핸즈온 머신 러닝 1장

2021.07.21 AAI Lab. 세미나

머신 러닝이란?

- “어떤 작업 T에 대한 컴퓨터 프로그램의 성능을 P로 측정했을 때 경험 E로 인해 성능이 향상되었다면, 이 컴퓨터 프로그램은 작업 T와 성능 측정 P에 대해 경험 E로 학습한 것이다.”

-톰 미첼

머신 러닝을 사용하는 이유

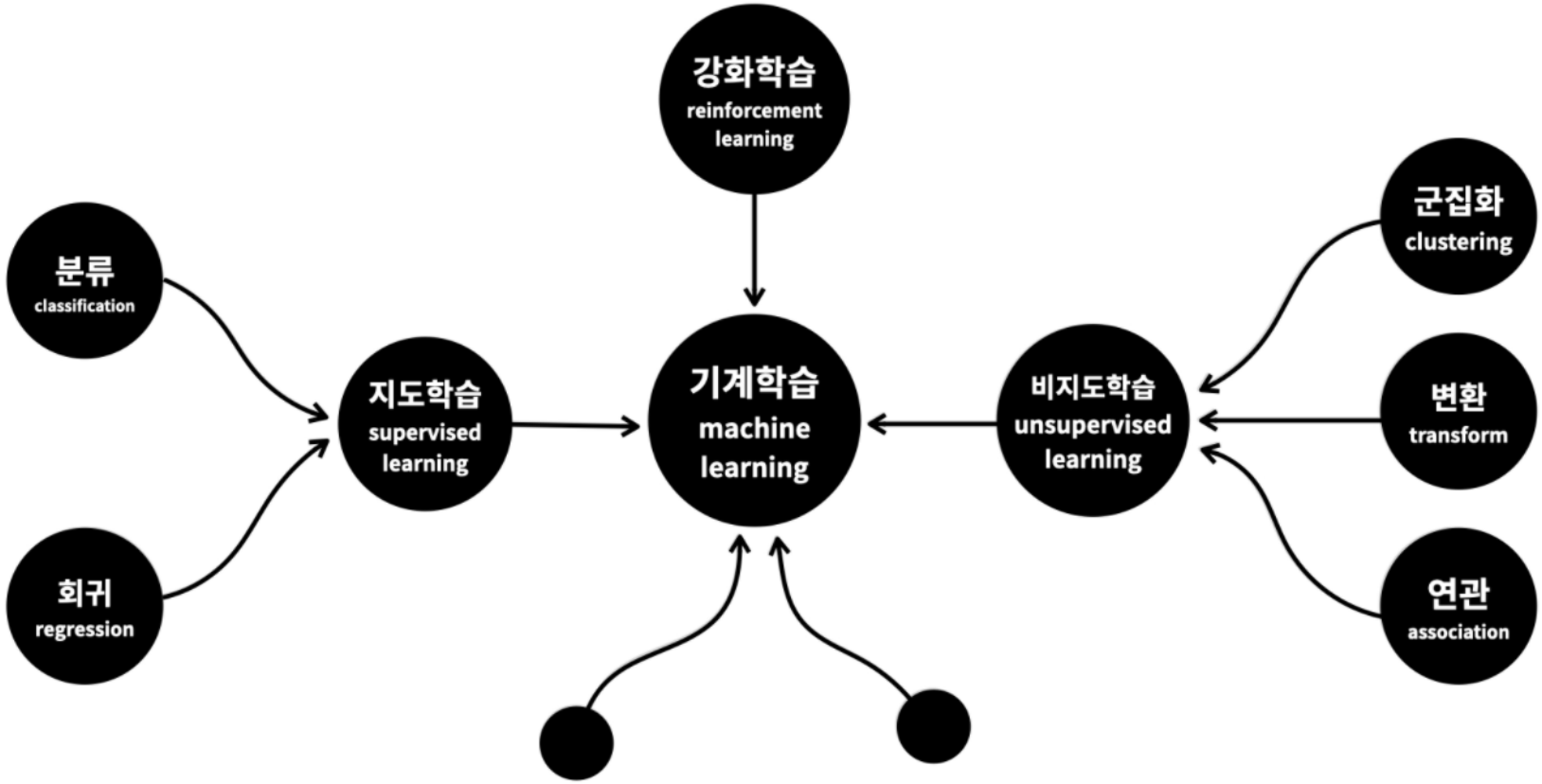
Ex) 스팸 메일 필터

- 머신 러닝 X : 직접 스팸 메일의 패턴감지 → 유지 보수 힘듦

- 머신 러닝 O : 어떤 단어와 구절이 스팸 메일을 판단하는 좋은 기준인지 **자동으로 학습** → 유지 보수 용이, 정확도 높음

머신 러닝을 통해 배운다. (data mining)

머신 러닝 시스템 종류


강화학습
reinforcement
learning
분류
classification
지도학습
supervised
learning
기계학습
machine
learning
비지도학습
unsupervised
learning
군집화
clustering
변환
transform
연관
association
회귀
regression

- 서로 다른 목적을 가진 여러 도구
- 지도학습, 비지도학습, 강화학습
- 지도학습은 분류, 회귀
- 비지도학습은 군집화, 변환, 연관
- 가장 중요하고 인기있는 것 들

지도 학습(Supervised Learning)

• 정답이 있는 문제를 해결

훈련 데이터에 정답 레이블 포함

- 분류
- 회귀

지도 학습(Supervised Learning)

- 독립변수 → 종속변수 예측
- 특성 → 타깃 예측
- 원인 → 결과 예측

- 두 변수(혹은 여러 변수) 사이의 상관관계

지도학습 - 회귀(Regression)

- 연속적인 값(숫자) 예측
- Ex) 주행거리, 연식 → 중고차 가격

지도학습 - 분류(Classification)

- 0과 1(참 거짓)으로 분류
- Discrete 한 값으로 분류

- Ex) 공부시간 → 합격 여부

비지도 학습(Unsupervised Learning)

- 기계에게 데이터에 대한 통찰력 부여
- 훈련 데이터에 정답 레이블이 없음

- 계층 군집
- 시각화
- 차원 축소
- 이상치 탐지
- 특이치 탐지

강화 학습(Reinforcement Learning)

- 더 좋은 보상을 받기 위해서 학습
- 행동 → 보상 or 벌점
- 가장 큰 보상을 받기 위해 "**정책**"이라고 불리는 최상의 전략을 **스스로 학습**

배치 학습(Batch Learning)

- 학습 시 가용한 데이터를 모두 사용

→시간과 자원을 많이 소모

→오프라인에서 수행 (오프라인 학습)

- 제품에 학습된 내용을 적용하면 더 이상의 학습 없이 사용만 된다.
- 새로운 데이터가 등장하면 새로운 데이터를 포함한 전체 데이터를 학습시킨다.

온라인 학습(Online Learning)

- 학습 시 미니배치(mini batch)라 부르는 작은 단위를 사용

→학습 단계가 빠르고 비용이 적게 든다.

→연속적으로 데이터를 받는 상황에서 적합하다.

- 문제점 : 안 좋은 데이터가 들어오면 성능이 점진적으로 감소

사례 기반 학습(Instance-Based Learning)

- 시스템이 훈련 샘플을 기억하는 것이 학습
- 학습 후 새로운 데이터와 학습한 샘플의 유사도 비교

- Ex) 스팸 메일 필터

스팸으로 지정한 메일과 유사도를 측정하여 스팸인지 아닌지 구분

모델 기반 학습(Model-Based Learning)

- 샘플들의 모델을 만들어 예측

- Ex) 선형 회귀 → loss function을 통해 모델 설정

머신 러닝의 주요 도전 과제

- 충분하지 않은 양의 데이터
- 대표성이 없는 훈련 데이터
- 낮은 품질의 데이터
- 관련 없는 특성
- 훈련 데이터 과대적합(overfitting)
- 훈련 데이터 과소적합(underfitting)

테스트와 검증

- 데이터를 **훈련세트**와 **테스트세트**로 나눔

(보통 데이터의 80%는 훈련세트, 20%는 테스트세트로 떼어놓음)

- 훈련세트로 학습을 시킨 후 테스트세트로 검증

(얼마나 잘 일반화 되었다)