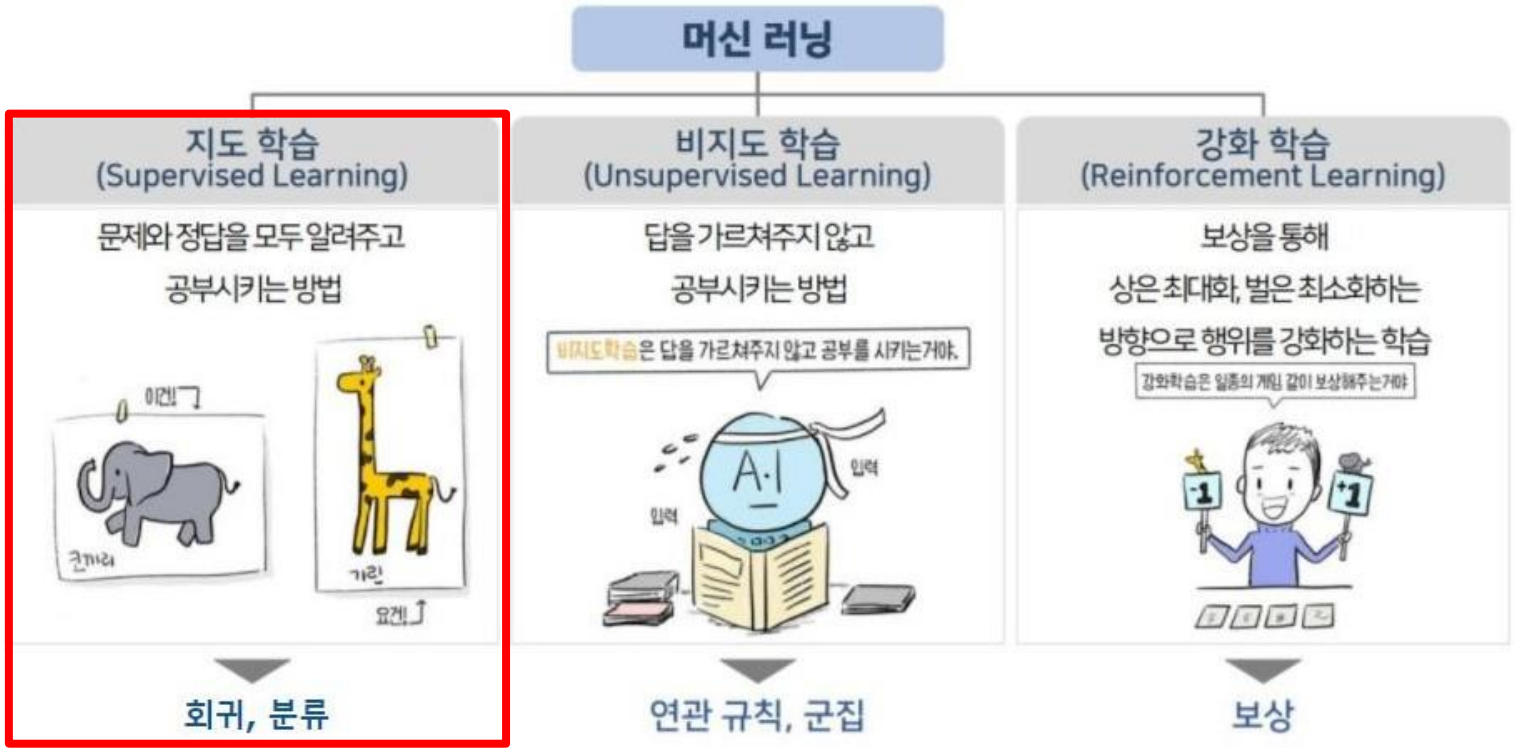
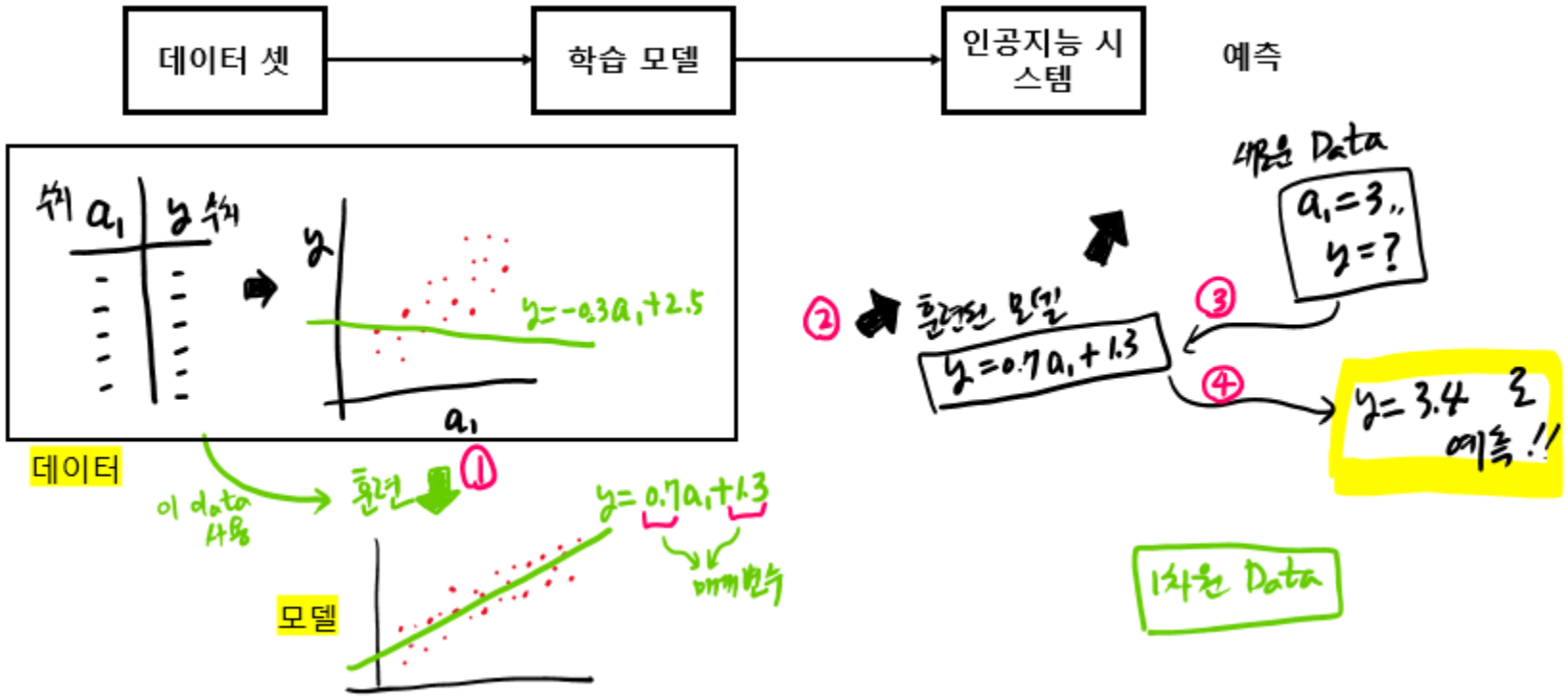
머신러닝 유형 : 지도학습(Supervised Learning)



회귀(Regression)



분류(classification)

입력 데이터를 미리 정의된 여러 클래스 중 하나로 할당하는 것

특징 feature	모델을 만들기 위해 필요한 속성
타깃 target	모델이 예측하려는 출력값
클래스 class	데이터가 분류되는 범주
레이블 label	각 입력 데이터에 대한 실제 클래스

feature

target

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

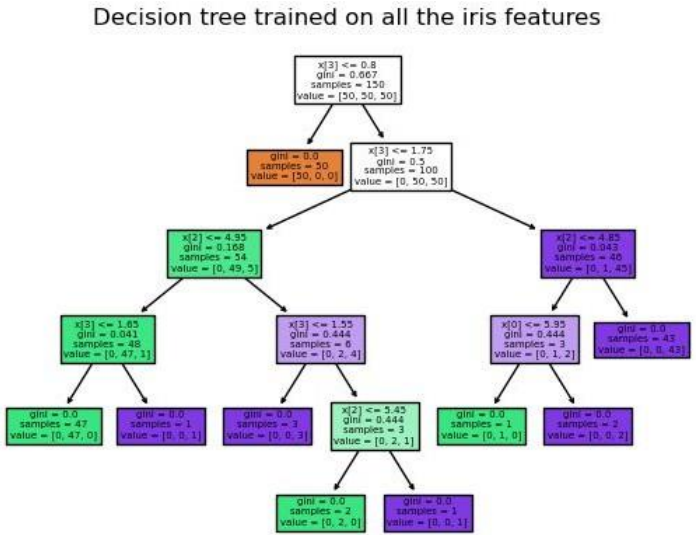
label

class : Iris-setosa, Iris-versicolor, Iris-virginica

분류 알고리즘(Classification algorithm)

결정트리 (Decision Tree)

- 데이터를 분할하여 특정 기준에 따라 분류하는 트리 구조의 모델
- 각 노드는 입력 특성의 값을 기반으로 예측을 수행
- 단말 노드(leaf node)에는 클래스 레이블이 할당됨

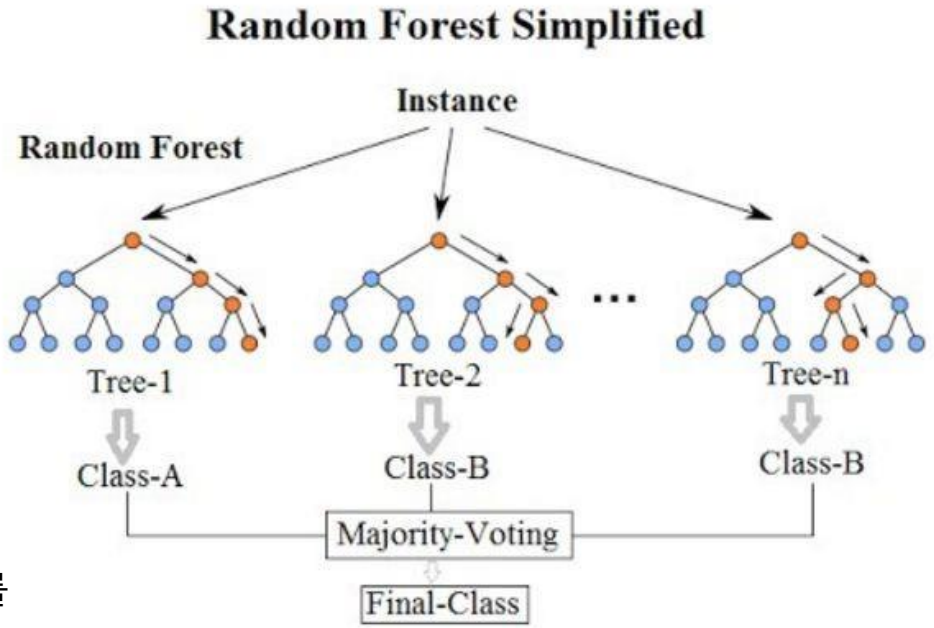


과적합(overfitting) : 결정트리는 데이터를 계속 해서 분할하게 되면 트리가 너무 세분화하고 복잡해져서 새로운 데이터에 대한 일반화 능력이 저하됨

분류 알고리즘(Classification algorithm)

랜덤 포레스트 (Random Forests)

- 여러 개의 결정 트리를 결합하여 만든 앙상블 모델
- 높은 정확도와 일반화 성능을 제공하며, 과적합 문제에 강함

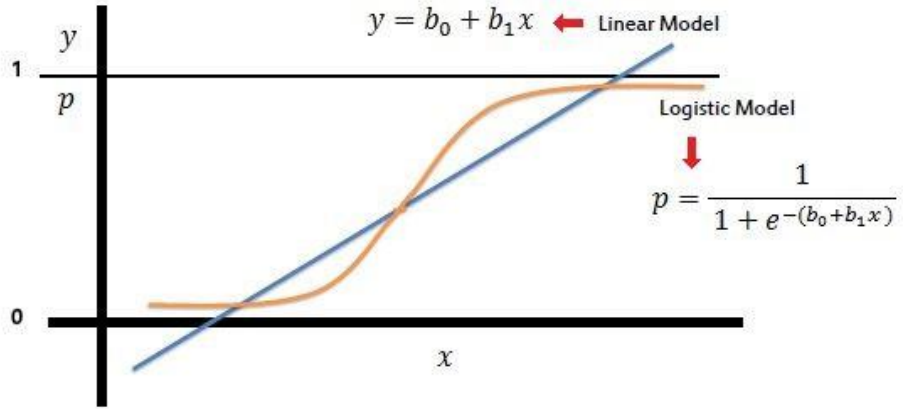


- 결정트리 : 훈련 데이터 일부만 사용해 학습 + 무작위 특징 선택 기법으로 데이터 분할
  -> 훈련 데이터에만 맞춰지는 것 방지 (과적합 감소)
- 새로운 데이터 입력하면 모든 결정트리가 해당 데이터를 분류 -> 각 결정트리가 예측한 분류결과를 다수결 표결을 통해 최종 분류 결과로 결정함

분류 알고리즘(Classification algorithm)

로지스틱 회귀 (Logistic Regression)

- 입력 변수의 선형 결합을 확률로 변환하여 데이터를 두 개의 클래스로 분류 (다중도 가능함)
- 간단하고 해석하기 쉽고, 적은 양의 데이터로도 학습 가능
- 데이터 스케일링 필요함



확률을 계산한 후 이 확률이 특정 수준(임계값, threshold value) 이상이면 1로, 그렇지 않으면 0으로 분류 / 기본 임계값 (0.5)

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

$$f(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

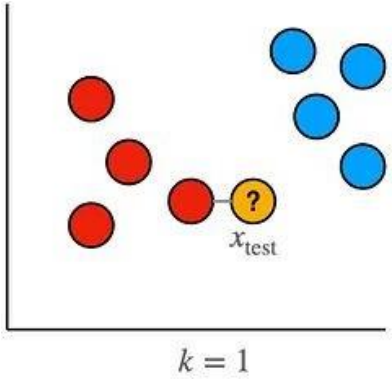
데이터 스케일링(data scaling)

- 데이터의 범위를 조정해 학습속도 향상 및 성능 향상

분류 알고리즘(Classification algorithm)

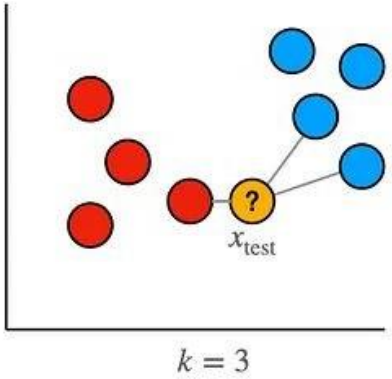
k-최근접 이웃(kNN)
(k-Nearest Neighbors)

- 새로운 데이터 포인트를 분류할 때 가장 가까운 이웃 데이터 포인트(K개)의 다수 클래스를 참조하여 분류
- 구현이 간단함
- 데이터 스케일링 필요함
- 거리측정 : 유클리디안, 맨하튼 거리측정



$k = 1$

Nearest point is red, so x_{test} classified as red

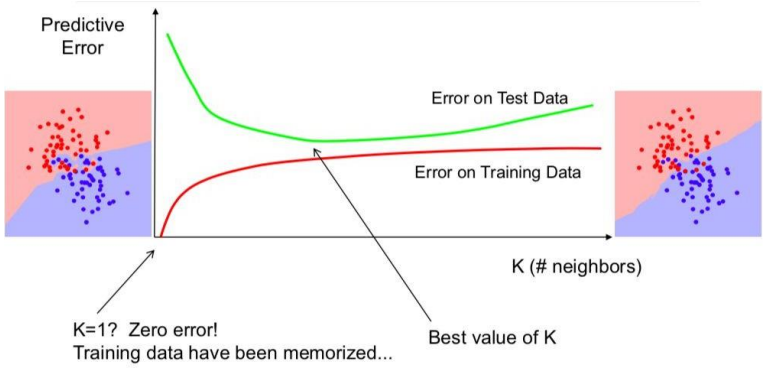


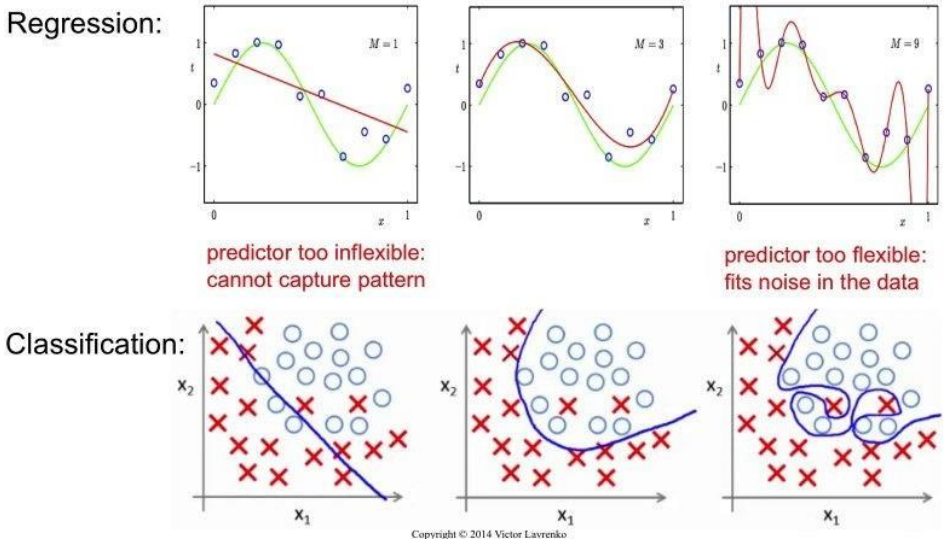
$k = 3$

Nearest points are {red, blue, blue} so x_{test} classified as blue

KNN : K

- K가 너무 작으면, 데이터의 지역적 특징을 너무 반영해 과적합(overfitting)
- K가 너무 크면, 다른 클래스의 데이터를 너무 많이 포함하게 되어 분류/회귀가 엉망이 됨 (underfitting)

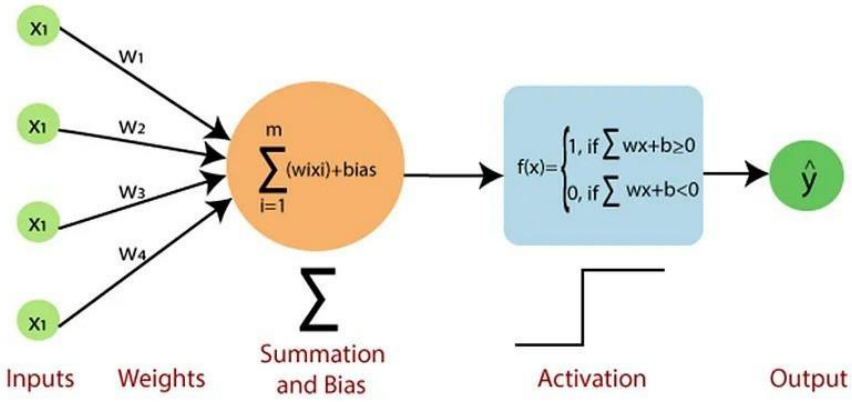




분류 알고리즘(Classification algorithm)

퍼셉트론(perceptron)

- 인공 신경망의 한 종류
- 이진 분류(binary classification) 문제 해결 (선형 분류 문제)



평가질문1

자신의 진로분야 혹은 관심 분야에서
분류, 회귀 기법을 사용하여 해결할 수 있는 문제를
각 1가지 제시하시오.

- 관심 분야:
- 분류 기법을 사용하여 해결할 수 있는 문제
- 회귀 기법을 사용하여 해결할 수 있는 문제