Google Developers

Experts

n단계 부트스트랩
n-step bootstrap

강화학습

송호연 @chris_loves_ai
sjhshy@gmail.com

n단계 TD 방법
n-step TD learning

단일 단계 TD

one-step TD

- 단일 단계 갱신에서는 최초의 이득과 다음 상태의 할인된 가치 추정값의 합이 목표다.
  이 합을 **단일 단계 이득(one-step return)**이라고 부른다.

$$q_t^{(1)} = R_{t+1} + \gamma Q(S_{t+1})$$

두 단계 TD

two-step TD

- **두 단계 이득(two-step return)**에서는 최초의 이득과 다음 상태의 할인된 가치와 두 단계 뒤 상태의 할인된 가치의 추정값의 합이 목표다.

$$q_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 Q(S_{t+2})$$

n단계 TD

n-step TD

- n단계 방법은 한쪽 끝에 MC 방법이 있고 다른 쪽 끝에는 단일 단계 TD 방법이 있는 스펙트럼을 포괄한다. 최선의 방법은 대체로 중간 쯤에 있는 방법이다.

- n단계 방법의 이점 중 하나는 n단계 방법이 시간 단계의 억압으로부터 자유롭게 해준다는 것이다.

$$q_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n Q(S_{t+n})$$

n단계 SARSA

n-step SARSA

- **n-step Q 함수**

$$q_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^n Q(S_{t+n})$$

- **n-step SARSA update**

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left(q_t^{(n)} - Q(S_t, A_t) \right)$$

$$\begin{aligned}
n = 1 \quad (Sarsa) \quad & q_t^{(1)} = R_{t+1} + \gamma Q(S_{t+1}) \\
n = 2 \quad & q_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 Q(S_{t+2}) \\
\vdots \quad & \vdots \\
n = \infty \quad (MC) \quad & q_t^{(\infty)} = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{T-1} R_T
\end{aligned}$$

n단계 비활성 정책 학습
n-step off-policy learning

n단계 비활성 정책 학습

n-step off-policy policy learning

- n단계 비활성 정책 학습은 다음 공식으로 표현할 수 있다.

$$V_{t+n}(S_t) = V_{t+n-1}(S_t) + \alpha\rho_{t:t+n-1}[G_{t:t+n} - V_{t+n-1}(S_t)], 0 \leq t < T$$

- 여기서 중요도추출비율(importance sampling ratio)라고 불리는 ρ는 두 정책 하에서 A_t로부터 A_t+n-1 까지의 n개의 행동을 취할 상대적 확률로서, 다음과 같이 계산된다.

$$\rho_{t:h} = \prod_{k=t}^{min(h,T-1)} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$$

E.O.D.