Reinforcement Learning

지난 논의에서 발생한 의문점

1. 그래서 RL이 사용되는 게 어느 부분?
2. 어떤 점에서 RL이 의미를 가지나?

=> Value가 어떻게 정해지는지를 기준으로 이해하기 보다는 다시 개념을 통해 살펴 보기로 했다.

Reinforcement learning

- 강화학습은 환경(environment)과 상호작용하는 agent가 시행착오를 통해 최적의 정책을 학습하는 과정이다.
⇒ 즉 RL 알고리즘이 주관하는 건 'policy optimization'이 된다.
⇒ 강화학습의 최종 목적: optimal policy 학습

- 강화학습이 다루는 문제는 Sequential decision-making problem
➔ 현재의 행동이 미래의 보상에 영향을 미침
➔ 각 행동의 결과는 명확하지 않을 수 있음(지금의 선택이 나중에도 좋은 선택인지 알 수 없음)

➔ 최종적으로 경험하지 않은 상황에도 좋은 선택을 할 수 있도록 학습하는게 목표!

Reinforcement learning-구성요소

- Agent: 환경과 상호작용하며 Action 선택
- Environment: agent의 행동에 반응, 다음 state와 reward 반환
- Policy: S_t에서 어떤 행동을 취할지 결정하는 전략

➔ $\pi(a|s)$: 상태 s에서 해동 a를 선택할 확률 분포

- Reward: agent가 특정 행동을 수행한 결과로 받는 값

➔ 목표: 누적 보상을 최대화

- Value Function: 상태, 혹은 상태-행동 쌍의 장기적인 기대 보상 측정

➔ V(s): 상태 s의 가치, Q(s,a): 상태 s에서 행동 a를 선택했을 때의 가치

Value function vs. return function
- Value function: reward의 기대값
- Return function: 실제로 얻는 보상값

Policy Optimization / RL Algorithm

- 정책을 학습하여 누적 보상을 최대화하는 방향으로 최적화
- 정책을 직접 학습하거나, 가치 함수를 통해 간접적으로 학습

1. Value-based Methods
- 가치함수 V(s), Q(s,a)를 학습하여 간접적으로 정책을 최적화
- Ex. Q-learning, Deep Q-Network

2. Policy-based Methods
- 정책을 직접 최적화
- Ex. REINFORCE, Proximal Policy Optimization

틱택토/바둑에서 RL이 어떻게 사용?

- 단순히 모든 상태를 시뮬레이션하거나, 모든 가능성을 계산하는 것 X
→ 시뮬레이션을 효율적으로 수행, 경험을 통해 학습. 궁극적으로 정책을 최적화.

-틱택토 Agent

1. 무작위 행동(Exploration)을 통해 게임 플레이
2. 승패 결과에 따라 각 상태에서의 행동에 대해 보상 계산
3. 얻어진 데이터로 Q(s,a) 업데이트
4. 정책 학습

⇒어떤 상태에서 어떤 행동이 승리로 이어질 가능성이 높은지(승률)를 학습!

-바둑 Agent

모든 상태를 탐색한 수행 -> 어려움.

⇒정책 신경망(가능성 높은 수 선택), 가치 신경망(특정 상태에서의 승률 추정) 활용

일반화 -> value function approximation 등의 기법 사용(TD, Q-learning)

Sutton and Barto

Reinforcement learning problems involve learning what to do—how to map situations to actions—so as to maximize a numerical reward signal.

In an essential way they are closed-loop problems because the learning system's actions influence its later inputs. Moreover, the learner is not told which actions to take, as in many forms of machine learning, but instead must discover which actions yield the most reward by trying them out.

actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards.

<RL의 세 가지 특징>
being closed-loop in an essential way, not having direct instructions as to what actions to take, and where the consequences of actions, including reward signals, play out over extended time period

Sutton and Barto

(Exploration and Expoitation)
One of the challenges that arise in reinforcement learning, and not in other kinds of learning, is the trade-off between exploration and exploitation. To obtain a lot of reward, a reinforcement learning agent must prefer actions that it has tried in the past and found to be effective in producing reward. But to discover such actions, it has to try actions that it has not selected before. The agent has to exploit what it already knows in order to obtain reward, but it also has to explore in order to make better action selections in the future. The dilemma is that neither exploration nor exploitation can be pursued exclusively without failing at the task. The agent must try a variety of actions and progressively favor those that appear to be best.

All reinforcement learning agents have explicit goals, can sense aspects of their environments, and can choose actions to influence their environments. significant uncertainty about the environment it faces. When reinforcement learning involves planning, it has to address the interplay between planning and real-time action selection, as well as the question of how Moreover, it is usually assumed from the beginning that the agent has to operate despite environment models are acquired and improved.