Multi-Armed-Bandit (MAB)

Reinforcement Learning : An Introduction Chapter2 by Sutton 도서 참고

KAIG 세미나 (2019-02-28)

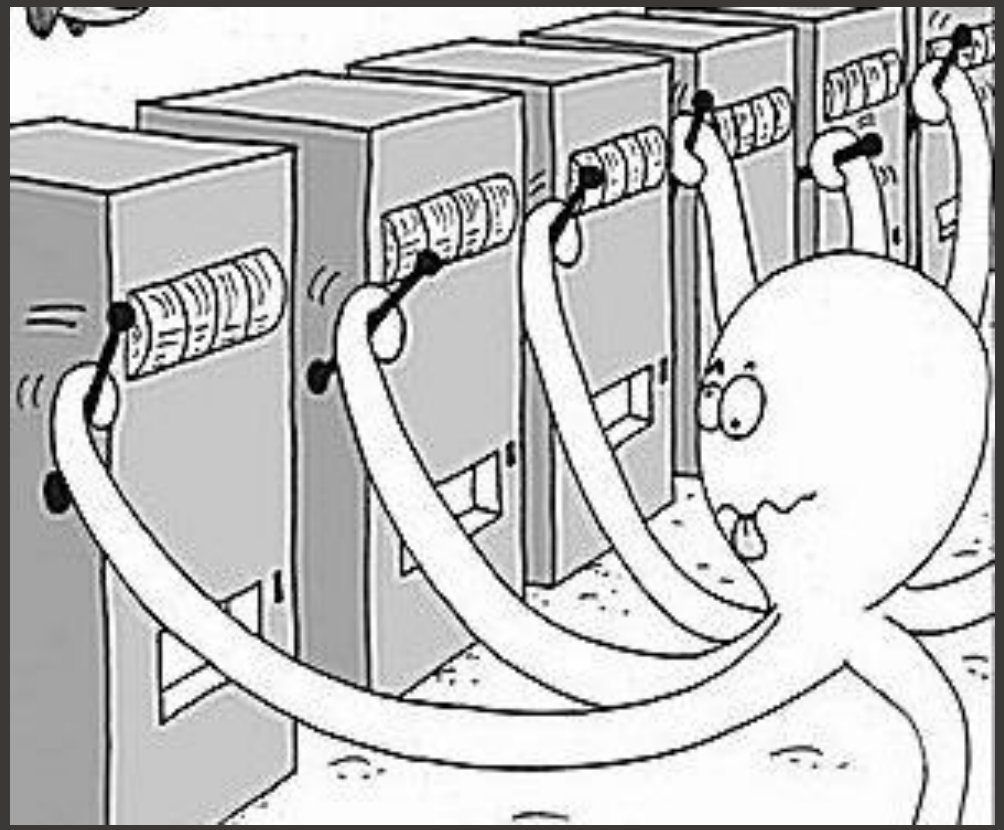
김수환

MAB의 개요

1. MAB란?
2. MAB 알고리즘
3. 정리

1. MAB란?

1. MAB란?



슬롯머신이 많은 **카지노**에서 등장한 개념이다.

여러 슬롯머신들 중에 어떤 식으로 선택을 해야

가장 최고의 수익을 낼 수 있을까?

라는 생각에서 시작된 문제다.

여러 개의 슬롯머신(Bandit) 중 어느 손잡이(Arm)를 당길

것인가라는 개념에서 이름이 비롯되었다.

이를 **Multi-Armed-Bandit Problem**이라고 부른다.

2. MAB 알고리즘

2. MAB 알고리즘

전략1. **greedy 알고리즘**:

여태까지 가장 좋은 결과를 낸 머신에 모두 투자하자!

2. MAB 알고리즘

슬롯머신1	슬롯머신2	슬롯머신3
10000원	5000원	500원

위와 같이 3개의 슬롯머신을 일정 횟수 테스트를 한 뒤,
결과를 통계를 냈다고 해보자.

위의 같은 결과를 냈다고 한다면, **앞으로는 슬롯머신1에 모두 투자**한다!

이러한 개념이 greedy 알고리즘이다.

2. MAB 알고리즘

greedy 알고리즘을 수식으로 표현

$$A_t = \underset{a}{\operatorname{argmax}}\, Q_t(a)$$

2. MAB 알고리즘

이러한 greedy 알고리즘의 문제점은

앞으로 일어날 수도 있는 **나머지 머신들의 가능성**에 대해

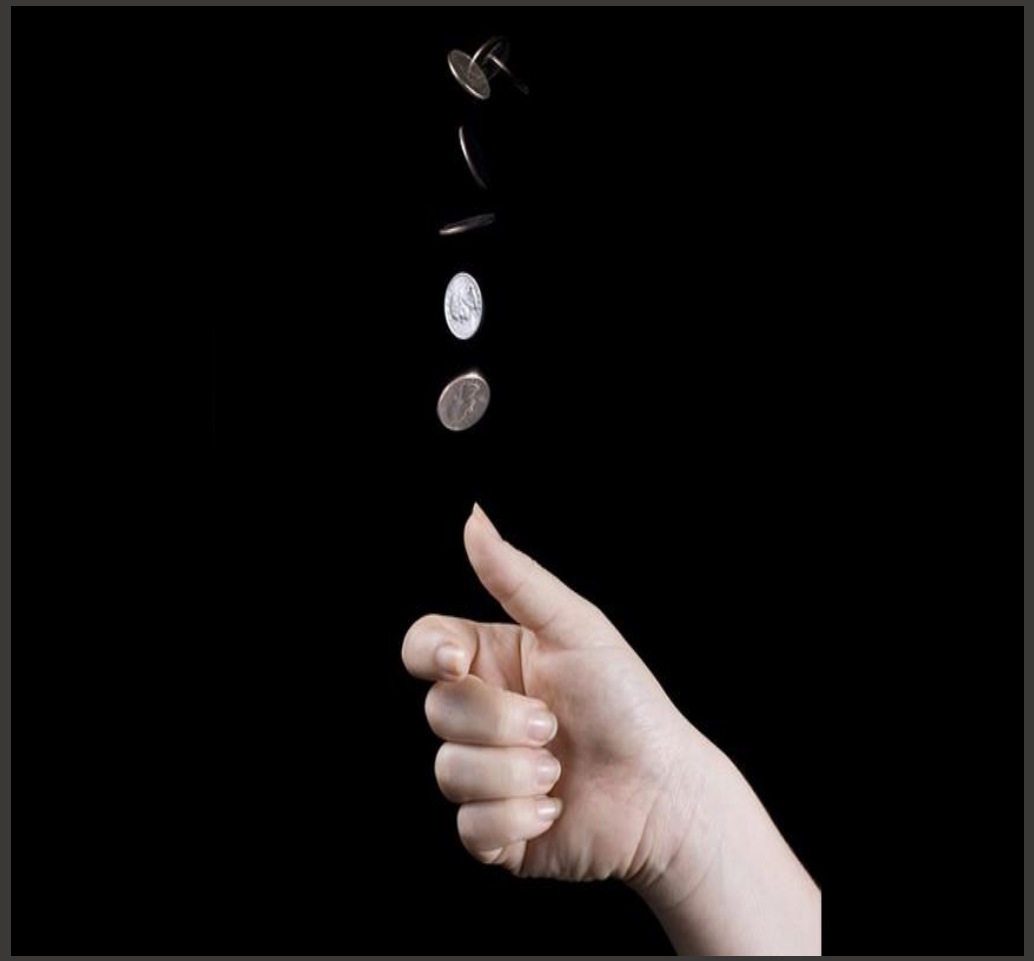
아예 닫아버린다는 점이다.

2. MAB 알고리즘

전략2. ε-**greedy 알고리즘**:

ε만큼의 확률은 아무 머신이나 해보지 뭐!

2. MAB 알고리즘



ε-greedy (Epsilon greedy) 알고리즘:

앞의 greedy 알고리즘에서 나머지 머신들에 대한

가능성을 열어두는 알고리즘이다.

동전을 던져서 앞면이 나온다면, greedy 알고리즘과 같이

가장 수익을 잘 낸 머신을 선택하고, 뒷면이 나온다면,

여러 머신들 중에 랜덤으로 선택한다.

(동전으로 든 예는 ε를 50%로 설정하였을 때)

2. MAB 알고리즘

ε-greedy 알고리즘을 수식으로 표현

$A_t = \underset{a}{\text{argmax}}\, Q_t(a)$ with Probablity 1-ε

Random Action with probability ε

2. MAB 알고리즘

이러한 ε-greedy 알고리즘은 강화학습의

Q-Learning에서 가장 대중적으로 쓰이는 알고리즘이다.

2. MAB 알고리즘

전략3. UCB **알고리즘**:

ε만큼의 확률은 랜덤보다는 **가능성**에 투자해보자!

2. MAB 알고리즘

UCB (Upper-Confidence-Bound) 알고리즘 :

ε-greedy 알고리즘에서 ε만큼의 확률에서 꼭 랜덤으로 해야할까?? 라는 생각에서 생긴 알고리즘이다.

ε-greedy 알고리즘은 ε만큼의 확률을 무작위 랜덤으로 선택하지만, UCB 알고리즘에서는

상대적으로 탐색이 덜 된 머신들을 우선 선택한다.

ε-greedy 알고리즘처럼 무작정 랜덤으로만 머신을 선택한다면 상대적으로 탐색이 덜 된 머신들이

생길 수 있기 때문에, **최대한 골고루 뽑히게 하는 알고리즘**이다.

2. MAB 알고리즘

UCB 알고리즘을 수식으로 표현

$$A_t = \underset{a}{\text{argmax}}[\, Q_t(a) + c\sqrt{\frac{logt}{N_t(a)}}\,]$$

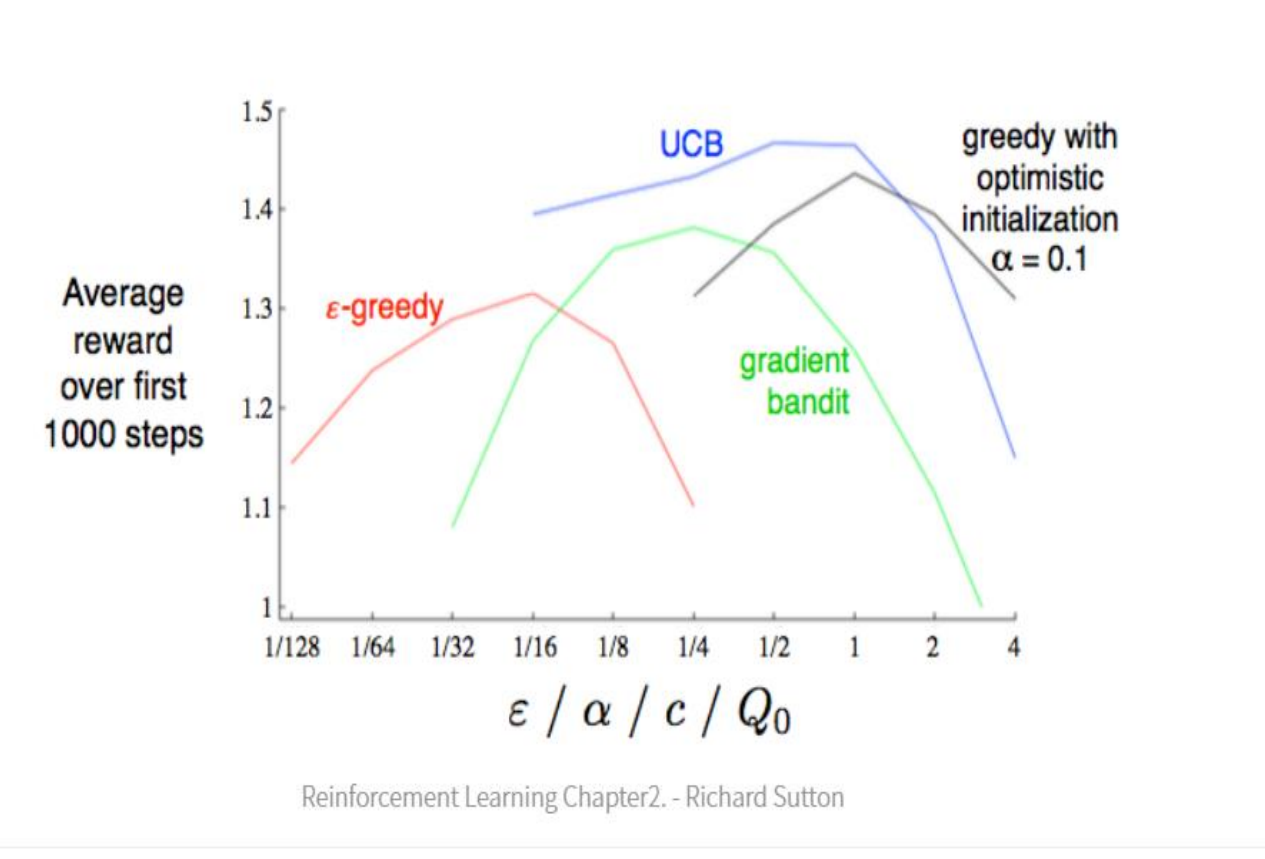
【c : 탐색의 정도를 조절하는 상수

【$N_t(a)$: 특정 슬롯머신을 선택한 횟수

【 t : 모든 슬롯머신을 선택한 횟수

3. 정리

3. 정리



이상으로 MAB 문제에 대한 3가지

알고리즘에 대하여 알아봤다.

참고한 책에서는 위 3가지 알고리즘 중

UCB 알고리즘이 해당문제에 대해

가장 효과가 좋았다고 나와있다.

이러한 MAB 알고리즘은 인터넷 기업들이

애용하고 실제로도 많이 사용한다고 한다.