Reinforcement learning

Geonhee Lee
gunhee6392@gmail.com

Outline

- Introduction to Reinforcement learning
- Markov Decision Process(MDP)
- Dynamic Programming(DP)
- Monte Carlo Method(MC)
- Temporal Difference Method(TD)
    - SARSA
    - Q-Learning
- Planning and Learning with Tabular Methods
- On-policy Control with with Approximation
- On-policy Prediction with Approximation
- Policy Gradient Method
- Actor Critic Method

Introduction to Reinforcement learn

RL 특성

다른 ML paradigms과의 차이점

- No supervisor, 오직 reward signal.
- Feedback이 즉각적이지 않고 delay 된다.
- Time이 큰 문제가 된다(연속적인, Independent and Identically Distributed(i.i.d, 독립항등분포) data가 아니다).
- Agent의 행동이 agent가 수용하는 연속적인 data에 영향을 준다.

Reward

- **Reward**: scalar feedback signal.
- agent가 step t에서 얼마나 잘 수행하는 지 나타냄.
- agent의 목표는 전체 reward의 합을 최대화하는 것

Sequential Decision Making

- Goal: Total future reward를 최대화하는 action 선택.
- Action들은 long term 결과들을 가질 것.
- Reward는 지연될 것.
- long-term reward를 더 크게 얻기 위해 즉각적인 reward를 희생하는 것이 나을 수도 있음.

History and State

- history: observations, actions, rewards의 연속.
- State: 다음에 어떤 일이 일어날 것인지 결정하기 위해 사용된 정보(다음 수식을 위한 정의로 보임)
- 공식으로는, state는 history의 함수이다.

$$S_t = f(H_t)$$

Information State

- Information state(a.k.a. Markov state)는 history로부터 모든 유용한 정보를 포함한다.

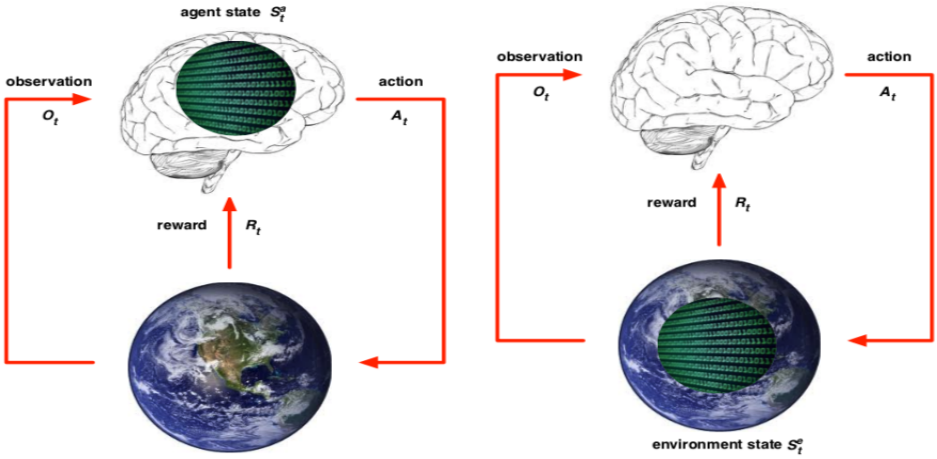
> Definition
>
> > state S_t는 Markov 이다 if and only if $P[S_{t+1}|S_t] = P[S_{t+1}|S_1, \ldots, S_t]$

- 미래는 현재의 과거와 독립적이다.
- State가 주어지면, history는 버려질 수 있다.

Fully Observable Environments

- **Full observability**: agent는 직접적으로 enviroment state를 관찰한다.

$$O_t = S_t^a = S_t^e$$



- Agent state = environment state = information state
- 형식적으로, 이것은 Markov decision precess(MDP).

Partially Observable Environments

- Partial observability: agent는 간접적으로 environment를 관찰.
  - robot이 카메라를 가지고 절대적인 위치를 알지못하는 것.
  - 포커를 하는 agent는 오직 오픈한 card들만 볼 수 있는 것
- 여기서 agent state $\neq$ environment state
- 형식적으로, 이것을 partially observable Markob decision process(POMDP)
- Agent는

Markov Decision Process(MDP)

Dynamic Programming(DP)

Monte Carlo Method(MC)

Temporal Difference Method(TD)

Planning and Learning with Tabular Methods

On-policy Control with with Approximation

Policy Gradient Method

Actor Critic Method

MP(X, P)

Reference

[1] UCL Course on RL [2] Reinforcement Learning: Tutorial(Seoul National University of Sceience and Technology)
[3] Reinforcement Learning : An Introduction, Sutton