Chapter07

강화학습

강화학습 방법 II

FASTCAMPUS
ONLINE
금융공학/퀀트 I

강사. 장순용

키포인트

- Q 함수.
- SARSA 강화학습.
- Q Learning 강화학습.

Copyright FASTCAMPUS Corp. All Rights Reserved

I Q 함수의 정의

- Q 함수란 상태-행동 함수 $Q_\pi(a, s)$를 일컫는다.

  $\Rightarrow$ 상태 s에서 행동 a를 선택하고 일련의 정책 $\pi(a|s)$를 따라서 행동을 이어갈 때 얻게되는 결과값을 의미한다.

- Q 함수를 사용하여 가치함수 다음과 같이 계산할 수 있다.

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \cdot Q_\pi(a, s)$$

Copyright FASTCAMPUS Corp. All Rights Reserved

I Q 함수의 정의

- Q 함수 또한 벨만의 기대값 방정식의 형태로 표현할 수 있다.

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

⇐ 이전에 알아본 벨만의 가치함수 기대값 방정식과 비교해 본다.

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

Copyright FASTCAMPUS Corp. All Rights Reserved

I SARSA 강화학습

- SARSA는 시간차 학습(TD)와 Q 함수를 사용한 가치반복의 조합이다.

⇒ 특정 정책을 전제하지 않고 현 상태에서 가장 큰 가치의 행동을 선택한다.

⇒ Agent는 가치함수가 아닌 Q 함수에 따라서 행동한다.

⇒ 갱신의 대상은 가치함수가 아니라 Q 함수이다.

$$Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)) \rightarrow Q(S_t, A_t)$$

⇒ 필요한 표본은 $S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}$이다: 순서대로 읽어가면 SARSA가 된다.

Fast campus

Copyright FASTCAMPUS Corp. All Rights Reserved

I Q Learning 강화학습

- SARSA는 소위 on-policy 시간차 제어이며, agent 자신이 행동하는 대로 학습하게 된다.
- 반면에 Q 러닝은 off-policy 시간차 제어 강화학습 방법이다.

⇒ 행동하는 정책과 학습하는 정책을 분리해 두는 것이 그 특징이다.

⇒ Agent는 행동하는 정책으로 지속적인 탐험을 하면서 별도의 목표 정책을 두어서 학습은 목표 정책에 의해서 하게 된다.

Copyright FASTCAMPUS Corp. All Rights Reserved

끝.

감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.

Copyright FASTCAMPUS Corp. All Rights Reserved