Notes for Reinforcement Learning

Synferlo

Apr. 4, 2020

Instructor: Emma Brunskill (Professor)

Univ: Stanford University

Dept: Computer Science

1 Introduction

DEF: Reinforcement Learning (RL) is a learning to make good sequence of decisions under uncertainty.

There are four parts get involved:

1. Optimization
2. Delayed Consequences
3. Exploration
4. Generalization

1.1 Intuition of RL

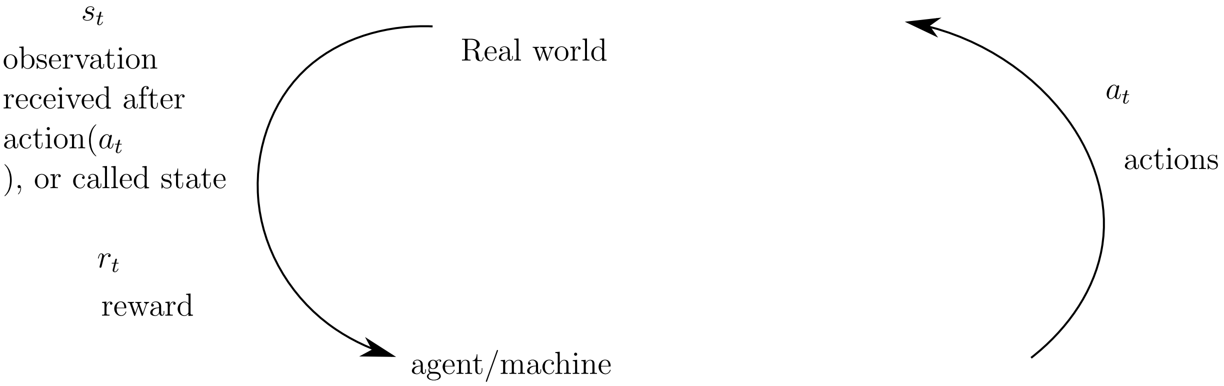

Figure 1: How RL work

The goal of RL is to maximized the expected discounted sum of the rewards.

Reward function: $r(s, a) \to \mathbb{R}$

1.2 Markov Process

In a dynamic model,

$$Pr(s_t|s_{t-1}, a_{t-1}, s_{t-2}, a_{t-2}, ...) = Pr(s_t|s_{t-1}, a_{t-1})$$

It says that the probability of state s_t appears given all previous states and actions is equal to the probability of s_t appears given only last period's state and action, s_{t-1}, a_{t-1}.
In Bandits case, it is just equal to $Pr(s)$,

$$Pr(s_t|s_{t-1}, a_{t-1}, s_{t-2}, a_{t-2}, ...) = Pr(s_t|s_{t-1}, a_{t-1}) = Pr(s)$$

1.2.1 Elements in MDP

There are five elements in Markov Decision Process (MDP).

1. $\mathcal{S}$: state space
2. A: action space
3. R: reward space, e.g., $r(s, a), r(s), r(s, a, s') \in \mathbb{R}$
4. T: dynamic model, e.g., $Pr(s'|s, a)$
5. γ: discounted factor, $\gamma \in (0, 1)$

1.3 Three methods for RL

There are three main approach in RL:

1. Model based:

Directly estimate and use R and T.

2. Value based:

Deal with the value function

$$V^{\pi} = \mathbb{E}_{s' \sim \pi} \left(\sum_{t=0}^{\infty} \gamma^{t} r_{t} \middle| s_{t=0} = s_{0} \right) \tag{1}$$

where s_0 stands for the initial state.

Clearly, equation(1) is the expected discounted sum of reward! And $s' \sim \pi$ indicates that if you follow policy π, then you will switch to state s' in the next period.

3. Policy based:

$$\arg\max_{\pi \in \Pi} V^{\pi}$$

The first two approaches leverage and assume Markovism, but Policy based approach does not!

The biggest difference between value based and policy based approach is how we compute the value function V^{π}.

1.4 Some Key Elements you need

1.5 Horizon

DEF:

Horizon H: number of times/steps/decisions/actions in an episode. It can be finite or infinity. Recall what we have learned in 805. In Macro, we normally assume infinite horizon.