0806

천정민

서울과학기술대학교
SEOULTECH SEOUL NATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

1. I.RL

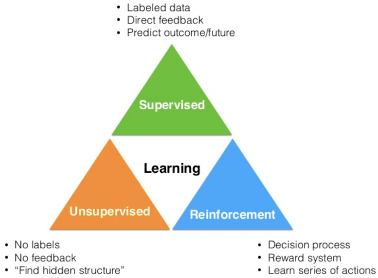
2. II. Markov chain

3. III. MRP

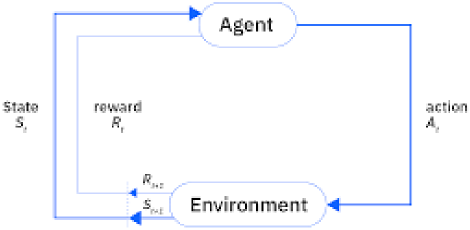
I.RL

Motivation

- What is RL?



- “시행착오를 통해 발전해 나가는 과정”
- “순차적 의사결정 문제에서 누적 보상을 최대화 하기 위해 시행착오를 통해 행동을 교정하는 학습 과정”



- 예를 들어 두발 자전거를 처음 배우는 A와 B가 있다고 하자.
- A는 부모님이 직접 타는 방법을 가르쳐주고, B는 스스로 아무 도움 없이 배우는 상황.
- 이 때, B를 agent라고 하면, B를 제외한 모든 것은 Environment로 볼 수 있음.
- B는 현재상태의 자전거의 위치, 기울어진 정도, 핸들의 각도 등등에 따라 넘어지거나, 그대로 균형을 유지하거나 등의 상태 변화를 겪음.

II. Markov chain

Motivation

- I drink a bottle of soda everyday. I drink either Coke or Pepsi everyday. When I choose what to drink for today, I only consider what I drank yesterday.

- Specifically,
  - Suppose I drank Coke yesterday, then the chance of drinking Coke again today is 0.7.
  - (What is the chance of drinking Pepsi today then?)
  - Suppose I drank Pepsi yesterday, then the chance of drinking Pepsi again today is 0.5.
  - (What is the chance of drinking Coke today then?)

Representation

- How would you describe this situation in diagram?

- How would you represent this situation to mathematical form?

$$\mathbf{P} = \begin{array}{c} \text{coke} \\ \text{pepsi} \end{array} \begin{pmatrix} 0.7 & 0.3 \\ 0.5 & 0.5 \end{pmatrix}$$

Definitions

- Stochastic process
  - Stochastic means time and randomness combined.
  - Stochastic process includes multiple random variables indexed by time.
- State: value of $S_\cdot$.
  - It may be deterministic.
    - Ex) $S_t = c \Leftrightarrow$ I drink coke on day-t, or say, "The state of S_t is c".
    - Ex) $S_1 = p \Leftrightarrow$ On day-1, I drink pepsi, or say, "The state of S_1 is pepsi".
  - It may be random. (not deterministic)
    - Ex) $\mathbb{P}(S_2 = p) = 0.6 \Leftrightarrow$ The probability that I drink pepsi on day-2 is 0.6.
  - It may be random and often described as a distribution.
    - Ex) $(\mathbb{P}(S_3 = c), \mathbb{P}(S_3 = p)) = (0.3, 0.7) \Leftrightarrow$ The probability that I drink coke on day-3 is 0.3 and pepsi is 0.7.
- State space: a set of all possible states that $S_\cdot$ can take.
  - Ex) A set of all possible kind of sodas that I might drink, i.e. $S = \{c, p\}$.

- Discrete time stochastic process
  - Discrete time stochastic process includes multiple random variables indexed by discrete time.
  - For example,
    - $S_0, S_1, S_2, \ldots$, where each implies day-0, day-1, and day-2,$\cdots$
    - $S_t, S_{t+1}, S_{t+2}, \ldots$, where each implies year-t, year-$t+1$,$\cdots$
  - Formally, $\{S_t : t \geq 0, t \in \mathbb{N}\}$

- Continuous time stochastic process
  - Continuous time stochastic process includes multiple random variables indexed by continuous time.
  - For example,
    - $\{S_t, t \in [0, \infty)\}$ where each implies daily or yearly evolution of certain quantity.
  - Formally, $\{S_t : t \in \mathbb{R}^+\}$

Markov Property

- Intuitively,
  - The nearest future only depends on the present. Past does not matter.
  - S_{t+1} depends only on the state of S_t.
  - S_{t+1} is function of S_t and some randomness, i.e. $S_{t+1} = f(S_t, \text{randomness})$.

- A bit rigorously,
  - The future only depends on the recent history that are known.
  - Future is independent of the past, given the present.

- Formally, Markov property holds if

$$\mathbb{P}(S_{t+1} = j | S_0 = i_0, S_1 = i_1, ..., S_t = i) = \mathbb{P}(S_{t+1} = j | S_t = i)$$

- Transitions depend only on the nearest past.
- Transitions depend only on the recent history.

III. MRP

Reward

- Let r_t be the spending on day-t. That is, r_t is cost or reward for time t.
- The reward r_t is fully determined by the state at time t, by a function $R(\cdot)$ such as $r_t = R(s)$.

Definition 1 (reward function)

A real-valued function $R : S \to \mathbb{R}$ is called a reward function that determines the reward given the state. That is, $R(s) = \mathbb{E}[r_t | S_t = s]$

Markov reward process (MRP)

Definition 2 (Markov reward process (MRP))

A MRP refers to a reward process where the underlying stochastic process is characterized with Markov property.

Remark 1

In other words, MRP is a reward process where the reward is determined by DTMC's state.

Return

- We were asked to find the expected value of $r_0 + r_1 + \cdots + r_9$.

Definition 3 (return)

The return G_t is the sum of remaining reward at time t.

- Using this notation, our problem has following returns.
  - $G_0 = r_0 + r_1 + \cdots + r_9$
  - $G_1 = r_1 + \cdots + r_9$
  - $G_2 = r_2 + \cdots + r_9$
  - $\cdots$
  - $G_9 = r_9$
- In other words, we were asked the value of $\mathbb{E}[G_0 | S_0 = c]$.

Dependence

- In our problem, we were asked to find the expected value of G_0 starting from state c at time 0.
- At time 0, the value of r_0 is known, but $r_1, ..., r_9$ are random variables. So, G_0 is random variable as well.

- The random variable G_0 depends on
  - the current state S_0
  - and the randomness along the stochastic path.
- In general, the random variable G_t depends on
  - the last-known state S_t
  - and some randomness along the remaining path.

- Since G_t is a random variable, we want to evaluate $\mathbb{E}[G_t]$.
- In general, considering its dependence structure, we are interested in evaluating $\mathbb{E}[G_t|S_t = s]$.

State-value function

- The current problem is $\mathbb{E}[r_0 + r_1 + \cdots + r_9 | S_0 = c]$ or $\mathbb{E}[G_0 | S_0 = c]$.
- This motivates the following definition.

Definition 4 (state-value function)

A state-value function $V_t(s)$ is the expected return given state s at time t. That is, $V_t(s) = \mathbb{E}[G_t | S_t = s]$

- Then, we are interested in finding $V_0(c) = \mathbb{E}[G_0 | S_0 = c] = \mathbb{E}[r_0 + \cdots + r_9 | S_0 = c]$.

Next week···

- How to calculate state value function?
- using Bellman equation which is the most important equation for Reinforcement Learning

```
"Thank you for listening"
```