AI1811 Reinforcement Learning Introduction

Xiyuan Yang

2025.11.20

Introduction to Reinforcement Learning, AI1811.

目录

§1. Introduction

The agent is interacting with the environment:

- Getting Observation: O_t
- Getting Reward R_t
- Doing Actions: A_t

§1.1. Simple Components

§1.1.1. History

A sequence of **observation, actions and rewards.**

$$H_t = O_1, R_1, A_1, O_2, R_2, A_2, ..., O_{t-1}, R_{t-1}, A_{t-1}, O_t, R_t$$

Based on the History:

- agents will operate actions A_t.
- environments will choose the observations and reward (based on A_t and H_t)

§1.1.2. State

State is a function (or mapping based on the history), due to the Markov Process.

$$S_t = f(H_t)$$

> **Recordings.**

- Observation is the sample of the environment and history.

§1.1.3. Policy

Policy is the mapping from State Space to Action Space. (For the current state s, we will do operations a)

- Deterministic Policy: $a = \pi(s)$
- Stochastic Policy: $\pi(a|s) = P(A_t = a|S_t = s)$

§1.1.4. Reward

Based on the state and the actions. (From the current state, we do specific actions, then we will get reward $R(s, a)$)

Based on the reward functions, we can define the value functions (in the long term):

$$\begin{aligned} Q_\pi(s, a) &= \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \ldots \mid S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma Q_\pi(s', a') | \ S_t = s, A_t = a] \end{aligned}$$

s' is the next state for $t + 1$.

Recordings.
在非确定性环境中，s_{t+1}是不确定的，哪怕s_t, a_t在给定的确定下，因此，上式的贝克曼方程需要对 s' 取期望。

Based on the model, we can predict:

- Predict the next state:

$$P^a_{ss'} = \mathbb{P}[S_{t+1} = s'|S_t = s, A_t = a]$$

- Predict the next step:

$$R^a_s = \mathbb{E}[R_{t+1}|S_t = s, A_t = a]$$

§1.2. Categories

- Value Based RL
  - 知道什么是好的 什么是坏的
  - 基于价值设计价值函数
- Policy Based
  - We have the policy but no value functions.
- Actor-Critic

§1.2.1. Value Based RL

The training goal is maximize the value function:

$$Q(s, a) = \mathbb{E}[R_{t+1}|s, a] + \gamma \max_{a'} Q(s', a')$$

Then, we do greedy search:

$$\pi(a|s) = \arg\max_a Q(s, a)$$

§1.2.2. Policy Based RL

Optimize $\pi_{\theta(a|s)}$ directly!

- REINFORCE
- Actor-Critic（A2C、A3C）
- PPO
- TRPO
- SAC（soft actor-critic）

§1.2.3. Advanced RL

- Deep Reinforcement RL: Use Deep Neural Network to do function approximations.
- Model-Based RL: learning with trajectory data in virtual environment.
- Goad-oriented RL (Long Horizon, Intermediate Steps and Simple Task Splitting)
- Imitation Learning
- Multi-Agent RL
- RL for Large Language Model Fine-tuning.

§1.3. Markov Process

Definition 1.3.1 Random State.

$$\mathbb{P}[S_{t+1}|S_1, S_2, ..., S_t]$$

§2. Conclusion