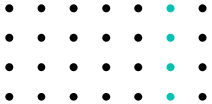
03

강화학습 기본 알고리즘

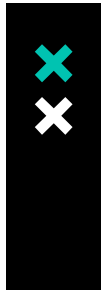
3. 몬테카를로 방법

몬테카를로 방법 기본개념



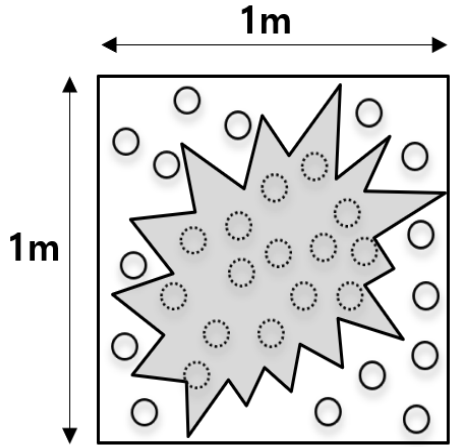
몬테카를로 방법(MC: Monte-Carlo Method)

- 모델프리 환경에서 사용하는 MDP 해결 방법
- 정확한 수학 수식에 의해 계산하거나 측정하는 것이 아니라 확률적인 방법에 의해 값을 통계적으로 계산하는 방법
- 계산하려는 값이 복잡할 때 정확한 결과를 얻기 보다는 근사적인 결과를 얻을 경우에 사용
- 에이전트가 동작하는 환경은 시작과 끝이 있는 에피소드 단위의 환경에서 사용 가능



몬테카를로 방법 사례

몬테카를로 방법(MC: Monte-Carlo Method)



사각형면적 : 1m x 1m = $1m^2$

사각형안 공 개수 : 30개

다각형안 공 개수 : 15개

몬테카를로 메소드 : $1m^2$: x ≒ 30개 : 15개

다각형면적 ≒ $0.5m^2$

몬테카를로 방법 MDP 해결하기

MDP 해결하기

MDP

$$v_\pi(s) = \mathbf{E}_\pi[G_t \mid S_t = s] \quad ①$$

$$= \mathbf{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$$

$$= \sum_{a \in A} \pi(a|s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')\right)$$

$$= \sum_{a \in A} \pi(a|s)\, R_s^a + \gamma \underbrace{\sum_{a \in A} \pi(a|s)}_{②\text{-}1} \underbrace{\sum_{s' \in S} P_{ss'}^a v_\pi(s')}_{②\text{-}2}$$

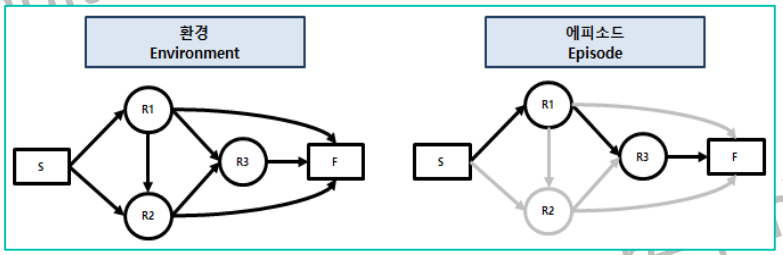
MC

$v_\pi(s) = V(s)$ when $N(s) \to \infty$ ③

누적 Count : $N(s) \leftarrow N(s) + 1$ (하나의 episode 수행) ④

누적 Return : $S(s) \leftarrow S(s) + G_t$ ⑤

평균 Return : $V(s) \leftarrow S(s) / N(s)$ ⑥



정책평가

몬테카를로 방법

증분평균

증분평균

$$\mu_k = \frac{1}{k}\sum_{j=i}^{k} x_j \quad ①$$

$$= \frac{1}{k}\left(x_k + \sum_{j=i}^{k-1} x_j\right) \quad ②$$

$$= \frac{1}{k}\left(x_k + \left(\frac{k-1}{1}\right)\underbrace{\left(\frac{1}{k-1}\right)\sum_{j=i}^{k-1} x_j}_{③\text{-}1}\right) \quad ③$$

$$= \frac{1}{k}\left(x_k + (k-1)\mu_{k-1}\right) \quad ④$$

$$= \mu_{k-1} + \frac{1}{k}\left(x_k - \mu_{k-1}\right) \quad ⑤$$

$$\fallingdotseq \mu_k + \frac{1}{k}\left(x_k - \mu_k\right) \quad ⑥$$

수학적 계산 편의를 위해 변경
from k-1 to k

- 이전 타임스텝까지 계산된 평균을 활용하여 새로운 값이 들어왔을 때 전체 평균을 빠르게 구할 수 있는 방법

$$\mu_k + \frac{1}{k}\left(x_k - \mu_k\right)$$

새로운 값 / 과거의 값

몬테카를로 방법 증분평균과 MC

증분평균과 MC

MC $v_\pi(s) = V(s)$ when $N(s) \rightarrow \infty$

정책평가

누적 Count : $N(s) \leftarrow N(s) + 1$ (하나의 episode 수행)

누적 Return : $S(s) \leftarrow S(s) + G_t$

평균 Return : $V(s) \leftarrow S(s) / N(s)$ ①

증분평균 Incremental Mean Return : $V(s) \leftarrow V(s) + \frac{1}{N(s)}(G_t - V(s))$ ②

몬테카를로 방법

프로그램위한 MC

프로그램위한 MC

MC

$$V(s) \leftarrow V(s) + \frac{1}{N(s)}(G_t - V(s)) \quad ①$$

$\frac{1}{N(s)}$을 $\propto$로 변경

$$V(s) \leftarrow V(s) + \propto(G_t - V(s)) \quad ②$$

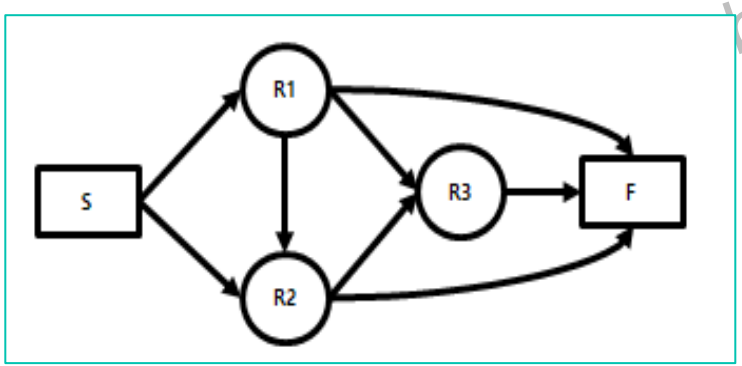
정책평가

에피소드를 반복하면서 계속해서 G를 구하고, 증분평균을 계산해서 가치함수를 업데이트 하다 보면 결국에는 참 가치함수(신(God)만이 알고 있는 값)를 구할 수 있다.

몬테카를로 방법 DP와 MC

다이나믹 프로그래밍(DP)과 MC

DP



MC