03

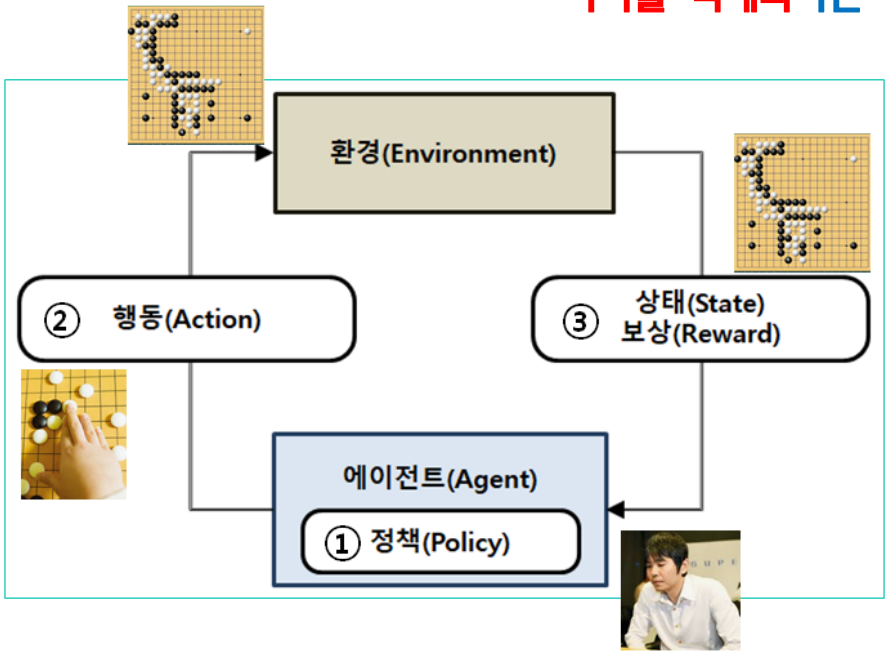
강화학습 기본 알고리즘

2. 다이나믹 프로그래밍

MDP 다시 보기 기본개념

강화학습 MDP

가치를 극대화하는 정책을 찾는 것



- 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

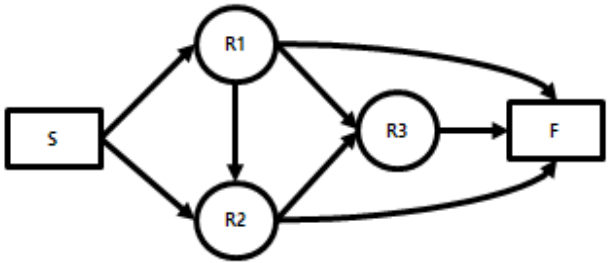
  $\gamma \in [0, 1]$
- A : 행동(Action)의 집합
- π : 정책 함수

MDP 다시 보기 가치계산

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

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

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



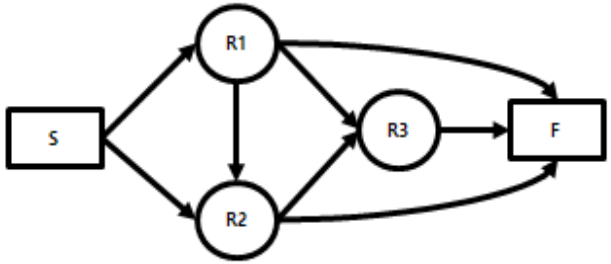
$$q_\pi(s, a) = E_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t=s, A_t=a] \quad ②$$

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

MDP 다시 보기

정책결정

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



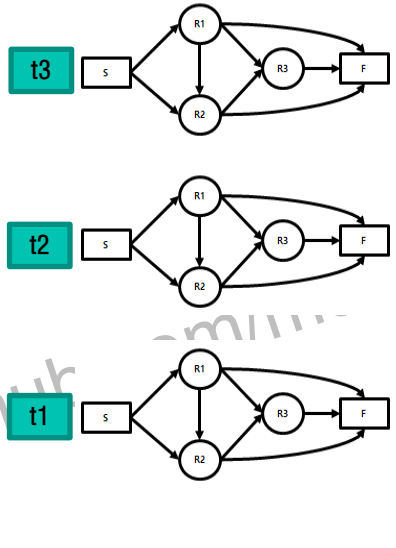
다이나믹 프로그래밍 기본개념

기본개념

- 동적계획법, 환경에 대한 모든 정보를 알고 있는 모델기반(Model Based) 방법론
- 모델기반 환경에서 사용하는 MDP 해결 방법

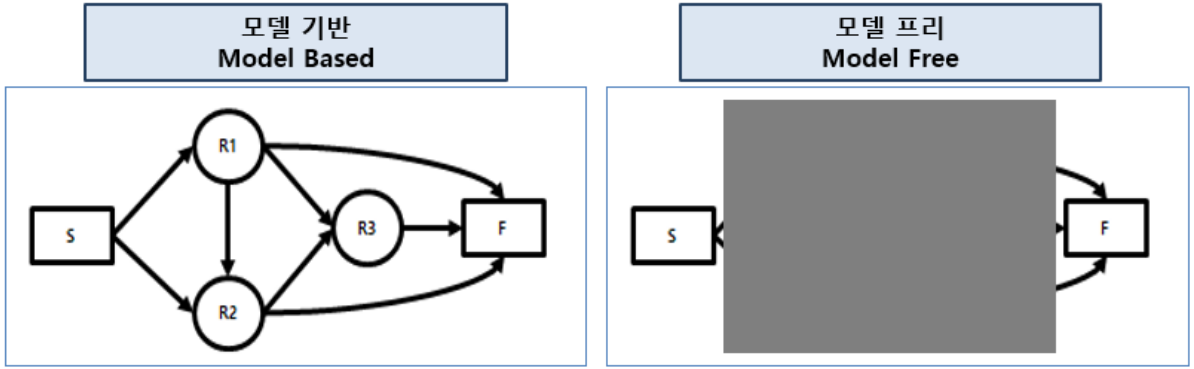
- 정책평가
  (1) 정책을 고정하고 처음 타임스텝과 뒤 따르는 스텝들에 대한 가치를 각각 구해서 합산
  (2) 마지막 타임스텝까지 반복 수행
  (3) 현재 타임스탭의 가치를 업데이트

- 정책제어
  (1)계산한 가치함수를 사용해서 탐욕적(greedy)으로 정책을 선택해서 현재 정책을 갱신(update)



다이나믹 프로그래밍 모델기반 vs 모델프리

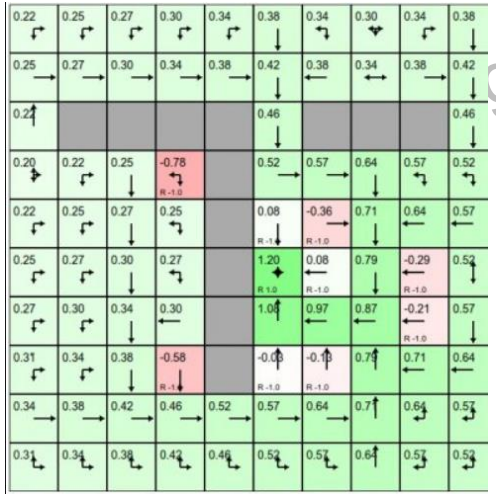
모델기반 vs 모델프리

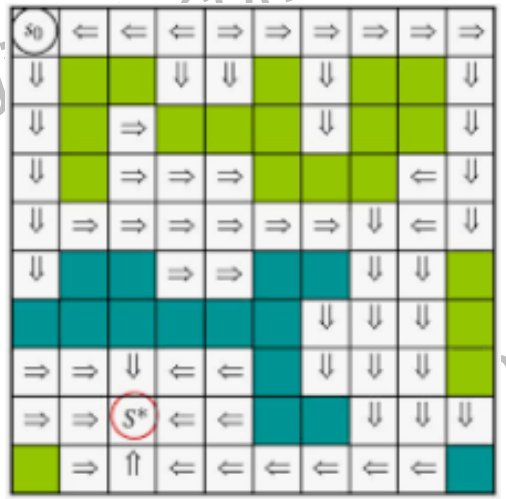


마르코프 결정 과정 그리드월드

그리드월드

- 바둑판처럼 정사각형으로 나누어진 환경에서 에이전트가 목적지를 찾아가는 게임
- 게임의 목적은 최단거리로 에이전트가 목적지를 찾아가도록 정책을 설정하는 것





마르코프 결정 과정 그리드월드

그리드월드 예제

① $k = 0$

0.0	0.0	0.0	0.0
0.0	목적지		0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0

목적지

② $k = 1$ 초기화

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0

③ $k = 2$ 좌표(0,1)

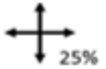
0.0	-1.7	-2.0	-2.0
-1.7	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.7
-2.0	-2.0	-1.7	0.0

.....

$k = \infty$

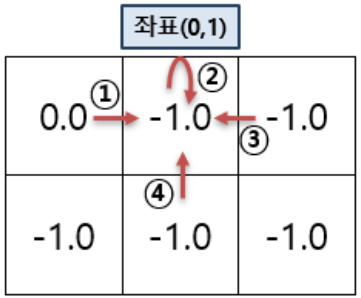
0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0

- 상태전이확률 : 1로 가정
- 보상 : 타임스텝에 따라 -1
- 초기정책 : 랜덤(상/하/좌/우 : 0.25) actions

actions 

마르코프 결정 과정 그리드월드

상태가치계산



k=1

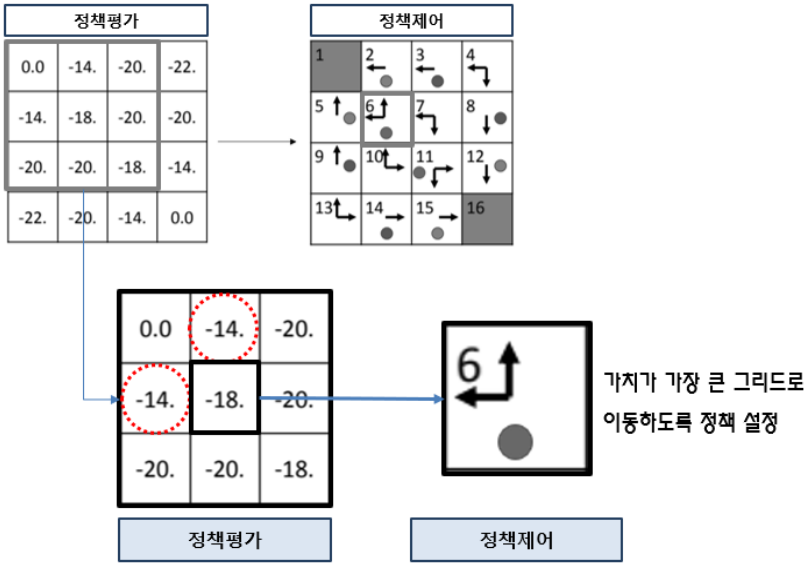
정책평가

-1.0 + (0.0*0.25 + -1.0*0.25 + -1.0*0.25 + -1.0*0.25) = -1.75

현재상태의 가치 ① ② ③ ④

마르코프 결정 과정 그리드월드

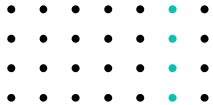
정책 업데이트



정책제어

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

마르코프 결정 과정 그리드월드



참고사이트

https://cs.stanford.edu/people/karpathy/reinforcejs/gridworld_dp.html