Chapter07

강화학습

강화학습 방법 I

FASTCAMPUS

ONLINE

금융공학/퀀트 I

강사. 장순용

I 키포인트

- 몬테카를로 (Monte Carlo, MC).
- 시간차 학습 (Time Difference Learning, TD).

I 몬테카를로

- 다음과 같은 두 가지 스텝을 반복하게 된다.

⇒ 평가 스텝 (Evaluation Step):

몬테카를로 평가를 적용하여 agent가 가치함수를 학습한다.

⇒ 제어 스텝 (Control Step):

가치함수를 바탕으로 정책을 계속해서 갱신해 나가며 최적 정책을 학습한다.

I 몬테카를로 평가

- 몬테카를로 평가의 **장점**은 다음과 같다:

⇒ 전체 상태의 개수와는 무관하다. 즉, 문제의 규모와 무관하다는 것이다.

⇒ 관심있는 상태부터 시작하여 많은 표본 episode를 생성할 수 있다.

⇒ 환경 모형을 정의해 주는 $P_{s\,s'}^{a}$,와 같은 수치를 모르더라도 적용이 가능하다.

I 몬테카를로 평가

- 몬테카를로 평가의 **단점**은 다음과 같다:

⇒ 모든 정책에 대해서 종료가 보장되지 않아서 무한반복 상황에 빠질 수도 있다.

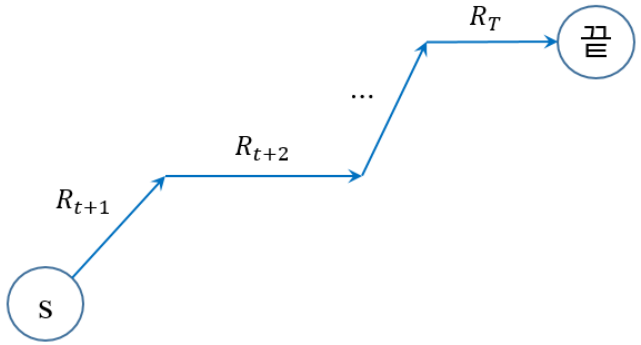
⇒ 가장 효율적인 방법은 아니다.

I 몬테카를로 평가

- 몬테카를로는 표본을 시뮬레이션 하는 방식이며 표본의 수가 많아 질수록 실제값에 수렴한다.
- 표본이 경험치를 대체하게 된다. 경험은 episode로 축적된다.

⇒ Episode는 아래와 같이 현재부터 끝까지의 상태의 연쇄를 의미한다.

⇒ 이렇게 가치함수를 학습 (계산)하는 것을 평가라 부른다.



I 몬테카를로 평가

- Episode의 반환값은 다음과 같다.

$$G_t(s) = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t+1} R_T$$

- 그러면 가치함수는 이들의 평균으로 구할 수 있다. N_s는 표본의 크기이다.

$$v(s) = \frac{1}{N_s} \sum_{i=1}^{N_s} G(s)_i$$

- Episode의 반환값 표본이 축적될 수록 가치함수의 참값에 수렴해 간다.

$\Rightarrow$ 수렴 과정을 다음과 같이 나타낼 수 있다. α는 조정 가능한 스텝 사이즈.

$$v(s) + \alpha(G(s) - v(s)) \rightarrow v(s)$$

I 제어 스텝

- 예를 들어서 탐욕 정책 발전 (Greedy Policy Improvement) 방법을 적용할 수 있다.
- 계산된 가치함수를 최적 가치함수인 것처럼 사용하여 정책을 갱신할 수 있다.

$$\pi(s) = \underset{a}{argmax} \sum_{s'} P^{a}_{s\,s'} \cdot v(s')$$

I 시간차 학습 (TD)

- 몬테카를로 평가와 같이 시뮬레이션 된 경험치로 가치함수를 평가하는 방법이다.

⇒ 하지만 episode의 완결을 기다리지 않는다.

⇒ 매 episode 단위로 갱신하는 것이 아니라 매 time step 가치함수를 평가한다.

I 시간차 학습 (TD)

- Episode 전체의 반환값인 $G(s)$를 사용할 수 없다.

  $\Rightarrow$ 대신 $R + \gamma v(s')$를 사용한다.

- 매 time step의 수렴 과정을 다음과 같이 나타낼 수 있다.

$$v(s_t) + \alpha(R + \gamma v(s_{t+1}) - v(s_t)) \rightarrow v(s_t)$$

  $\Rightarrow$ Episode가 완결되지 않고도 time step별 갱신이 가능하다.

끝.

감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.