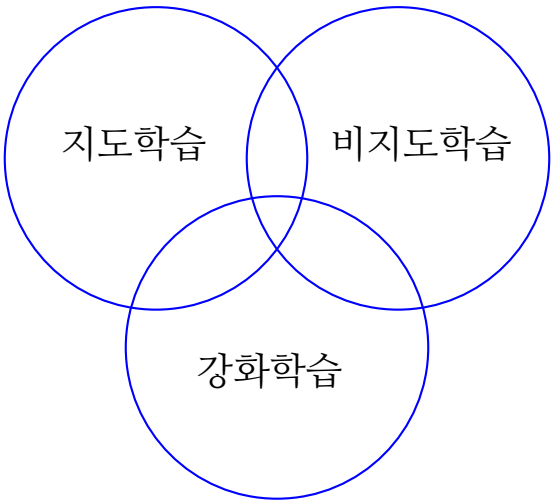
01. 강화학습 소개

김호철

Contents

---

1 머신 러닝의 구분



2 강화 학습의 특징

- 감독자는 없고 보상(reward)만 존재한다.
- 피드백이 즉각적이지 않고 지연된다.
- 시간이 정말 중요(순차적, 비 iid 데이터- independent, identically distributed:독립적이고 동일하게 분포된)
- 에이전트의 action(행동)이 수신하는 후속 데이터에 영향을 미친다.
- 강화학습의 예시
  - 헬리콥터에서 비행 스턴트 기동
  - 투자 포트폴리오 관리
  - 발전소 제어
  - 휴머노이드 로봇을 걷게 하기
  - 다양한 아타리 게임에서 인간 수준 이상으로 플레이

3 보상(Reward)

- reward R_t는 스칼라 피드백 신호
- 에이전트가 어떤 t단계에서 얼마나 잘하고 있는지 나타내는 것
- 에이전트의 임무는 누적 보상(reward)를 극대화하는것
- 강화 학습은 보상(reward) 가설을 기반으로 함

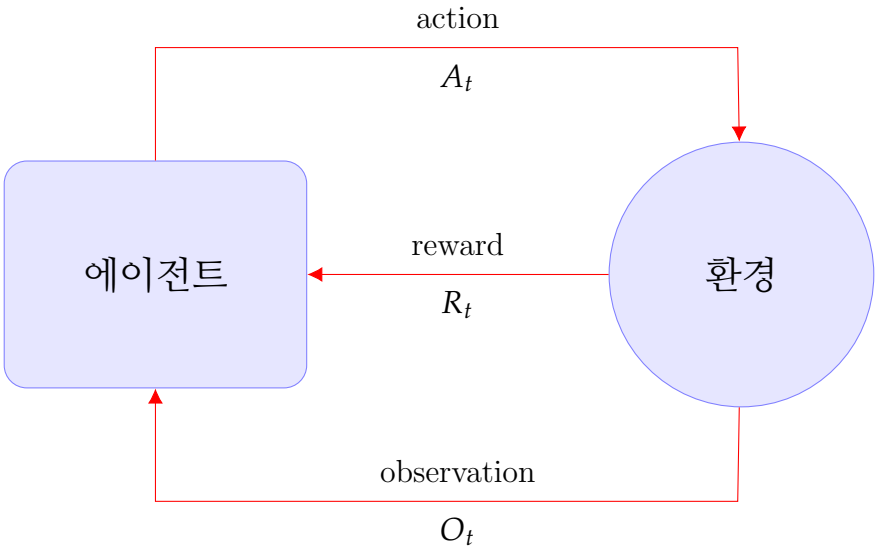
보상 가설의 정의

모든 목표는 예상되는 누적 보상(reward)을 극대화하는 것

4 순차적 의사 결정(Sequential Decision Making)

- 목표 : 미래의 총 reward를 극대화하기 위한 action을 선택
- 액션은 장기적인 결과가 될 수 있음
- 보상(reward)은 지연 될 수 있다
- 장기적인 보상(reward)을 얻기 위해 즉각적인 보상을 희생하는 것이 더 나을 수 있음

5 에이전트와 환경



- 에이전트는 각 단계 t에서, 액션 A_t를 실행하고, 관측치 O_t를 받고, 스칼라 보상 R_t를 받는다.
- 환경은 각 단계 t에서, 액션 A_t를 받고, 관측치 O_{t+1}를 내보내고, 스칼라 보상 R_{t+1}를 내보낸다.

6 상태(State)

6.1 히스토리(History)와 상태(State)

- 히스토리는 observation, action, reward들의 순서

$$H_t = O_1, R_1, A_1, ..., A_{t-1}, O_t, R_t \tag{1}$$

- 즉, 시간 t 까지 관측가능한 모든 변수들
- 히스토리에 의해, 에이전트가 액션을 선택하고, 환경이 observations, rewards를 선택한다.
- 상태는 다음에 일어날 일을 결정하는데 사용되는 정보이다.
- 상태는 히스토리의 함수이다.

$$S_t = f(H_t) \tag{2}$$

6.2 환경 상태(Environment State)

- 환경 상태 S_t^e는 환경 내부에서 사용되는 표현이다.
- 환경에서 다음 관측/보상을 선택하는데 사용되는 모든 데이터
- 환경 상태는 일반적으로 에이전트에게는 보이지 않는다.
- S_t^e가 보이더라도 관련없는 정보가 포함될 수 있다.

6.3 에이전트 상태(Agent State)

- 에이전트 상태 S_t^a는 에이전트의 내부에서 사용되는 표현이다.
- 에이전트에서 다음 액션을 선택하는데 사용되는 모든 정보
- 강화학습 알고리즘들에서 사용되는 정보들이다.
- 히스토리의 어떤 함수가 될 수도 있다.

$$S_t^a = f(H_t) \tag{3}$$

6.4 정보 상태(Information State)

- 정보 상태(마르코프 상태)는 히스토리에서 모든 유용한 정보들이다.

정의

상태 S_t는 다음의 경우에만 마르코프이다.

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, ..., S_t]$$

- 미래는 현재가 주어지면 과거와는 독립적이다.

$$H_{1:t} \rightarrow S_t \rightarrow H_{t+1:\infty} \tag{4}$$

- 상태를 알게되면, 히스토리는 버려진다.
- 상태는 미래에 대해 충분한 통계이다.
- 환경 상태 S_t^e는 마르코프이다.
- 히스토리 H_t는 마르코프이다.

6.5 완전 관찰 가능 환경(Fully Observable Environments)

- 에이전트가 환경의 상태를 직접 관찰하는 형태

$$O_t = S_t^a = S_t^e \tag{5}$$

- 에이전트 상태=환경 상태=정보 상태(마르코프)
- 이것이 의사 결정 프로세스(MDP:Markov decision process)

6.6 부분 관찰 가능 환경(Partially Observable Environments)

- 에이전트가 환경을 간접적으로 관찰하는 형태
- 주식 거래 에이전트는 현재 가격만 관찰하고, 포커 게임 에이전트는 오픈된 카드만 관찰한다.
- 현재의 에이전트 상태 ≠ 환경 상태
- 이것은 부분 관찰 가능 마르코프 의사 결정 프로세스 (POMDP)
- 에이전트는 자체적인 상태 표현 S_t를 구성해야함

7 RL 에이전트

- RL 에이전트는 다음 중 하나 이상을 가진다.
  - 정책(Policy) : 에이전트에서 액션(action) 함수
  - 가치 함수(Value Function) : 각 상태 또는 행동이 얼마나 좋은지
  - 모델(Model) : 에이전트에서 보는 환경에 대한 표현

7.1 정책(Policy)

- 정책은 에이전트의 행위이다.
- 상태를 입력으로 액션을 출력으로 주는 맵이다.
- 결정적(Deterministic) 정책 : $a = \pi(s)$
- 통계적(Stochastic) 정책 : $\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$

7.2 가치 함수(Value Function)

- 가치 함수는 미래 보상에 대한 예측이다.
- 상태들의 좋고 나쁨을 평가하는데 사용됨
- 그러므로 액션들 중에서 선택한다.

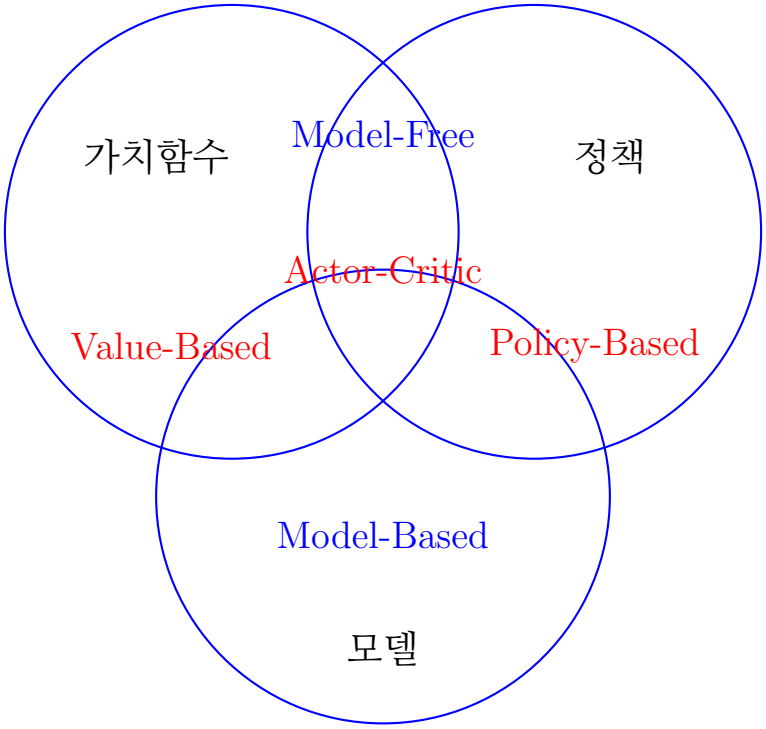
$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \ldots | S_t = s] \tag{6}$$

7.3 모델(Model)

- 모델은 환경이 다음에 무엇을 할지 예측한다.
- P는 다음 상태를 예측한다.
- R는 다음 보상을 예측한다.

$$\begin{aligned} P^a_{ss'} &= \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a] \\ R^a_s &= \mathbb{E}[R_{t+1} | S_t = s, A_t = a] \end{aligned} \tag{7}$$

7.4 RL 분류



8 RL 문제

8.1 Learning과 Planning

- 순차적 의사 결정의 두 가지 근본적인 문제

- Learning : 환경을 모르는 상태에서, 에이전트가 환경과 상호작용하여, 정책을 개선
- Planning : 환경을 아는 상태에서, 에이전트는 외부와의 상호작용 없이, 모델을 사용하여 정책을 개선(심의, 추론, 성찰, 숙고, 생각, 검색)

8.2 Exploration(탐색)과 Exploitation(사용)

- 강화학습은 시행 착오 학습이다.
- 에이전트는 환경을 탐색하는데, 보상을 잃지않고, 좋은 정책을 발견해야 한다.
- Exploration(탐색)은 환경에서 더 많은 정보를 찾고,
- Exploitation(사용)은 획득한 정보를 사용하여 보상을 극대화 한다.
- 강화학습에서 Exploration(탐색)과 Exploitation(사용) 정책이 매우 중요하다.

8.3 예측(Prediction)과 제어(Control)

- 예측(Prediction)은 정책(policy)이 주어졌을때 미래를 평가하는것, 가치 함수(value function)을 잘 학습 시키는 것 .
- 제어(Control) 문제는 미래를 최적화하는것, 최고의 정책(policy)을 찾는 문제

9 DRL(Deep Reinforcement Learning) 맵