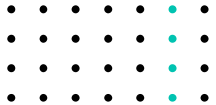
03

강화학습 기본 알고리즘

1. 마르코프 결정 과정

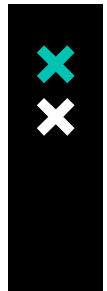
마르코프 결정 과정 기본개념



마르코프 결정 과정(Markov Decision Process)

- 마르코프 결정 과정(MDP: Markov Decision Process)은 마르코프 보상 과정(MRP: Markov Reward Process)에 행동(A: Action)과 정책(π: Policy)이 추가된 개념
- MDP = MRP + Action + Policy

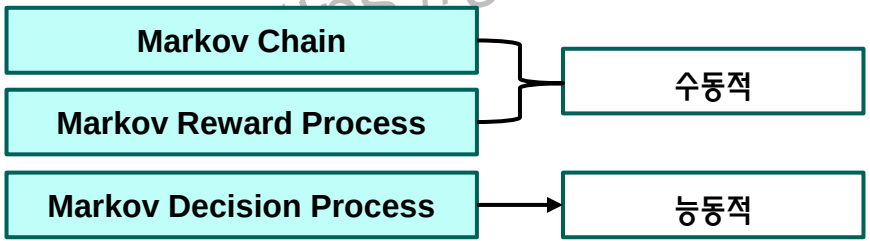
Markov Chain	확률
Markov Reward Process	가치
Markov Decision Process	가치, 정책

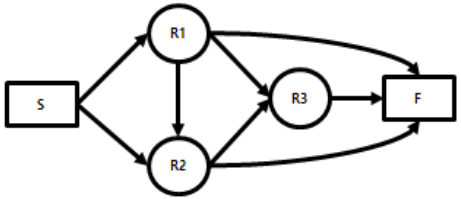


마르코프 결정 과정 기본개념

에이전트(Agent)

- 행위자, 어떤 행동을 하는 주체
- MDP에서 에이전트는 정책(π)에 따라 행동(Acton)을 하며 상태(State)는 에이전트가 취한 행동과 상태 전이 확률(P)에 따라 바뀌게 된다.



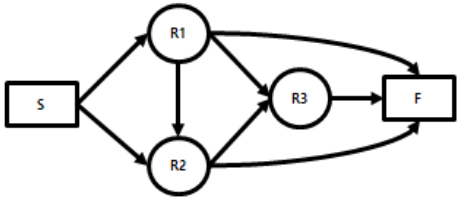


마르코프 결정 과정 정책

정책(π Policy)

- 행동을 선택하는 확률(상태전이 매트릭스와 같은 형태)
- 4가지 종류의 행동이 있다면 에이전트가 한 상태에서 각각의 행동을 할 확률의 합은 1이 되어야 한다.

$$\pi = P[A_t = a \mid S_t = s]$$



마르코프 결정 과정 구성요소

MDP 구성요소

- S : 상태(State)의 집합
- ***P***: 상태 전이 매트릭스

  $P^{a}_{ss'} = \mathbf{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$
- ***R***: 보상 함수

  $R^{a}_{s} = \mathbf{E}[R_{t+1} \mid S_t = s, A_t = a]$
- γ : 감가율

  $\gamma \in [0, 1]$
- A : 행동(Action)의 집합
- π : 정책 함수

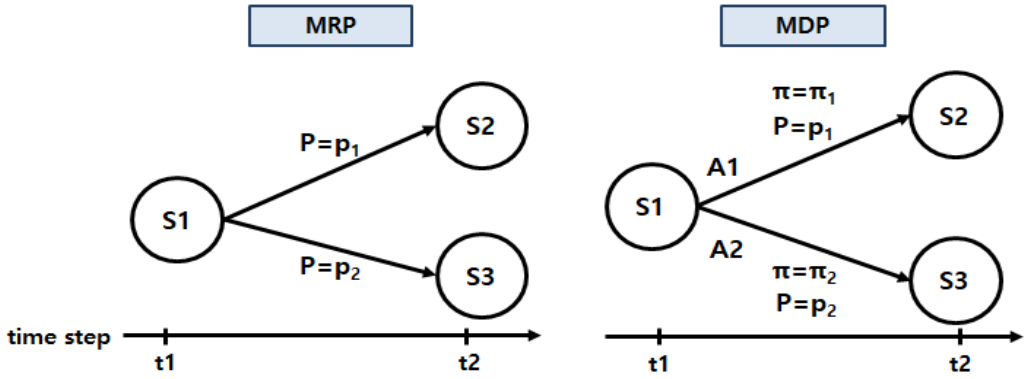
MRP

*행동(Action)
에이전트의 행동, 다음 상태에 영향을 미치는 행위

*정책(π Policy)
행동을 선택하는 확률(상태전이 매트릭스와 같은 형태)

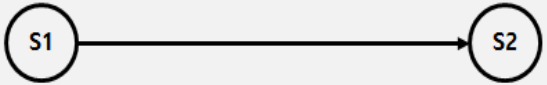
마르코프 결정 과정 사례

MRP와 MDP 비교사례



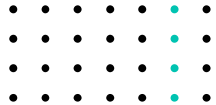
t1에서 S1일 때 t2에서 S2 확률

π_1 (0.4)+ π_2 (0.6) = 1 p_1 (0.7)+ p_2 (0.3) = 1



MRP	P1 = 0.7
MDP	π1 * p1 + π2 * p1= 0.4*0.7 + 0.6*0.7 = 0.28 + 0.42 = 0.7

마르코프 결정 과정 정책을 고려

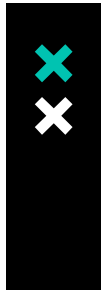


정책을 고려한 상태전이매트릭스와 보상함수

$$P_{ss'}^{\pi} = \sum_{a \in A} \pi(a|s) P_{ss'}^{a}$$

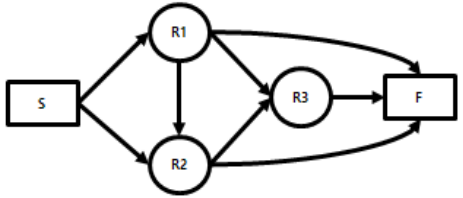
$$R_{s}^{\pi} = \sum_{a \in A} \pi(a|s) R_{s}^{a}$$

*모두 기댓값의 형태로 표현



마르코프 결정 과정 상태가치함수

MDP에서의 상태가치함수



MRP

$$v(s) = \mathbf{E}[\, R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s \,] \quad ①$$

$$= R_{t+1} + \gamma \mathbf{E}[\, v(S_{t+1}) \mid S_t = s \,] \quad ②$$

$$= R_{t+1} + \gamma \sum_{s' \in S} P_{ss'} v(s') \quad ③$$

*상태가치함수
환경 전체에 대한 가치를 측정하는 것

$$\mathbf{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + .. \mid S_t = s \,]$$

MDP

$$v_\pi(s) = \mathbf{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s \,] \quad ①$$

$$= \sum_{a \in A} \pi(a|s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')\right) \quad ②$$

$$= \underbrace{\sum_{a \in A} \pi(a|s) R_s^a}_{③\text{-}1} + \underbrace{\gamma \sum_{a \in A} \pi(a|s) \sum_{s' \in S} P_{ss'}^a v_\pi(s')}_{③\text{-}2} \quad ③$$

$$v(s) = v_\pi(s)$$

마르코프 결정 과정 행동가치함수

행동가치함수

- Q함수, Action Value Function
- 행동에 따른 가치를 평가하는 함수
- 선택할 수 있는 여러 가지 행동 중에 하나를 선택했을 때의 가치를 계산하는 함수

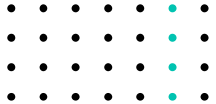
$$v_\pi(s) = \mathbf{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t=s]$$

$$= \underbrace{\sum_{a\in A}\pi(a|s)}_{①\text{-}1} R_s^a + \gamma \underbrace{\sum_{a\in A}\pi(a|s)}_{①\text{-}2} \sum_{s'\in S} P_{ss'}^a v_\pi(s') \quad ①$$

$$q_\pi(s, a) = \mathbf{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t=s, A_t=a] \quad ②$$

$$= R_s^a + \gamma \sum_{s'\in S} P_{ss'}^a \pi(s', a') q_\pi(s', a') \quad ③$$

마르코프 결정 과정 행동가치함수



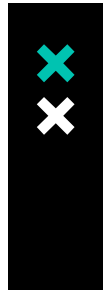
행동가치함수와 상태가치함수 관계

- Q함수, Action Value Function
- 행동에 따른 가치를 평가하는 함수

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) q_\pi(s,a) \quad ①$$

$$q_\pi(s, a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s') \quad ②$$

$$q_\pi(s, a) = \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t=s, A_t=a]$$
$$= R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \pi(s', a') q_\pi(s', a')$$



마르코프 결정 과정 최적가치함수

최적가치함수

- 최적상태가치함수(v*(s): Optimal State Value Function)
  여러 가지 정책을 따르는 상태가치함수가 있을 때 가치를 최대로 하는 정책을 따르는 상태가치함수
- 최적행동가치함수(q*(s,a): Optimal Action Value Function)
  다양한 정책을 따르는 행동가치함수 중에서 가치를 최대로 하는 정책을 따르는 행동가치 함수

$$v^*(s) = \max_{\pi} v_\pi(s) \quad ①$$

$$q^*(s,a) = \max_{\pi} q_\pi(s,a) \quad ②$$

마르코프 결정 과정 최적정책

최적정책

- 최적의 가치를 얻도록 행동할 수 있게 만드는 정책이 바로 최적정책
- 정책이란 행동을 선택할 수 있는 확률이기 때문에 값이 크다는 얘기는 확률이 높다는 얘기

$\pi^* : \pi^* \geq \pi, \forall\pi$ ①

$v_{\pi^*}(s) = v^*(s)$ ②

$q_{\pi^*}(s) = q^*(s)$ ③

$$\pi^*(a|s) = \begin{cases} 1 & \text{if } a = \operatorname{argmax}_{a \in A} q*(s,a) \quad ① \\ 0 & otherwise \quad ② \end{cases}$$

수학기호 ∀는 임의의 또는 전체의 의미를 가지고 있다. ∀π란 모든 정책에 대해 해당한다는 의미를 가지고 있다. argmax함수(x)는 조건을 만족하는 함수의 값을 가장 크게 만드는 x를 찾는 것이다.

마르코프 결정 과정 용어정리

정책평가와 제어

정책평가 Policy Evaluation	상태가치함수 계산
정책제어 Policy Control	정책변경

마르코프 결정 과정 용어정리

모델기반과 모델프리

모델 기반 Model Based	환경에 대한 모든 정보 알고 있는 경우
모델 프리 Model Free	환경에 대한 일부 정보만 알고 있는 경우

정 리

마르코프 결정 과정

에이전트와 정책

MDP에서 상태가치함수

MDP에서 행동가치함수

최적가치함수

최적정책