03

강화학습 기본 알고리즘

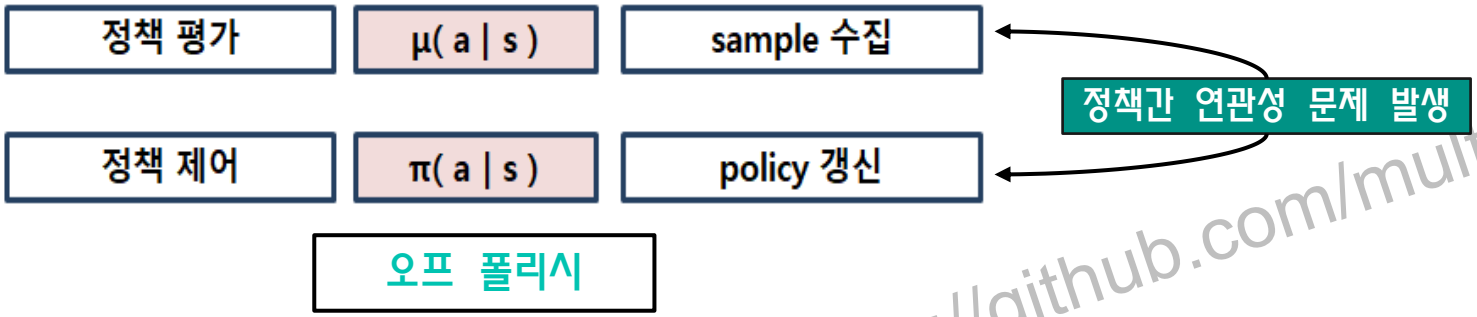
5. Q러닝

Q러닝

온 폴리시와 오프 폴리시

온 폴리시(On Policy)와 오프 폴리시(Off Policy)

- 온 폴리시 : 정책을 평가하는데 사용하는 정책(π)과 정책을 제어하는데 사용하는 정책(π)이 모두 같은 경우
  한번 평가에 사용한 정책은 다음에 재사용할 수 없다.
- 오프 폴리시 : 정책 평가에 사용되는 정책과 정책 제어에 사용되는 정책을 각각 따로 사용

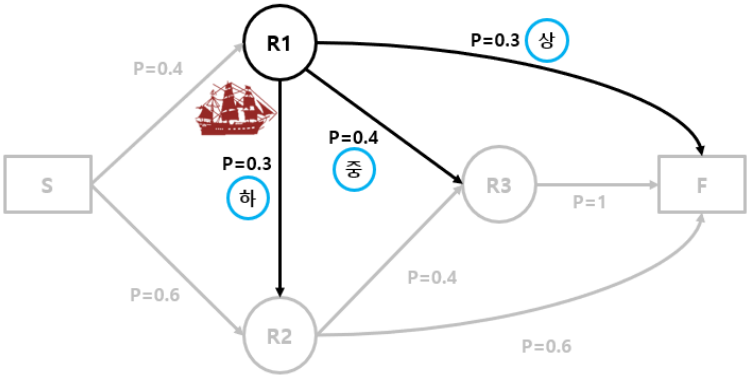


Q러닝

확률변수와 확률분포

확률변수와 확률분포

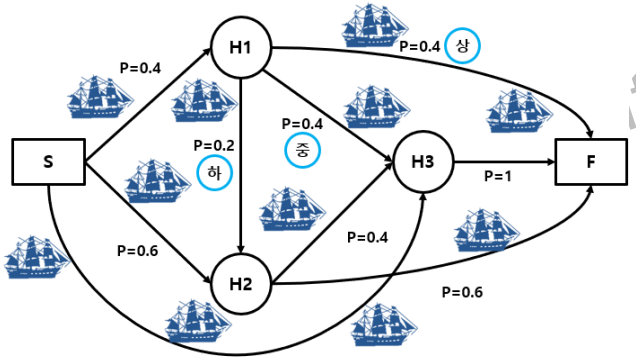
- 확률변수 : 행동의 종류(A: Set of Actions), 확률적인 과정에 따라 값이 결정되는 변수
- 확률분포 : 정책(π: Policy), 확률 변수가 특정한 값을 가질 확률을 나타내는 함수



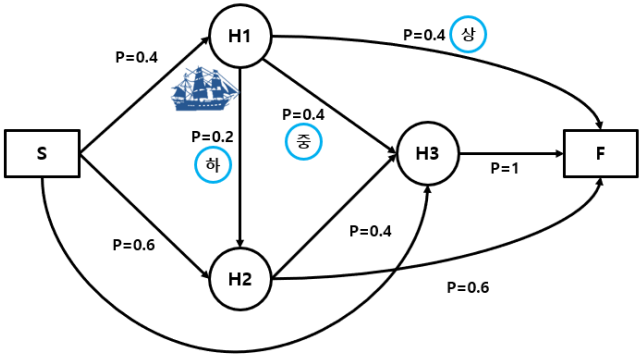
확률변수	{상, 중, 하}	Action
확률분포	{0.3, 0.4, 0.3}	Policy

Q러닝 중요도 샘플링

For Example



기존 항로 데이터



신규 항로

Q러닝 중요도 샘플링

중요도 샘플링

$$\sum P(X)f(X) = \sum Q(X) \left[\frac{P(X)}{Q(X)} f(X) \right]$$

P(X)	어떤 환경에서 변수 X의 확률분포 P
Q(X)	다른 환경에서 변수 X의 확률분포 Q
f(X)	X의 함수 어떤 함수도 가능(sin, cos, 2x+1 등)
∑P(X)f(X)	변수X의 함수 f(X)에 대한 확률분포 P의 기댓값

Q러닝 중요도 샘플링

MC와 TD에서 중요도 샘플링

μ	정보가 풍부한 환경에서 사용하고 있는 정책
π	학습하는 환경에 대한 정책(알고 싶어하는)

MC

$$G_t^{\pi/\mu} = \frac{\pi(A_t|S_t)\pi(A_{t+1}|S_{t+1})}{\mu(A_t|S_t)\mu(A_{t+1}|S_{t+1})} \cdots \frac{\pi(A_n|S_n)}{\mu(A_n|S_n)} G_t$$

$$V(S_t) \leftarrow V(S_t) + \propto (G_t^{\pi/\mu} - V(S_t))$$

TD

$$V(s_t) \leftarrow V(s_t) + \propto (\frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}(R_{t+1} + \gamma V(s_{t+1})) - V(s_t))$$

Q러닝 기본개념

Q러닝 기본개념

- 경험을 쌓을 때 다음 행동은 정책을 따라가는 것이 아니라 Q값을 max로 만드는 행동을 선택한다. 이것이 SARSA와 Q러닝의 차이점
- 중요도 샘플링(Importance Sampling)을 사용하지는 않지만 정책을 평가할 때 사용하는 정책(max)과 정책을 제어(π)할 때 사용하는 정책이 다르기 때문에 오프 폴리시 방법

SARSA $Q(S,A) \leftarrow Q(S,A) + \alpha(R_{t+1} + \gamma Q(S',A') - Q(S,A))$

Q-Learning $Q(S,A) \leftarrow Q(S,A) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S',a) - Q(S,A))$