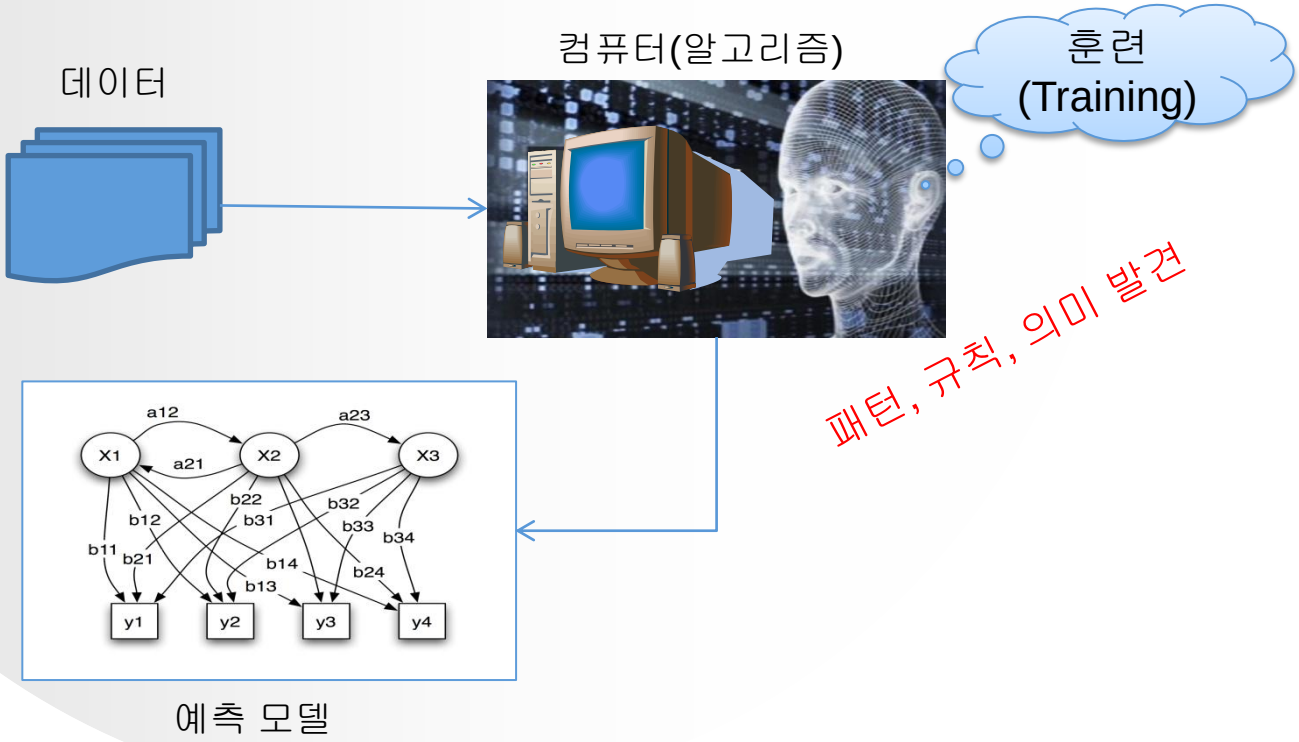
기계학습(Machine Learning)?

목차

- 기계학습(Machine Learning)?
- 학습방법에 따른 분류
- 알고리즘에 따른 분류

기계학습(Machine Learning)?

- **기계학습(Machine Learning)**
  - 인공지능(Artificial Intelligence)분야, 1950년대 연구 시작
  - 빅 데이터 분석과 Deep Learning 이슈로 최근 새롭게 조명
  - data에 내재된 패턴, 규칙, 의미 등을 알고리즘을 기반으로 컴퓨터가 스스로 학습하고, 이를 토대로 new data 결과 예측 프로그래밍

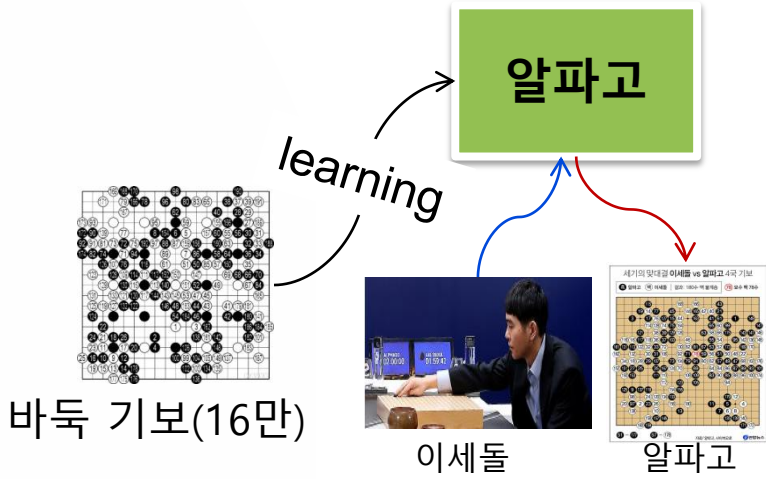


학습방법에 따른 분류

- **학습방법에 따른 기계학습 분류**
  - ➢ 지도학습((supervised learning)
    - ✓ 학습데이터가 정해짐
    - ✓ 예 : 동물 인식(cat, dog) : 레이블이 있는 상태에서 학습
  - ➢ 비지도학습((unsupervised learning)
    - ✓ 특정 레이블이 없는 상태 : 구글 뉴스(유사한 뉴스 그룹핑)
    - ✓ 레이블이 없는 데이터를 학습하여 패턴 인식 : 유사단위 그룹핑
  - ➢강화학습(reinforcement learning)
    - ✓ 지도학습 + 비지도학습
    - ✓ 기본 입출력 정보 제공, 출력 결과에 보수(reward) 정보 제공

Supervised learning

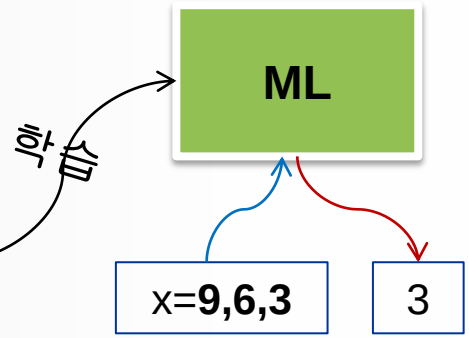
- **Supervised learning**
  - ➢ 일반적인 문제
  - ➢이미지 레이블링
  - ➢이메일 ham/spam 분류
  - ➢ 시험 점수 예측



- **Training data set**
  - ➢ 레이블이 정해진 값(y)

x	y
2,6,5	1
3,7,4	2
3,6,9	3

Training data set



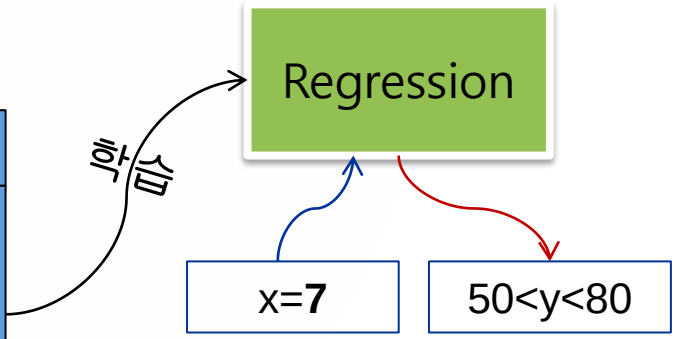
Supervised learning

- **데이터 셋 유형에 따른 Regression 분류**
  - 연속형(0~100) : Liner regression
  - pass and non-pass : Logistic regression(classification)
  - A,B,C,D,F : multi label 분류

- **Linear regression**
  - 공부시간과 점수 관계

hours	score
10	90
9	80
5	50
3	30

Training data set

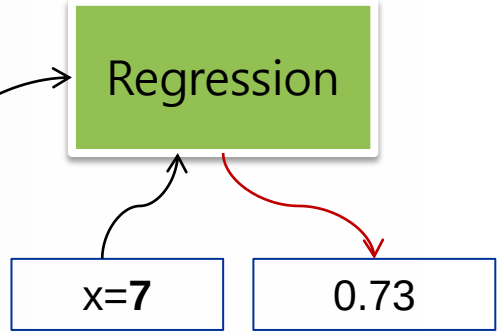


Supervised learning

- **Logistic regression(classification)**
  - ➢ 합격/불합격 분류

hours	score
10	pass
9	pass
5	fail
3	fail

학습



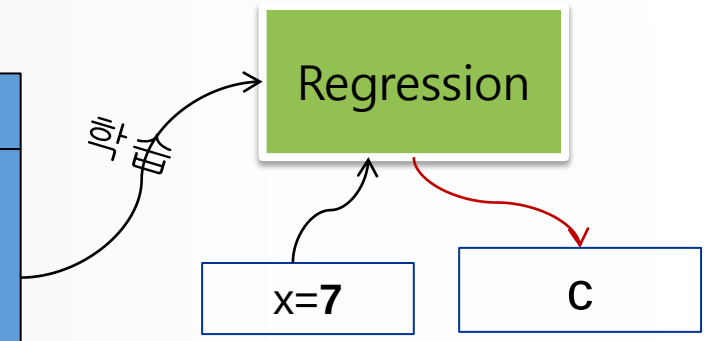
Training data set

Supervised learning

- **multi label classification**
  - 학점분류

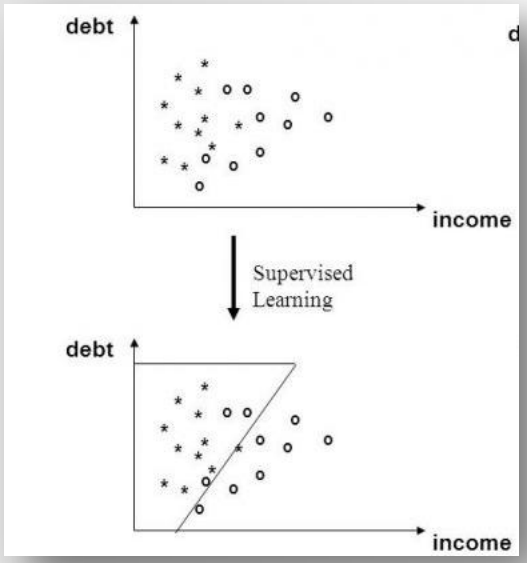
hours	score
10	A
9	B
5	D
3	F

Training data set



Supervised learning

- **multi label classification**
  - 데이터들을 정해진 몇 개의 부류(class)로 대응시키는 문제



분류 함수
- 학습 데이터를 잘 분류할 수 있는 함수
- 수학적 함수일 수도 있고, 규칙일 수도 있음

분류기(classifier)
- 분류모델 알고리즘이 적용된 함수를 이용하여 데이터를 분류하는 프로그램

Supervised learning

- **classification 모델 알고리즘**
  - 결정트리(decision tree) 알고리즘
  - K-근접이웃 (K-nearest neighbor, KNN) 알고리즘
  - 서포트 벡터 머신(Support Vector Machine, SVM)
  - 임의 숲(random forest)
  - 에이다부스트(AdaBoost)
  - 확률 그래프 모델 (probabilistic graphical model)

Supervised learning

- 혼돈(confusion) 매트릭스 : 예측치와 관측치 비교 표

		관측치 NO	관측치 YES
예측치	NO	참 부정(TN)	거짓 긍정(FP)
예측치	YES	거짓 부정(FN)	참 긍정(TP)

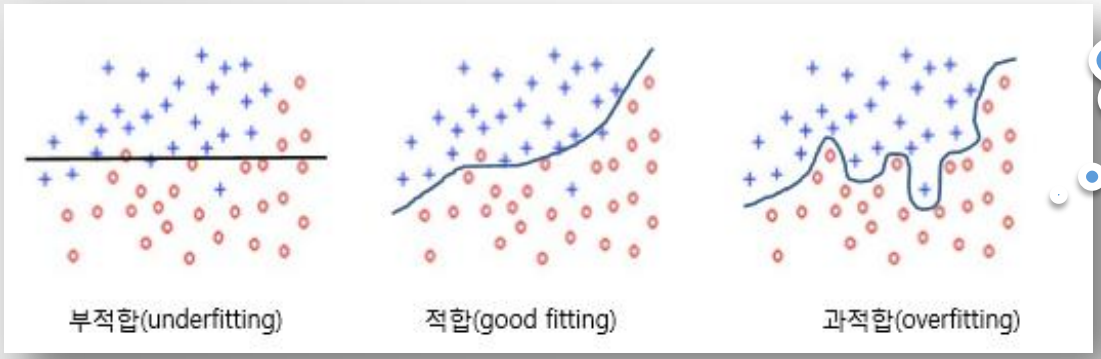
정분류율(Accuracy)= (TN+TP) / 전체관측치
오분류율(Inaccuracy) = (FP+FN) / 전체관측치
정확률(Precision) = TP / (FN + TP)
재현율(Recall) = TP / (FP + TP)

F 측정치(F measure)=2 x $\frac{\text{Precision x Recall}}{\text{Precision + Recall}}$

정분류율(Accuracy) : 알고리즘의 성능평가 척도
오분류율(Inaccuracy) : 알고리즘의 오차 비율
정확률(Precision) : 알고리즘이 Yes로 판단한 것 중에서 실제로 Yes인 비율
재현율(Recall) : 실제값이 Yes인 것 중에서 알고리즘이 Yes로 판단한 비율
F 측정치(F measure) : 정확률과 재현율을 동시에 고려하는 측정치

Supervised learning

- **과적합(overfitting)과 부적합(underfitting)**
  - 과적합
    - ✓ 학습 데이터에 대해서 지나치게 잘 학습된 상태
    - ✓ 오류나 잡음을 포함할 개연성이 큼
    - ✓ 학습되지 않은 데이터에 대해 좋지 않은 성능을 보일 수 있음
  - 부적합
    - ✓ 학습 데이터를 충분히 학습하지 않은 상태



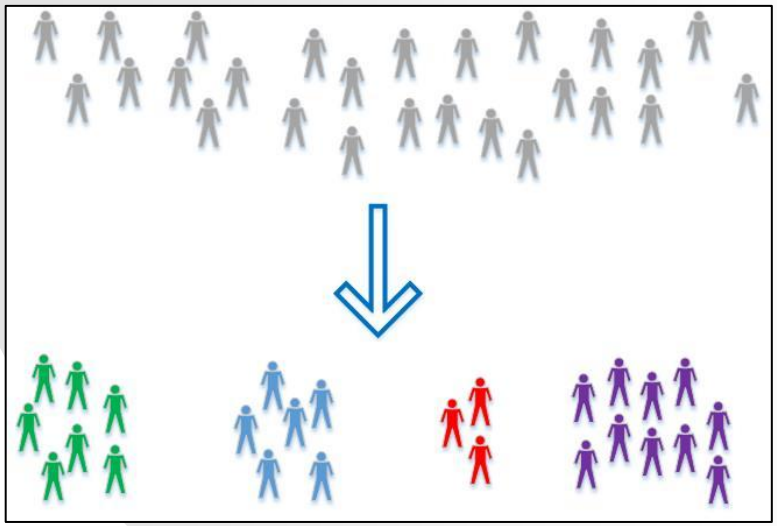
해결법:
정규화

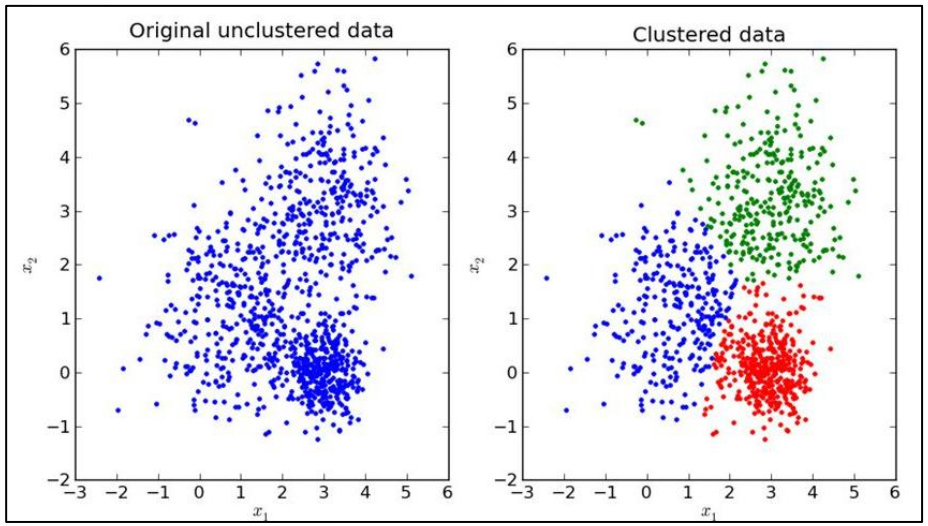
Supervised learning

- **모델 성능평가 방법**
  - K-겹 교차검증(k-fold cross-validation)
    - ✓ 전체 데이터를 k 등분
    - ✓ 각 등분을 한번씩 테스트 데이터로 사용하여, 성능 평가
    - ✓ 평가 결과 평균값 선택

UnSupervised learning

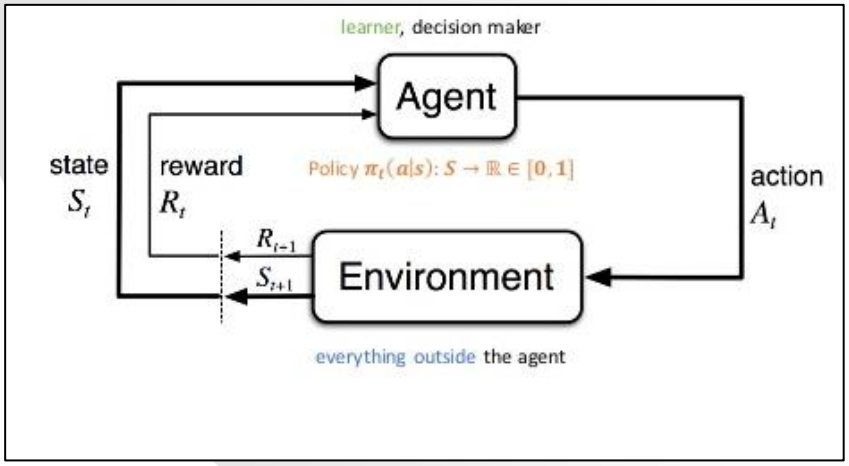
- **비지도학습**
  - 컴퓨터에게 답을 알려줄 수 없다.
  - 훈련용 데이터를 통해 함수를 추론할 수 없다
  - 컴퓨터가 알아서 분류를 하고, 의미 있는 값 제공
  - 예측 등이 아닌, 데이터가 어떻게 구성되어 있는지 밝히는데 목적

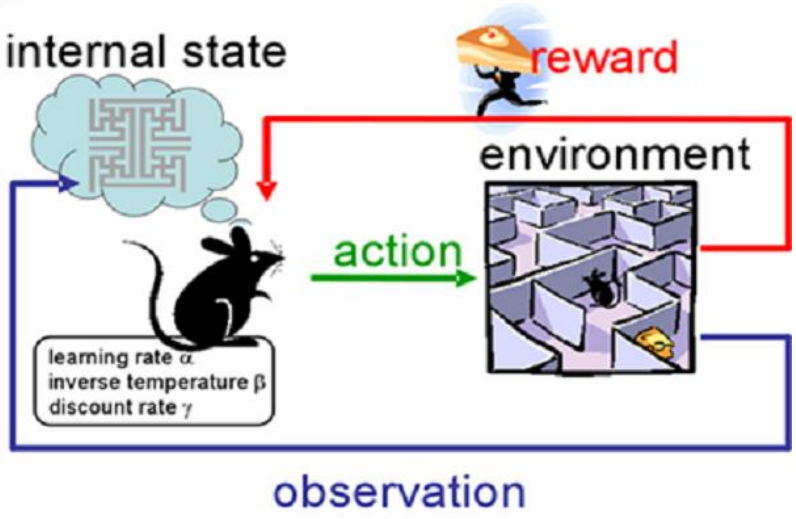




Reinforcement learning

- **강화학습**
  - 이름 그대로 자신이 수행한 행동에 대하여 보상값을 받아 조금씩 좋은 방향으로 행동을 강화시키는 학습방법
  - 현재 상태에서 최적의 행동을 계산을 통해 결정하지 않고, 여러 번의 시행착오에 기반한 경험에 의해 각 상태에서의 최적의 행동을 조금씩 학습해 나간다.
  - 스스로 경험을 통해 자율적으로 학습





알고리즘에 따른 분류

- **알고리즘에 따른 기계학습 분류**

분류		알고리즘	적용분야
지도	회귀모델	선형회귀, 회귀트리(CART)	수치예측
		로지스틱회귀	분류측
	분류모델	최근접 이웃(kNN)	분류예측
		나이브 베이즈(NB)	
		결정트리(DT)	
	블랙박스모델	앙상블(랜덤 포레스트(RM), XGBoost) 서포트 벡터 머신(SVM)	다중 용도
비지도	군집모델	K평균 군집화	군집 예측
	연관모델	연관규칙	연관 예측
강화	인공신경망	ANN, DNN, CNN, RNN	다중 용도