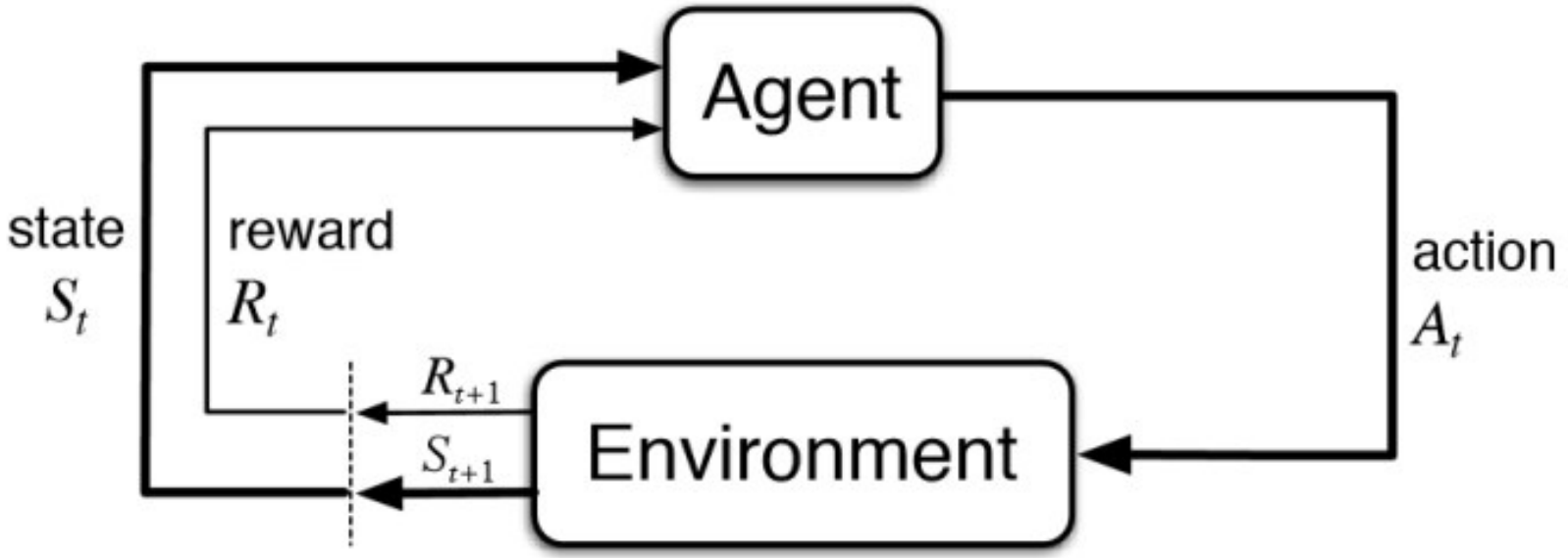
강화학습 기초

이제희

강화학습이란?

강화학습은 현재의 상태에서 어떤 행동을 취하는 것이 최적인지 학습하는 것
이는 보상을 최대화하는 방향으로 학습



마르코프 의사 결정 과정(MDP)

- 마르코프 과정

다음의 상태는 현재 상태에 의해서만 결정

- 마르코프 의사 결정 과정

강화학습은 순차적으로 행동을 계속 결정해야 하는 문제를 푸는 것
이 문제를 수학적으로 표현한 것이 MDP

상태 & 행동(State & Action)

- 상태(state)

에이전트가 관찰 가능한 상태의 집합

$$S_t = s$$

- 행동(action)

의사 결정을 통해 취할 수 있는 행동

$$A_t = a$$

보상(Reward)

- 보상

에이전트가 한 행동에 대한 환경의 피드백

$$R_s^a = E[R_{t+1}|S_t = s, A_t = a]$$

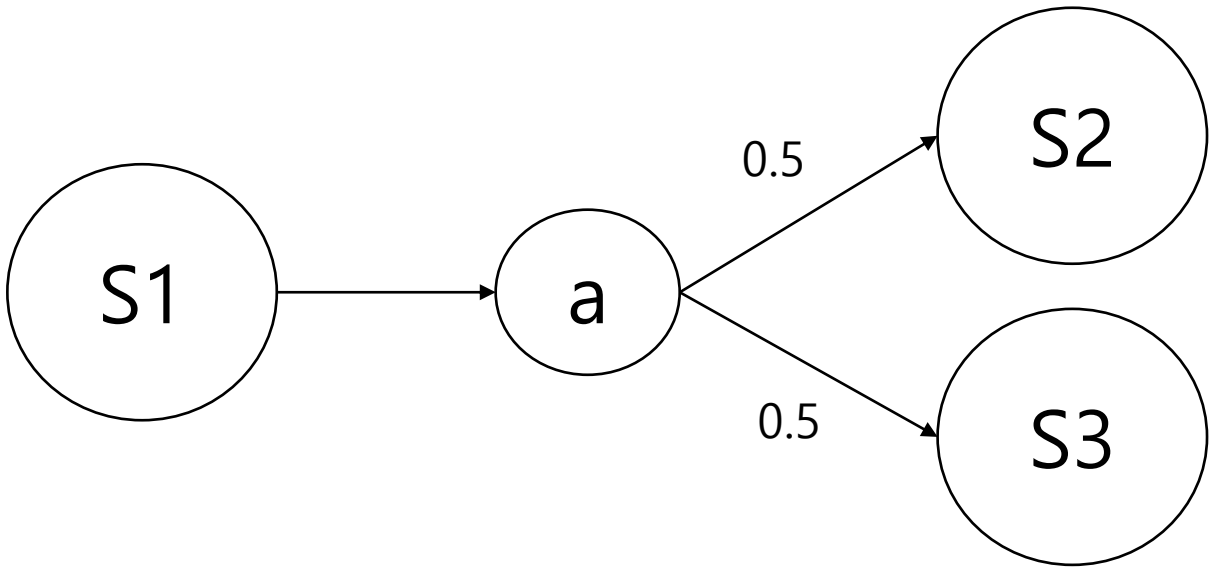
현재 환경 이후의 시간에 보상을 받음

상태 변환 확률(State transition probability)

- 상태변환확률

상태 s 에서 행동 a를 했을 때 다음 상태 s′로 갈 확률

$$P_{ss'}^{a} = P[S_{t+1} = s' | S_t = s, A_t = a]$$



감가율(Discount factor)

- 감가율

미래에 받은 보상을 현재의 시점에서 고려할 때 감가하는 비율

$\gamma \in [0,1]$ 감가율 : $0 \sim 1$ 사이

$\gamma^{k-1} R_{t+k}$ 현재의 시간 t로부터 k만큼 지난 후 받을 보상의 현재 가치

정책(Policy)

- 정책

에이전트가 각 상태에서 하는 행동에 대한 정보를 나타내는 것
상태 s에서 행동 a를 선택할 확률로 정의

$$\pi(a|s) = P[A_t = a|S_t = s]$$

가치함수(상태 가치함수)

- 가치함수

특정 상태에서의 반환값들의 기댓값(반환값 : 보상으로 생각)

$$G_t(\text{반환값}) = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$v(s) = E[G_t | S_t = s]$$

Q함수(행동 가치함수)

• Q함수

특정 상태 s에서 특정 행동 a를 취했을 때 받을 반환값에 대한 기댓값

$$q_\pi(s,a) = E_\pi[G_t|S_t = s, A_t = a]$$

PI를 사용하는 이유는 정책에 따라 행동의 가치를 평가하기 위해서