Chapter09

통계 모델링

I 마르코프 의사결정 과정

FASTCAMPUS
ONLINE

금융공학/퀀트 I

강사. 장순용

키포인트

- 마르코프 의사결정 과정 (Markov Decision Process, MDP).
- 행동, 상태, 보상, 전이확률, 감가율.
- 반환값.
- 격자 세상 (Grid world).

I 마르코프 의사결정 과정 : 기초

- 순차적 행동결정 문제를 수학적으로 모델링한 것이다 ⇒ 정책의 최적화.
- 환경은 전체가 완전히 관찰 가능해야 한다.
- 현재 상태가 과정의 특성을 모두 반영하여야 한다. 미래의 상태는 현재의 상태에 의해서 결정되고 더 이상의 과거는 돌아볼 필요가 없다.

⇐ 마르코프 과정 전제.

- 마르코프 의사결정 과정의 해는 다이내믹 프로그래밍 방법 또는

I 마르코프 의사결정 과정 : 사례

⇒ 주어진 상태에서 그 다음 최적 움직임이 무엇이지 결정해야 하는 경우.

바둑

I 마르코프 의사결정 과정 : 사례

⇒ 주어진 상태에서 그 다음 최적 움직임이 무엇이지 결정해야 하는 경우.

I 마르코프 의사결정 과정 : 용어

- 에이전트 (agent): 환경과 상호작용하는 알고리즘 또는 시스템.
- 정책 (policy): 정책은 특정 상태에서 에이전트가 취해야 할 행동을 정의한다. 즉, 상태에 행동을 매핑해 놓은 것이다. 정책은 확정적일 수도 있고, 확률적일 수도 있다.

??

Left??

Right??

마르코프 의사결정 과정 : 정의

- 마르코프 의사결정 과정 (MDP)은 다음과 같이 정의된다.

$$MDP = (S, A, P^a, R^a, \gamma)$$

- MDP 는 조건부 확률에 대한 주장이다. 마르코프 과정을 따른다는 가정하에 상태 S_{t+1}은 바로 이전의 상태 S_t에 대한 조건으로 만들어 진다. 즉, 더 먼 과거로 부터의 모든 필요한 정보가 반영된다는 것을 알 수 있다. 이것을 수식으로 다음과 같이 표현할 수 있다.

$$P(S_{t+1}|S_t, S_{t-1}, S_{t-2}, \dots, S_1) = P(S_{t+1}|S_t)$$

마르코프 의사결정 과정 : 상태

- 마르코프 의사결정 과정 (MDP)은 다음과 같이 정의된다.

$$MDP = (\boldsymbol{S}, A, P^a, R^a, \gamma)$$

→ $\boldsymbol{S} = \{s_1, s_2, s_3, \dots\}$: 에이전트가 취할 수 있는 상태 (state)의 집합.

I 마르코프 의사결정 과정 : 행동

- 마르코프 의사결정 과정 (MDP)은 다음과 같이 정의된다.

$$MDP = (S, \boldsymbol{A}, P^a, R^a, \gamma)$$

→ $\boldsymbol{A} = \{a_1, a_2, a_3, \dots\}$: 에이전트가 취할수 있는 행동 (action)의 집합.

→ 특정 시점 t 에서 행동 a를 취하는 경우$A_t = a$와 같이 표기한다.

→ 격자세상 (grid world)에서는 $A = \{$위, 아래, 왼쪽, 오른쪽$\}$와 같다.

I 마르코프 의사결정 과정 : 전이확률

- 마르코프 의사결정 과정 (MDP)은 다음과 같이 정의된다.

$$MDP = (S, A, \boldsymbol{P^a}, R^a, \gamma)$$

→ $\boldsymbol{P^a_{s\,s'}} = P(S_{t+1} = s' | S_t = s, A_t = a)$: 시점 t 에서 행동 a를 취한 경우 상태가 s에서s'로 전이되는 확률이다.

→ 환경과 에이전트의 상호작용: 에이전트가 행동을 취하면 전이확률을 통해서 다음으로 에이전트가 갈 상태를 알려준다.

I 마르코프 의사결정 과정 : 보상

- 마르코프 의사결정 과정 (MDP)은 다음과 같이 정의된다.

$$MDP = (S, A, P^a, \boldsymbol{R^a}, \gamma)$$

→ $\boldsymbol{R_s^a} = E[R_{t+1} | S_t = s, A_t = a]$: 시점 t의 상태 s에서 행동 a를 취한 대가로 에이전트가 받는 보상이다.

→ R_s^a와 같이 표기함은 기대값의 의미로의 보상이고 R_t, R_{t+1}등과 같이 표기함은 확률 변수의 의미로의 보상이다.

→ 시점 t에서의 행동에 따라서 보상이 전달되는 시점은 $t+1$이다.

I 마르코프 의사결정 과정 : 감가율

- 마르코프 의사결정 과정 (MDP)은 다음과 같이 정의된다.

$$MDP = (S, A, P^a, R^a, \gamma)$$

→ γ : 과거 또는 미래의 행동에 대한 보상을 얼마만큼 반영할지 정하는 0과 1사이의 수치.

→ 예를 들어서 $\gamma = 0.9$ 라면 $\gamma^2 = 0.81, \gamma^3 = 0.729, \ldots$ 과 같다.

I 마르코프 의사결정 과정 : 감가율

- 마르코프 의사결정 과정 (MDP)은 다음과 같이 정의된다.

$$MDP = (S, A, P^a, R^a, \gamma)$$

→ 현재 시점 t에서 k만큼 지난 후에 보상 R_{t+k}을 받는다면 γ^{k-1} 만큼의 감가가 적용된다. 미래의 가치를 현재의 가치로 환산하는 것과도 같다.

→ 할인된 미래 $t+k$의 보상 $= \gamma^{k-1}R_{t+k}$.

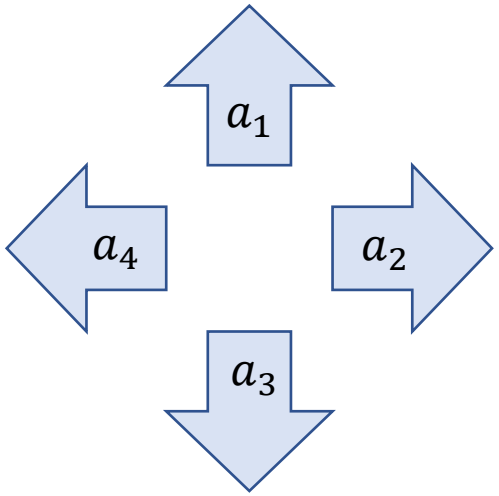
→ 다음과 같이 감가된 미래의 보상 "반환값"을 정의할 수 있다.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

I 마르코프 의사결정 과정 : 격자 세상 (Grid World)

- 다음과 같이 상태의 집합 S와 행동의 집합 A가 있다고 가정해 본다.

s_1 출발점	s_2	s_3 함정
s_4	s_5	s_6
s_7 함정	s_8	s_9 목적지

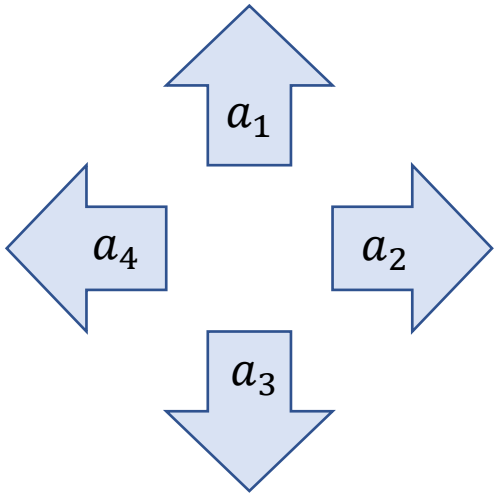


- 감가율 $\gamma = 0.9$임을 전제한다.

I 마르코프 의사결정 과정 : 격자 세상 (Grid World)

- 다음과 같은 보상 구조를 전제한다.

−0.02 출발점	−0.02	−1 함정
−0.02	−0.02	−0.02
−1 함정	−0.02	+1 목적지

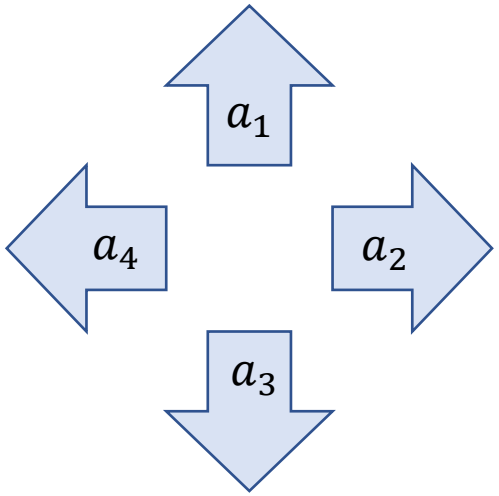


- s_1에서 행동 a_2를 연이어서 두번 취해서 상태를 $s_1 \rightarrow s_2 \rightarrow s_3$와 같이 변이하여 함정에 빠졌다면, -0.02, -1 과 같은 순서로 보상을 받는다.

I 마르코프 의사결정 과정 : 격자 세상 (Grid World)

- 다음과 같은 보상 구조를 전제한다.

-0.02 출발점	-0.02	-1 함정
-0.02	-0.02	-0.02
-1 함정	-0.02	$+1$ 목적지



- 감가율을 적용한 반환값은 $-0.02 + 0.9 \times (-1) = -0.92$ 와 같다.

I 끝.

Life Changing Education

감사합니다.

Fast campus

FAST CAMPUS
ONLINE
장순용 강사.

Fast campus