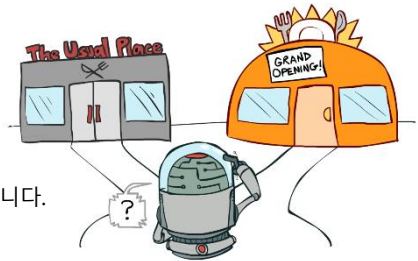
Monte Carlo

Kihwan Lee

Contents

1. Exploitation vs Exploration

Exploitation vs Exploration



- Exploitation

: 이미 알고 있는 정보 내에서 가장 최선의 선택을 하는 것

ex) 이미 알고 있는 밥집 중에서 가장 맛있는 집을 찾는 것

이미 맛있는 곳을 가기에 안전하지만 더 맛있는 집을 찾을 방법은 없습니다.

- Exploration :

알고 있는 정보 이외에 더 최선의 방법을 찾아 떠나는 것

ex) 알고 있는 최선의 맛집 이외에 가 보지 않은 곳을 선택해 보는 것

=> 방문하지 않았던 식당을 찾기에 실패할 수도 있지만 새로운 맛집을 찾을 수 있습니다.
장기적으로 보면 Exploration이 좋습니다!

Exploitation vs Exploration

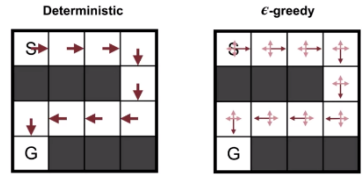
< 이를 그대로 강화학습에 적용! >

- Exploitation

: highest Q value를 갖는 action을 선택합니다. 기존에 배웠던 greedy policy와 같습니다. 이렇게 얻은 action들 즉 이 policy를 가지고 새로운 에피소드를 샘플링하게 됩니다. 취해보지 않은 state action pair에 대해서는 Q value 값을 모르기에 update가 되지 않습니다.

- Exploration

: 여기서는 ε만큼 다른 action을 취해봅니다. 예를 들어 ε=0.0001이라면,
1-ε=99.99%로 highest Q value를 택하고, ε만큼 새로운 action을 취하는 것 입니다.
이렇게 함으로 가보지 않았던 pair에 대해 방문하며 sampling해보는 것 입니다.
ε확률만큼 손해볼 수도 있지만 여러 번 반복하다 보면
더 좋은 Q value를 찾을 수도 있습니다.



=> ε-greedy policy!

2. MC control
(MC policy improvement)

MC control

- **MC control : MC policy improvement**

이제 더 이상 Action을 고정된 형식으로 policy를 정하는 것이 아니라, ε-greedy 형식으로!

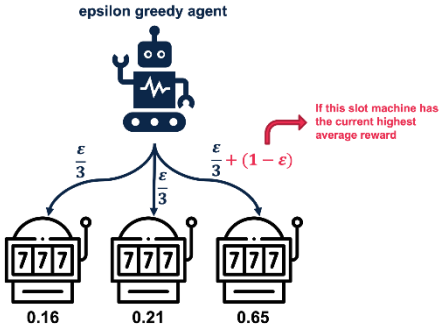
Q) 그렇다면 ε 은 어떤 값으로? Action의 개수

$$\pi'(a|s) = \begin{cases} 1-\epsilon+\frac{\epsilon}{m} & \text{if } a = \arg\max_{a'} Q^{\pi}(s,a') \\ \frac{\epsilon}{m} & \text{otherwise } (m-1 \text{ actions}) \end{cases}$$

=> Dynamic Programming 에서는 모든 것을 고려하기에 highest Q value를 갖는 action을 고정적으로 하나 선택했지만, 이제는 확률적으로 선택하게 됩니다.

: 1-ε+ ε/m 만큼 highest Q value를 갖는 action을, ε/m만큼 모든 action에 확률을 부여.

MC control



이와 같이 강화학습은 ε-greedy policy로 stochastic policy를 사용하기에 기존에 주어진 샘플 안에서의 가장 좋은 action을 취할 뿐 아니라, 가끔은 취하지 않았던 action도 취해보며 장기적으로는 더 좋은 action들을 선택하는 기회를 만들게 됩니다!

3. GLIE

GLIE

- **GLIE : Greedy in the Limit with Infinite Exploration**

학습을 해 나감에 따라 충분한 탐험을 했다면 greedy policy에 수렴하는 것

아래 두 조건을 만족하면 GLIE를 만족하며, 이를 만족해야 수렴된 policy를 가집니다.

- all state-action pairs (s, a) are explored infinitely many times.
  $\lim_{k\to\infty} n_k(s, a) = \infty$
- the learning policy converges to a greedy policy.
  $\lim_{k\to\infty} \pi_k(a \mid s) = 1$ where $a = \arg\max_{a'} Q_k(s, a')$

Q) 그렇다면 MC의 learning policy는 GLIE를 만족하게 어떻게 만들 수 있을까?

첫 번째 경우 : 무조건 데이터 양으로 늘리기만 하면 됩니다.

두 번째 경우 : ε값을 점점 0으로 수렴시키면 가능합니다. Ex) ε을 1/k로 설정

=> 이런 방식으로 MC가 GLIE를 만족하게 control 가능!