Chapter06

벨만의 방정식

마르코프 의사결정 과정과 벨만의 방정식

FASTCAMPUS
ONLINE
금융공학/퀀트 I

강사. 장순용

키포인트

- 마르코프 의사결정 과정 (Markov Decision Process, MDP).
- 행동, 상태, 보상, 전이확률, 감가율.
- 반환값.
- 벨만의 방정식 (Bellman’s Equation).

I 마르코프 의사결정 과정 : 기초

- 순차적 행동결정 문제를 수학적으로 모델링한 것이다.

  ⇒ 정책의 최적화를 목표.

- 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.

I 마르코프 의사결정 과정 : 사례

- 주어진 상태에서 최적의 전략을 결정해야 하는 경우.

바둑

I 마르코프 의사결정 과정 : 사례

- 주어진 상태에서 최적의 전략을 결정해야 하는 경우.

투자

FAST CAMPUS
ONLINE
장순용 강사.

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??

I 마르코프 의사결정 과정 : 정의

- 마르코프 의사결정 과정 (MDP)은 다음과 같이 정의된다.

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

- 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.

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

I 마르코프 의사결정 과정 : 정의

- 마르코프 의사결정 과정 (MDP)은 다음과 같이 정의된다.

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

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

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

→ $\boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{s}\,\boldsymbol{s'}}$: 행동 a를 취한 경우 상태가 s에서s'로 전이되는 확률이다.

→ $\boldsymbol{R}^{\boldsymbol{a}}_{\boldsymbol{s}}$: 상태 s에서 행동 a를 취한 대가로 에이전트가 받는 보상이다.

→ $\boldsymbol{\gamma}$: 과거 또는 미래의 행동에 대한 보상을 얼마만큼 반영할지 정하는 0과 1사이의 수치.

I 벨만의 반정식 (Bellman's Equation)

- 벨만 방정식을 사용하여 MDP를 수학적으로 정의할 수 있다.
- 이 방정식을 통해서 주어진 환경에서의 최적 정책을 알아낼 수 있다.
- "정책"이라 함은 다음과 같은 수식으로 정의 할 수 있다.

$\Rightarrow$ t 시점에서 에이전트의 상태가 s일 때 a라는 행동을 취할 확률이다.

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

I 벨만의 반정식 (Bellman's Equation)

- 다음과 같은 두가지 방법으로 상태가 주어졌을 때의 가치 $v_\pi(s)$를 표현할 수 있다:

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

"가치함수 기대값 방정식"

$$v_\pi(s) = R(s) + \gamma \times \sum_{s'} P_{s\,s'}^{\pi(s)}\, v_\pi(s')$$

⇒ 구하고자 하는 해는 최적 가치함수 $v^*(s)$와 최적 정책 $\pi^*(s)$이다.

⇒ 다이내믹 프로그래밍 방법은 쉽게 적용할 수 있지만 상태의 수가 증가하면 현실적이지 않다.

⇒ 강화학습 알고리즘은 에이전트가 기대 보상이 최대화 되도록 행동을 선택하게 하여 최적의 정책을 학습한다.

끝.

감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.