

머신러닝 기반 감성 분석

머신러닝 개요

인공지능, 머신러닝, 딥러닝

인공지능

- 예측/판단/추론/학습 등의 인간과 같은 **지적능력**을 컴퓨터를 통해 구현하는 기술

머신러닝

- 시스템이 **데이터를 통해 스스로 학습**하여 예측/판단/추론을 제공하는 기술
(지능을 구현하기 위한 SW 분야)

딥러닝

- 인공신경망 등을 심층화하여 특징도 스스로 추출하여 학습하는 머신러닝

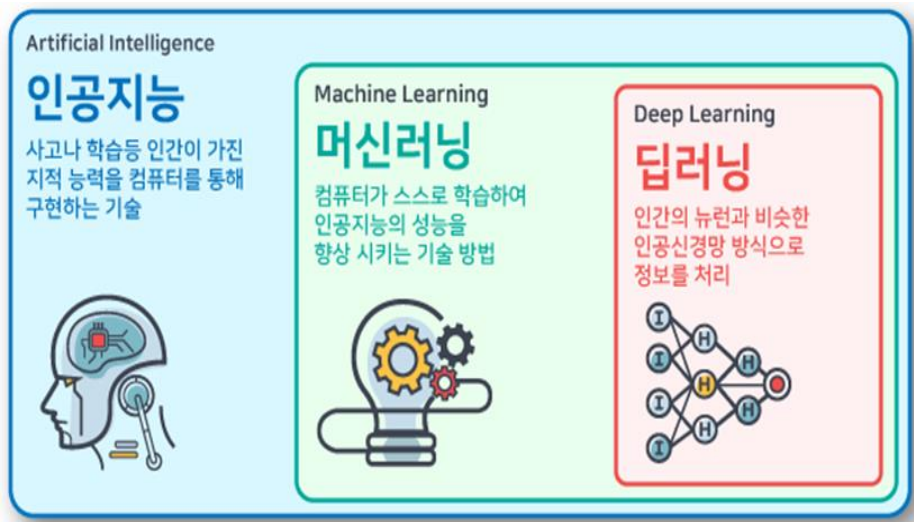
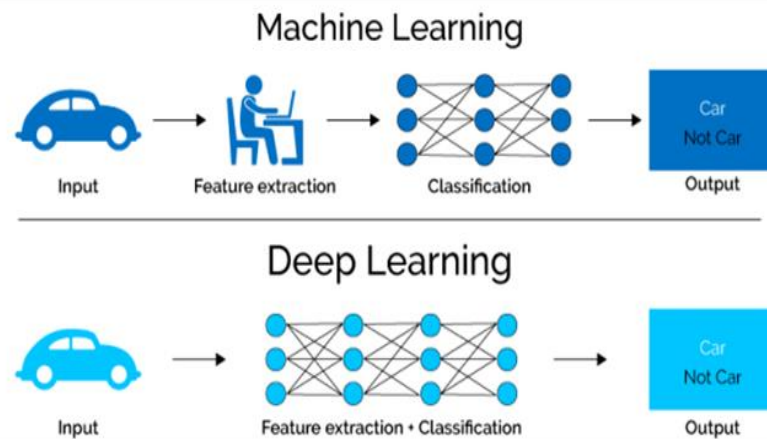


그림4 머신러닝 VS 딥러닝



자료: Towards Data Science, 메리츠증권증권 리서치센터

머신러닝Machine Learning 개념도

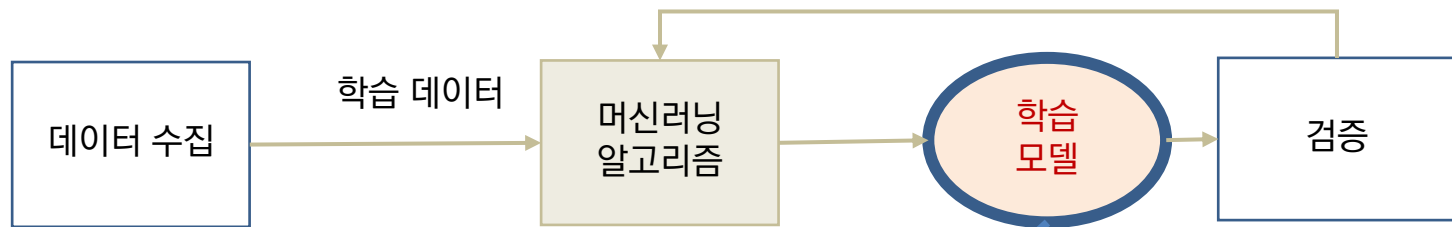
인공지능, AI

- 예측/판단/추론/학습 등의 인간과 같은 **지적능력**을 컴퓨터를 통해 구현하는 기술

머신러닝, 기계학습

- 시스템이 **데이터를 통해 스스로 학습**하여 예측/판단/추론을 제공하는 기술 (지능을 구현하기 위한 SW 분야)

학습 단계 (머신러닝)



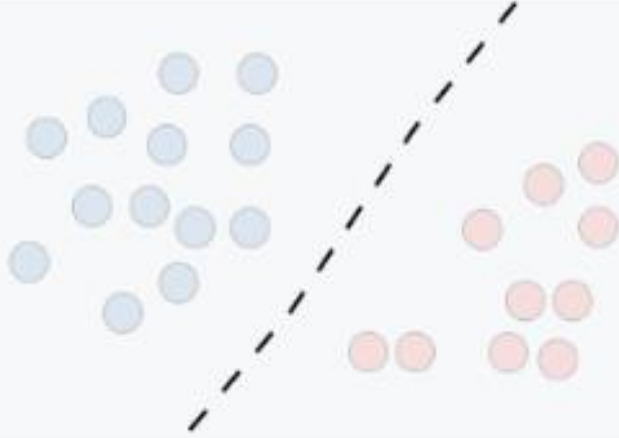
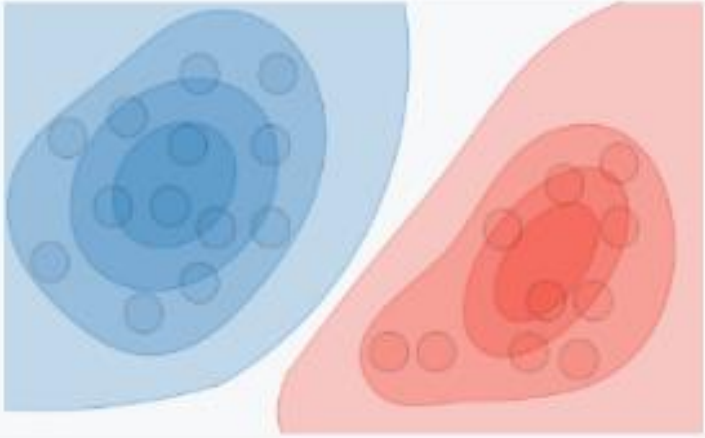
사용단계 (서비스, Application)



머신러닝의 방법



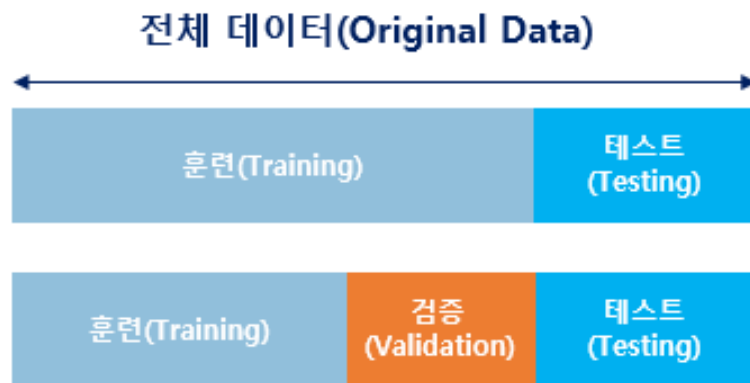
머신러닝을 통해 무엇을 학습하는가?

	판별형 학습모델	생성형 학습모델
	Discriminative model	Generative model
Goal	Directly estimate $P(y x)$	Estimate $P(x y)$ to then deduce $P(y x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
	Teachable Machine	WRTN, ChatGPT

머신러닝 모델의 평가

■ 머신러닝을 위한 데이터를 훈련용, 검증용, 테스트용으로 분리

- 훈련 데이터 : 머신러닝 모델 학습
- 검증 데이터 : 모델의 성능을 조정
- 테스트 데이터 : 머신러닝 모델의 성능을 평가



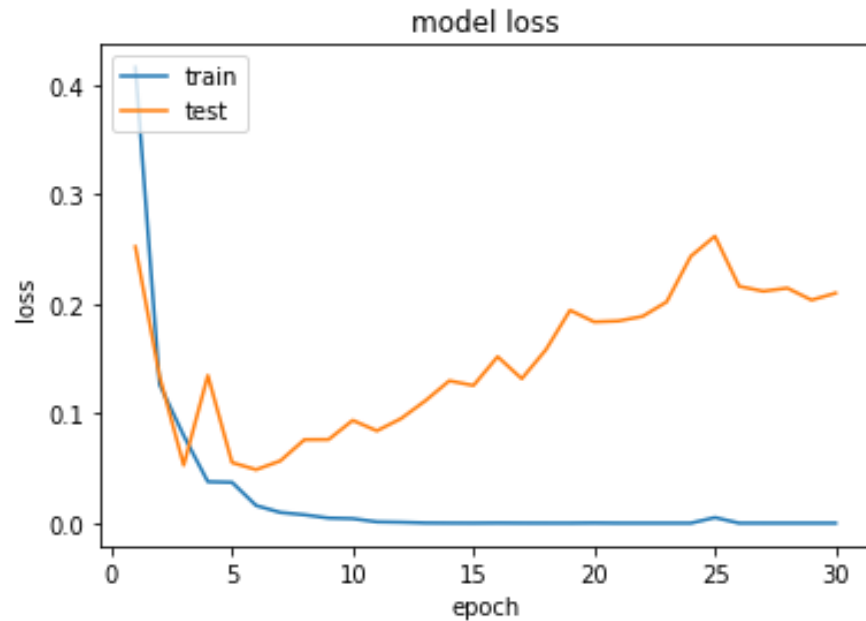
■ 성능 튜닝 : 하이퍼파라미터의 조정

- 하이퍼파라미터(초매개변수) : 모델의 성능에 영향을 주는 사람이 값을 지정하는 변수
 - 학습률, 학습 횟수, 한번에 학습할 데이터의 양 등
- 파라미터(매개변수) : 가중치와 편향. 학습을 하는 동안 값이 계속해서 변하는 수

과적합(Overfitting), 과소 적합(Underfitting)

■ 과적합(Overfitting)

- 훈련 데이터에 대해서는 오차가 낮지만, 테스트(신규) 데이터에 오차가 커지는 경우
- 훈련을 많이 한 경우, 데이터가 편향되어 있을 경우 등



■ 과소적합 (Underfitting)

- 테스트 데이터의 성능이 올라갈 여지가 있음에도 훈련을 덜 한 상태
- 훈련 자체가 부족한 상태

분류 머신러닝 알고리즘

의사결정나무(Decision Tree)

■ 의사결정 규칙을 나무 형태로 분류하는 분석 방법

- 상위 노드에서 시작하여 분류 기준값에 따라 하위 노드로 확장하는 방식이 ‘나무’를 닮았다고 하여 ‘의사결정나무’라고 불림
- 의사결정나무는 분석 과정이 직관적이고 이해하기 쉬움

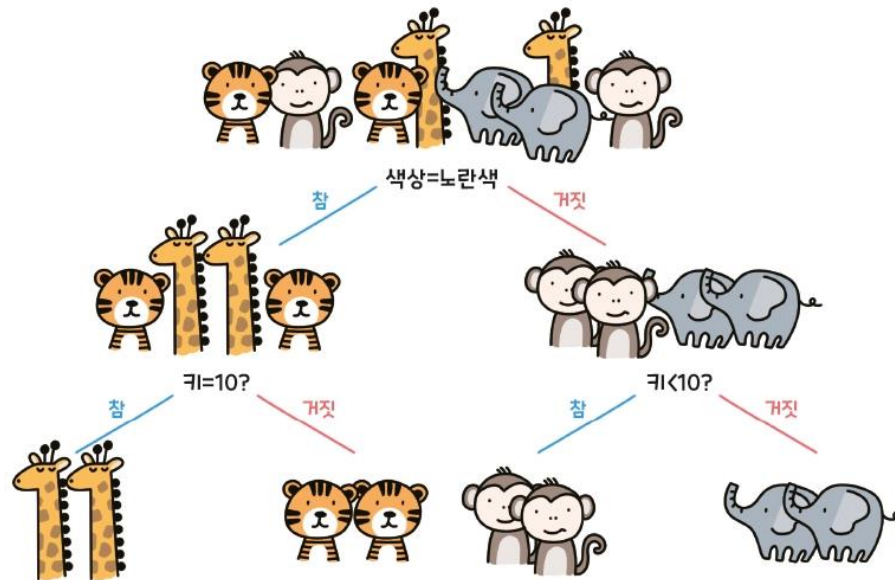
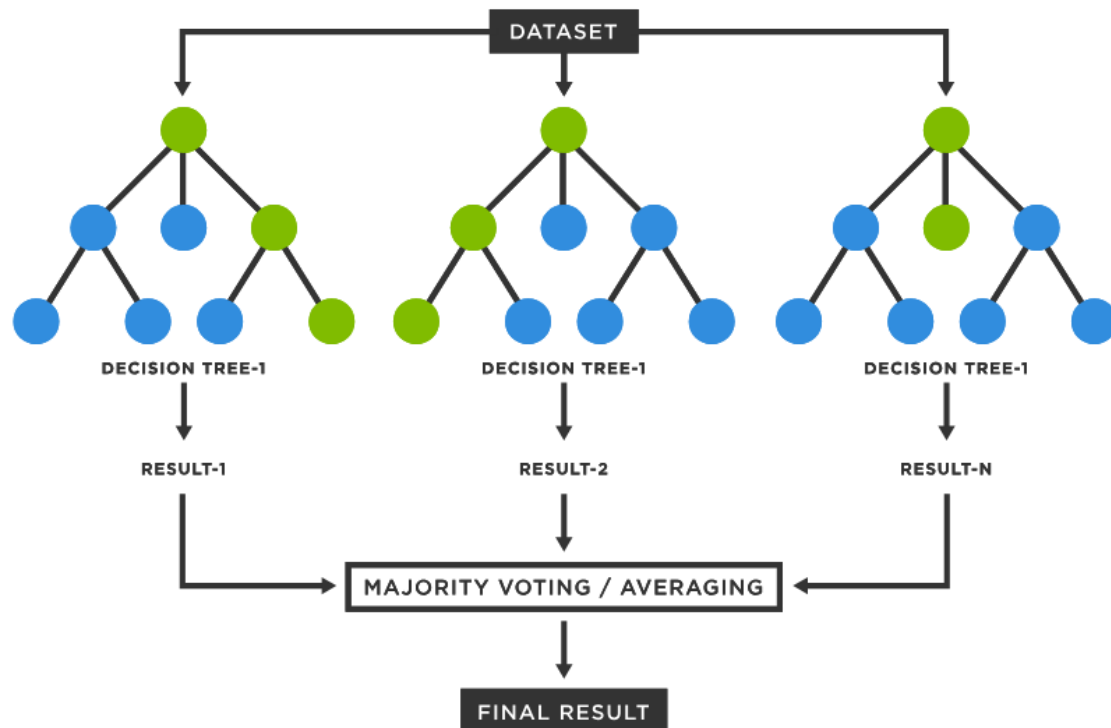


그림 8-19 의사결정나무 구조

랜덤 포레스트 (Random Forrest)

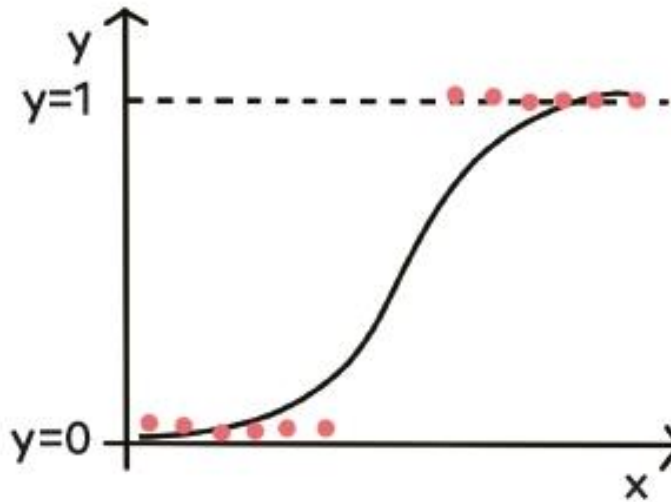
- Decision Tree를 random하게 여러개 만들어서 모델링
→ 각 Decision Tree의 예측을 사용해 최종 예측 수행
 - n_estimators : Decstion Tree의 수 (기본 100개)
 - Bootstrap Sampling : 학습 데이터에서 데이터를 중복해서 샘플링하는 방식



로지스틱 회귀(Logistic Regression)

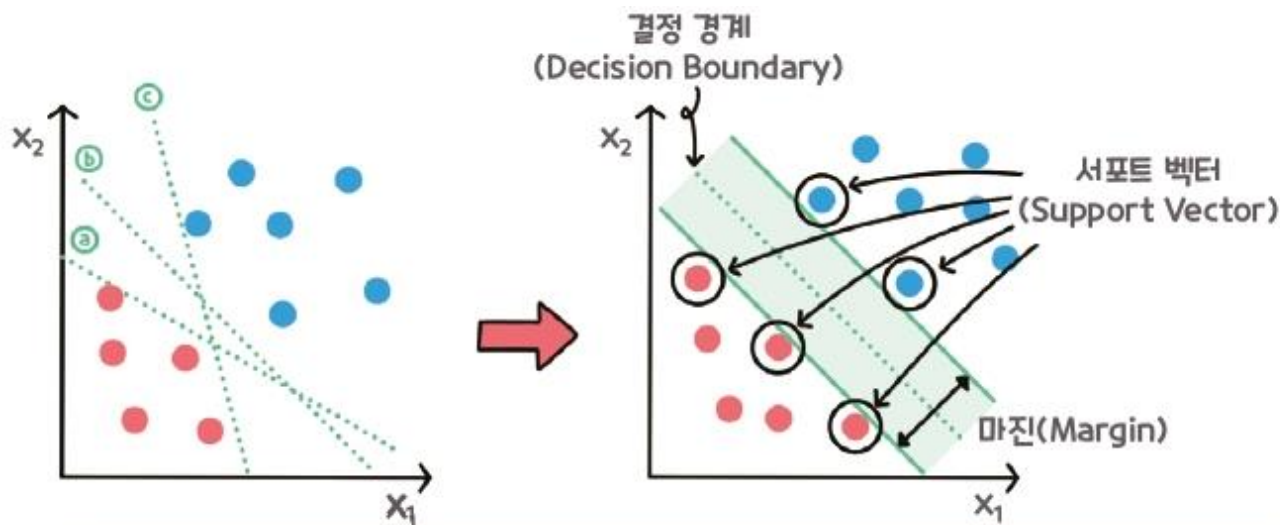
- 데이터가 어떤 범주에 속할 확률을 0~1 사이의 값으로 정해놓고, 그 확률에 따라 가능성이 더 높은 범주에 속하는 것으로 분류해 주는 지도학습 알고리즘
 - 입력 데이터가 주어졌을 때 해당 데이터의 결과가 0과 1 사이의 값을 가짐
 - 결괏값이 정해진 범주 내에서 나오므로 확률적인 의미에서 사건 발생 가능성을 예측하는 데 사용할 수 있음

로지스틱 회귀



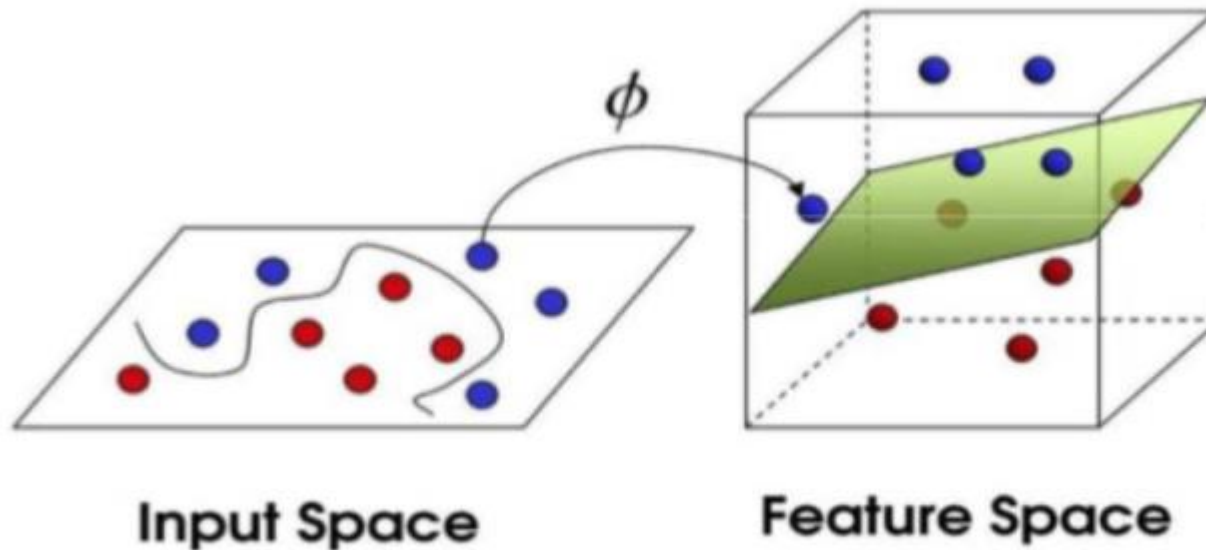
SVM, Support Vector Machine

- 두 분류 사이의 여백을 의미하는 마진을 최대화하는 방향으로 데이터를 분류
 - SVM은 마진을 극대화하는 선을 찾아 분류하므로 마진이 크면 클수록 새로운 데이터가 들어 오더라도 잘 분류할 가능성이 높아짐
 - SVM은 사용 방법이 쉽고 예측 정확도가 높다는 장점
 - 하지만 모델 구축에 시간이 오래 걸리고 결과에 대한 설명력이 떨어지는 단점



- 결정 경계(Decision Boundary) : 분류를 위한 기준선
- 서포트 벡터(Support Vector) : 결정 경계와 가장 가까운 위치에 있는 데이터
- 마진(Margin) : 결정 경계와 서포트 벡터 사이의 거리

SVM, Support Vector Machine



- 커널법(Kernel) : 비선형 분류 문제 -> 저차원의 입력 x 를 고차원의 공간의 값 $\phi(x)$ 로 변환

인공신경망

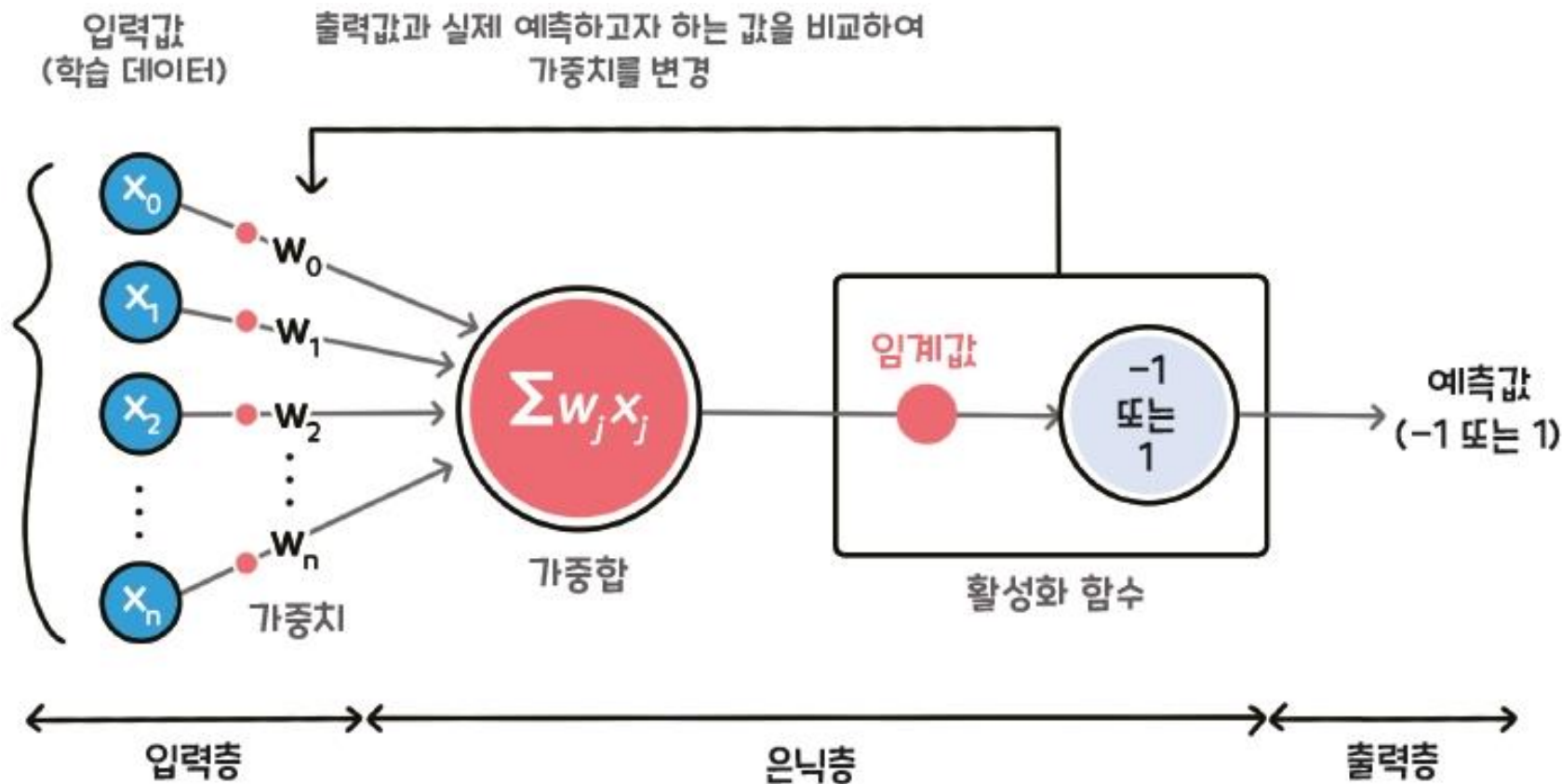
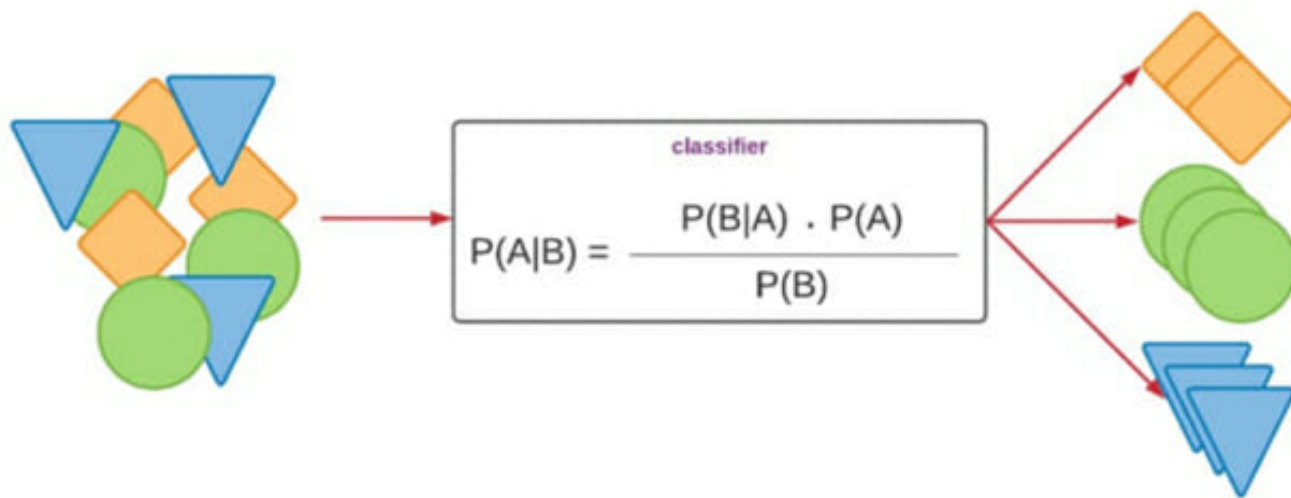


그림 9-9 인공신경망의 구조

베이지안 분류 (Bayes Classifier)

- 베이즈 이론을 기반으로 하는 확률적인 분류 알고리즘 확률 이론의 기본 원리 중 하나로, 조건부 확률을 계산하는데 사용 관찰된 이벤트들이 어느 클래스에 속할지를 분류
 - 주어진 데이터와 관련된 특징(feature)들을 기반으로 각각의 클래스에 속할 확률을 계산하고, 그 중에서 가장 높은 확률을 가진 클래스로 분류. 베이즈 이론을 사용하여 조건부 확률을 계산



■ 베이즈 이론

- $P(A|B)$: 사건 B가 일어났을 때, A 클래스에 속할 확률
- $P(A)$: 사건 A가 일어날 확률
- $P(B)$: 사건 B가 일어날 확률 = 사건 A가 발생하기 전 사건 B가 일어날 확률
- $P(B|A)$: 클래스 A에서 사건 B가 발생할 확률

베이지안 분류 (Bayes Classifier)

- Naïve Bayesian Classifier

- 광고(W_1), 자기야(W_2), 사랑해(W_3), 폭탄 세일(W_4)

$$\begin{aligned} P(SPAM|W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) &= \frac{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4|SPAM)P(SPAM)}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)} \\ &= \frac{P(W_1|SPAM)P(\neg W_2|SPAM)P(\neg W_3|SPAM)P(W_4|SPAM)P(SPAM)}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)} \end{aligned}$$

K-Nearest Neighbor (KNN)

- 가장 거리가 가까운 n 개의 데이터의 class를 조사하여 가장 많은 수에 해당하는 class로 결정

