파이썬 예제와 함께하는 강화학습 입문 1장

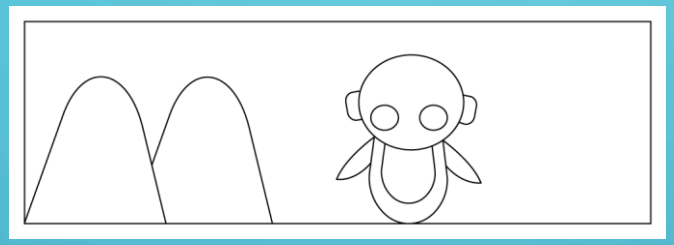
OPENAI GYM과 TENSORFLOW 실습 가이드

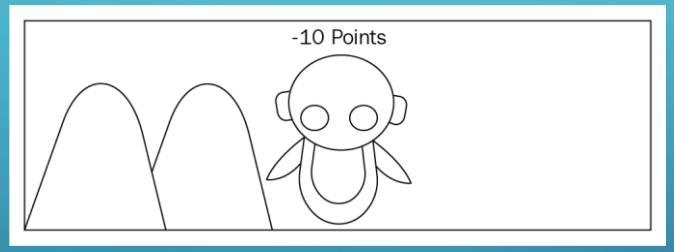
강화학습이란 ?

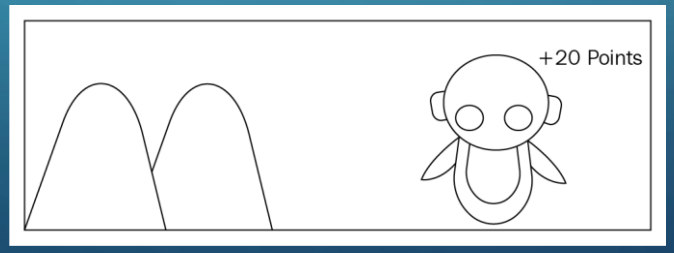
- 강아지 길들이는 방법과 비슷 (강아지한테 공 잡는 법을 가르칠때)
  - ✓강아지가 공을 잘 받으면 쿠키를 줌
  - ✓공을 못 받으면 쿠키를 주지 않음
  - ✓강아지는 쿠키를 얻을 수 있는 행동을 반복적으로 학습하여 결과적으로 공을 잘 받게 됨.

- 강화학습은 직접적으로 가르치는 대신, 행동에 따라 보상을 주는 간접방식
- 보상은 긍정적인 보상과 부정적인 보상 모두 가능
- 학습자인 에이전트(*agent*, 강아지) 는 시행착오(*trial and error*, 공 받기)를 겪으면, 긍정적인 보상(쿠키)을 받음
- 보상은 행동 이후 즉시 주어지지 않을 수 있음

로봇 에이전트에게 언덕을 피해 이동하는 법 가르치기 예시







강화 학습 에이전트의 2가지 전략

- 탐험(eplore) : 좋은 보상을 얻을 새로운 행동, 장점과 단점 ??
- 활용(exploit) : 좋은 보상을 얻을 수 있었던 이전의 행동, 장점과 단점 ??
- 탐험과 활용은 트레이드 오픈 관계

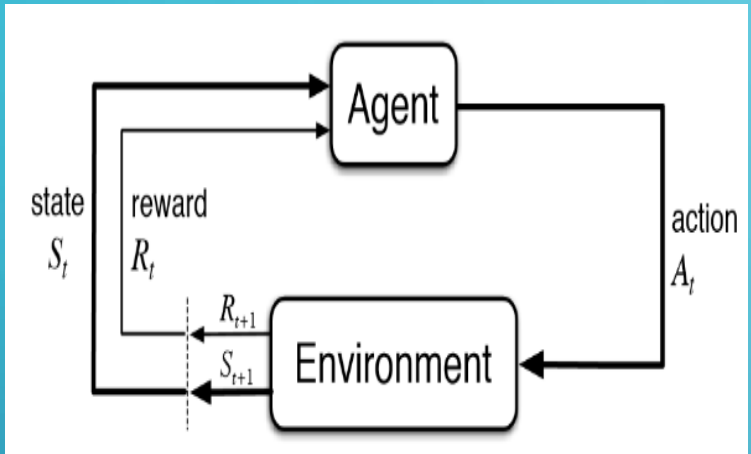
강화학습과 다른 기계학습과 다른점

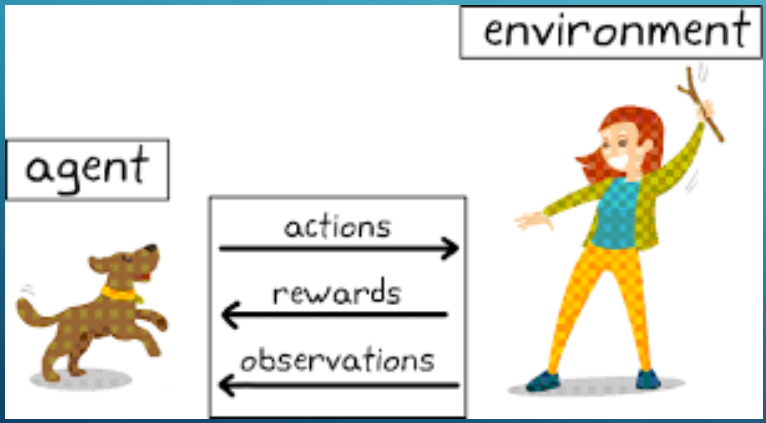
- 지도학습 : 정답이 주어진 데이터로 학습, 분류
- 비지도 학습 : 입력 데이터만을 학습에 활용하여 숨겨진 구조 학습, 군집
- 강화 학습 : 보상을 최대화하는 방향으로 학습

강화학습의 구성 요소

- 에이전트(*agent*)
  - ✓ 지능적으로 판단하고 결정하는 소프트웨어
  - ✓ 예) 슈퍼마리오 게임에선 슈퍼마리오가 에이전트에 해당
- 정책함수(*policy function*)
  - ✓ 에이전트가 어떤 행동을 할지 정의한 것
  - ✓ 정책은 기호 π 표시
  - ✓ 정책은 테이블 또는 트리 형태(complex search process)로 표현
- 가치함수(*value function*)
  - ✓ 에이전트가 특정한 상태(state)에 있을때 평가하는 것
  - ✓ 현재까지 정책에 따라 행동할때 기대할 수 있는 보상의 합
  - ✓ v(s) 로 표시
- 모델(*model*)
  - ✓ 모델 기반 학습은 행동에 따라 상태가 어떻게 달라지는지에 대한 정보(모델의 동적 특성)을 활용
  - ✓ 예) 집에서 회사까지 도착하려고 할때, 모델 기반 학습은 지도를 사용

에이전트와 환경의 인터페이스





- 미로 게임을 통해서 강화학습 이해
- 에이전트(*agent*)
  - ✓미로를 여행하는 소프트웨어 프로그램
- 환경 : 미로
- 상태 : 미로에서 현재 에이전트가 위치한 좌표
- 보상
  - ✓벽에 막히지 않으면 긍정적인 보상
  - ✓박에 막혀 목적지에 도달할 수 없게 되면 부정적 보상
- 목표 : 미로를 파악하여 목적지에 도달

강화학습 환경의 종류

- 결정적 환경 : 미로 환경, 체스게임
- 확률적 환경 : 바둑, 스타크레프트
- 완전히 관찰할 수 있는 환경 : 체스 게임
- 일부만 관찰할 수 있는 환경 : 포커 게임
- 이산환경 : 선택할 수 있는 행동의 경우의 수가 유한, 체스 게임
- 연속환경 : 선택 가능한 행동이 무한, 바둑
- 에피소딕 환경 : 순차적이지 않고 에이전트의 과거의 행동에 영향 받지 않음. 독립적
- 논에피소딕 환경 : 순차적 환경, 과거의 행동에 영향, 비독립적
- 단일 에이전트와 멀티 에이전트 환경

강화학습 플랫폼

- *OpenAI Gym*과 *Universe* : 이 플랫폼 이용
- *DeepMind Lab*
- *RL-Glue*
- *Project Malmo* : 마인크래프트 환경
- *ViZDoom* : *Doom* 게임 환경

강화학습 적용

- 교육 : 학생들에게 맞춤형 콘텐츠 제공
- 의료와 헬스케어 : 맞춤형 치료
- 제조 : 지능 로봇 학습
- 재고관리 : 구글에서 데이터선테의 전력 소모 줄이기 위해 사용
- 금융 : *JP Morgan*에서 거래 수익을 늘림
- 자연어 처리와 컴퓨터 비전