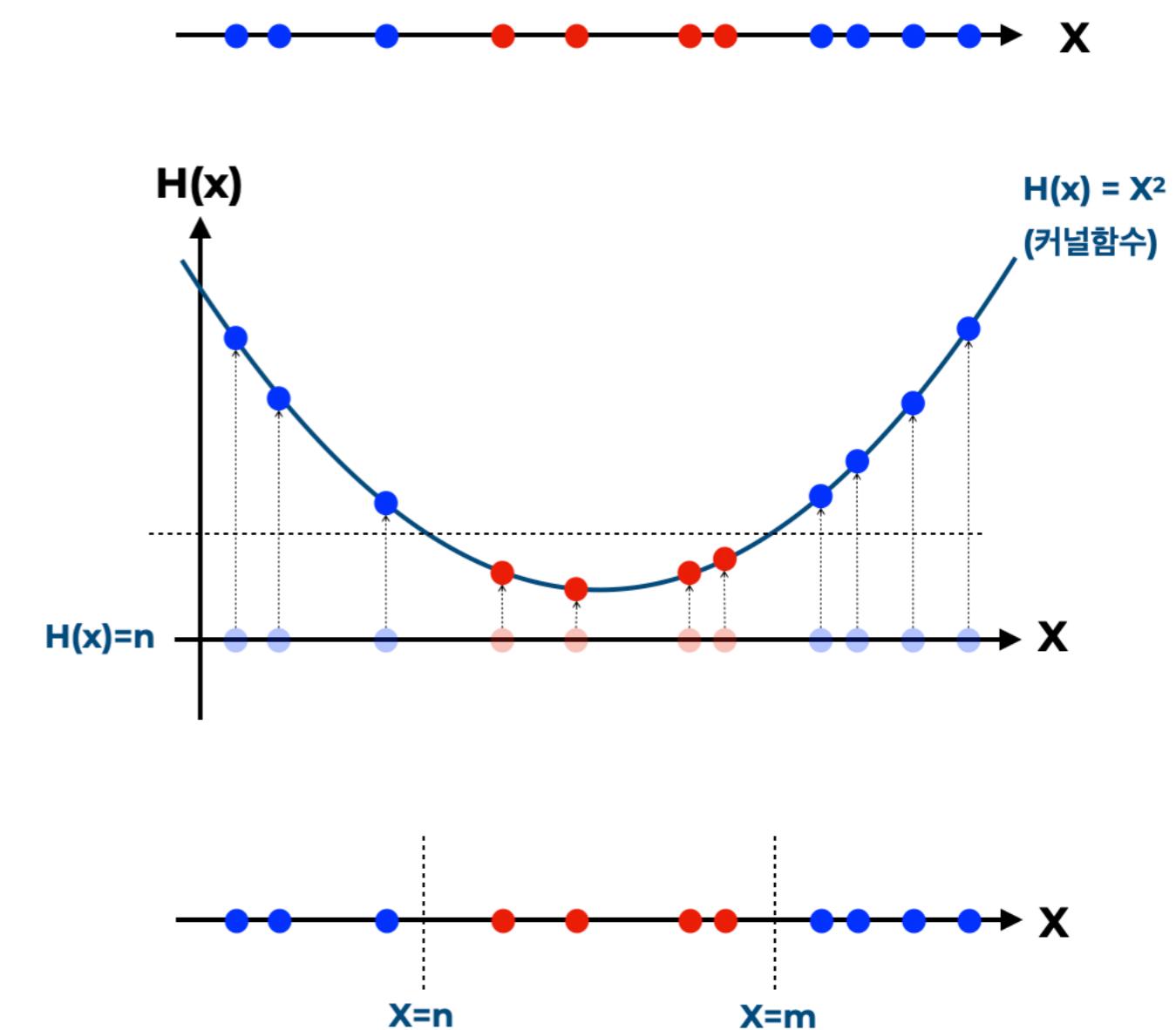
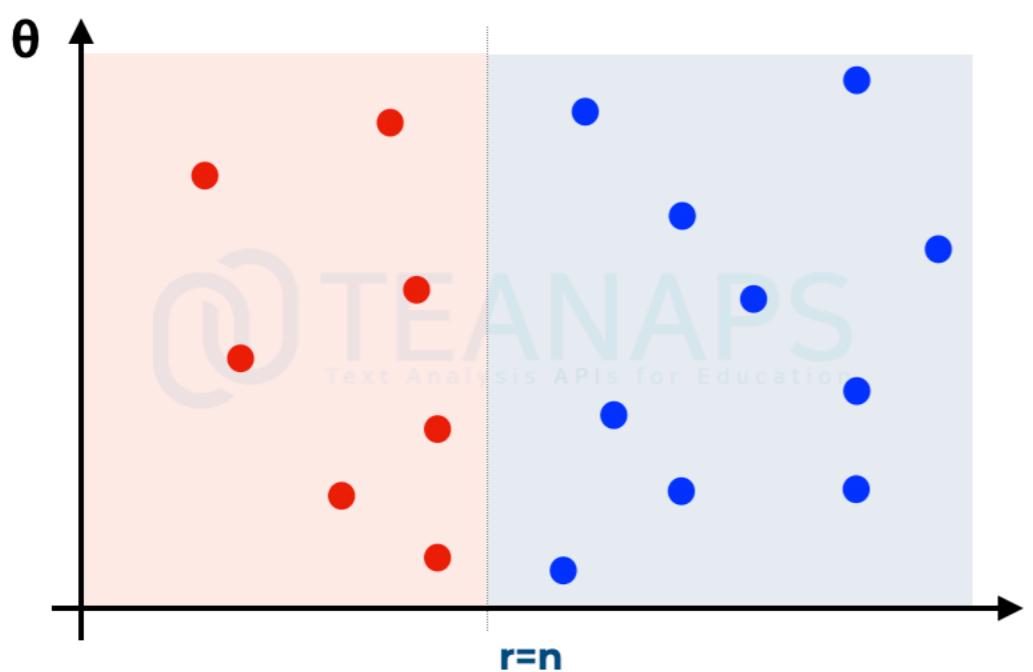
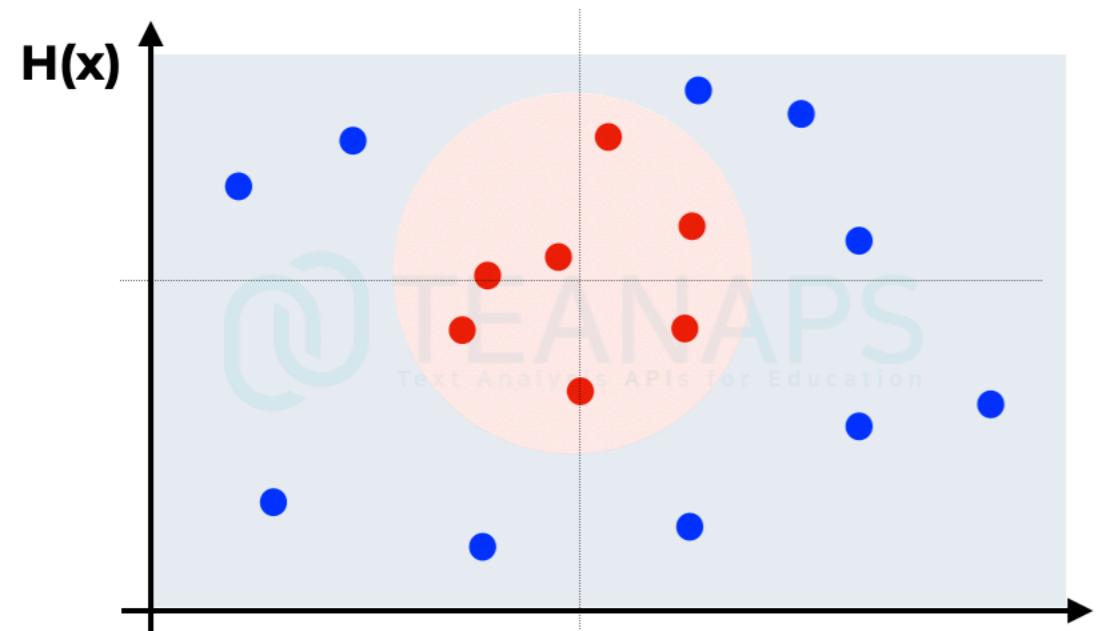
분류 알고리즘: 서포트 벡터 머신

커널함수 (Kernel Function)



분류 알고리즘: 서포트 벡터 머신

커널함수 (Kernel Function)



분류 알고리즘: 서포트 벡터 머신

| 커널함수 (Kernel Function)

Linear :

$$K(x_1, x_2) = x_1^T x_2$$

Polynomial :

$$K(x_1, x_2) = (x_1^T x_2 + c)^d \quad (c > 0)$$

Sigmoid :

$$K(x_1, x_2) = \tanh(a(x_1^T x_2) + b) \quad (a, b \geq 0)$$

Gaussian :

$$K(x_1, x_2) = \exp \left\{ \frac{\|x_1 - x_2\|_2^2}{2\sigma^2} \right\} \quad (a, b \geq 0)$$

기계학습 절차: 평가

(Validation)

2진분류 (Binary Classification) 모델을 평가하는 방법

- 모델의 분류 결과와 실제 정답과의 비교를 통해 모델의 성능을 평가할 수 있음
- **정확도 (Accuracy)** : 전체 분류결과 중 정답과 일치하는 분류결과의 비율
- **재현율 (Recall)** : 정답이 TRUE인 경우의 수 중 모델이 정답과 동일하게 분류한 결과의 비율
- **정밀도 (Precision)** : 모델이 TRUE로 분류한 결과 중 정답과 일치하는 분류결과의 비율
- **F1-score** : 재현율과 정밀도의 조화평균으로 두 가지 평가지표의 특성을 균등하게 반영할 수 있음

| | | 정답 | |
|-------|-------|---------------------------|---------------------------|
| | | TRUE | FALSE |
| 분류 결과 | TRUE | TRUE Positive (TP) | FALSE Positive (FP) |
| | FALSE | FALSE Negative (FN) | TRUE Negative (TN) |

1종오류
(Type1 Error)

2종오류
(Type2 Error)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1-score} = \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

기계학습 절차: 평가

(Validation)

다중분류 (Multi-class Classification) 모델을 평가하는 방법

- 다중분류 모델의 각 범주(Class)에 대해 2진분류와 동일한 평가를 진행하고, 각 범주 별 평가 결과의 평균값을 통해 전체 모델의 성능을 평가할 수 있음

| 정답 | | | 정답 | | | 정답 | | | 정답 | | | |
|---------|----------|-------|----|--------|-------|----|-----------|-------|----|----------|-------|----|
| | A | Not A | | B | Not B | | C | Not C | | D | Not D | |
| 분류 | A | TP | FP | B | TP | FP | C | TP | FP | D | TP | FP |
| 결과 | Not A | FN | TN | Not B | FN | TN | Not C | FN | TN | Not D | FN | TN |
| 범주 | Accuracy | | | Recall | | | Precision | | | F1-score | | |
| Class A | 89 | | | 88 | | | 92 | | | 89.9 | | |
| Class B | 100 | | | 100 | | | 100 | | | 100 | | |
| Class C | 83 | | | 84 | | | 78 | | | 80.8 | | |
| Class D | 65 | | | 62 | | | 67 | | | 64.4 | | |
| 최종 모델평가 | 84.25 | | | 83.5 | | | 84.25 | | | 83.8 | | |

E.O.D

Contact

-  <http://www.teanaps.com>
-  fingeredman@gmail.com