Machine Learning in Practice

#5: Reinforcement Learning #1: Exercise

Summer 2019

Definition 1 (Markov Process).

A **Markov process** *(or Markov chain) is a pair* $(\mathcal{S}, \mathcal{P})$

- $\mathcal{S}$ *: a (finite) set of states*
- $\mathcal{P} : \mathcal{S}^2 \to [0,1]$ *: a state transition probability*

Markov reward process = Markov process + reward

Definition 2 (Markov Reward Process).

A **Markov reward process** *(***MRP***) is a tuple* $(\mathcal{S}, \mathcal{P}, R, \gamma)$ *where*

- $\mathcal{S}$ *: a (finite) set of states*
- $\mathcal{P} : \mathcal{S}^2 \to [0,1]$ *: a state transition probability*
  - $(\mathcal{S}, \mathcal{P})$ *constitutes a Markov process*
- $R : \mathcal{S} \to \mathbb{R}$ *: a* **reward** *function*
  - $R(s)$ *represents the expected intermediate reward at next state*
  - *현재 state* s *에서의 intermediate reward로 해석해도 됨*
- $\gamma \in [0,1]$ *: a discount factor*

Definition 3 (State-Value Function of Markov Reward Process).

Given an MRP $(\mathcal{S}, \mathcal{P}, R, \gamma)$*, its* **state-value function** $v : \mathcal{S} \to \mathbb{R}$ *is*

- $v(s) = \mathbb{E}\Big[R(s) + \gamma R(N_1(s)) + \gamma^2 R(N_2(s)) + \cdots\Big] = \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k R\big(N_k(s)\big)\Big]$

where $N_k(s)$ *is the random variable describing the state after* k *steps from* s*, i.e.*

- $\mathbb{P}[N_k(s) = s'] = \sum_{s_i \in \mathcal{S}} \big(\mathcal{P}(s, s_1) \cdot \mathcal{P}(s_1, s_2) \cdot \ldots \mathcal{P}(s_{k-1}, s')\big)$

Markov decision process = Markov reward process + action

Definition 4 (Markov Decision Process).

A **Markov decision process** *(***MDP***) is a tuple* $(\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma)$ *where*

- $\mathcal{S}$ *: a (finite) set of states*
- $\mathcal{A}$ *: a (finite) set of* **actions**
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ *: a state transition probability*
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *: a reward function*
  - $R(s,a)$*는 현재 state* s*에서 action* a*를 택했을 때 다음 state들에서 받을 수 있는 expected reward를 나타냄*
- $\gamma \in [0,1]$ *: a discount factor*

Definition 5 (Policy).

A **policy** *of an MDP* $(\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma)$ *is a probability distribution over actions given states:*

- $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$
  - *non-random policy의 경우* $\pi : \mathcal{S} \to \mathcal{A}$

Given an MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma)$ and a policy π,

- the state/reward sequence is a Markov reward process $(S, \mathcal{P}^\pi, R^\pi, \gamma)$ where

  $\boxed{\mathcal{P}^\pi(s,s') = \sum_{a \in \mathcal{A}} \pi(s,a) \cdot \mathcal{P}(s,a,s')}$

  $\boxed{R^\pi(s) = \sum_{a \in \mathcal{A}} \pi(s,a) \cdot R(s,a)}$

  $\boxed{N_k^\pi(s) \triangleq \text{R.V. describing state after } k \text{ steps from } s \text{ under } \pi}$

  - $\mathbb{P}[N_k^\pi(s) = s'] = \sum_{s_i \in \mathcal{S}} \left(\mathcal{P}^\pi(s,s_1) \cdot \mathcal{P}^\pi(s_1,s_2) \cdot \ldots \mathcal{P}^\pi(s_{k-1},s') \right)$

Definition 6 (State-Value/Action-Value Functions of Markov Decision Process).

The **state-value function** $v^\pi : \mathcal{S} \to \mathbb{R}$ *is*

- $v^\pi(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R^\pi\left(N_k^\pi(s)\right) \right]$
  - *i.e. expected total reward starting from* s*, and then following* π

The **action-value function** $q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *is*

- $q^\pi(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} \left(\mathcal{P}(s,a,s') \cdot v^\pi(s') \right)$
  - *i.e. expected total reward starting from* s*, taking action* a*, and following* π

1. Derive the **recursive formula for the state-value function** from Definition 6:

- $\boxed{v^\pi(s) = R^\pi(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^\pi(s,s') \cdot v^\pi(s')}$

Hint: use $\boxed{v^\pi(s) = \sum_{a \in \mathcal{A}} \pi(s,a) \cdot q^\pi(s,a)}$

Definition 7 (Optimal State-Value/Action-Value Functions).

The **optimal state-value** *function $v^* : \mathcal{S} \to \mathbb{R}$ is defined by*

- $v^*(s) = \max\{v^\pi(s) \,|\, \text{policy } \pi \text{ of the MDP}\}$ *for each $s \in \mathcal{S}$*

The **optimal action-value** *function $q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined by*

- $q^*(s,a) = \max\{q^\pi(s,a) \,|\, \text{policy } \pi\}$ *for each $s \in \mathcal{S}, a \in \mathcal{A}$*

Definition 8 (Optimal Policy).

A **policy** *π^* of an MDP is said to be* **optimal** *if, for all policy π,*

- $v^{\pi^*}(s) \geq v^\pi(s)$ *for all $s \in \mathcal{S}$*

Theorem 9. *For every MDP,*

- *there exists an optimal policy*
- *every optimal policy π^* achieves the optimal state-value function, i.e. $v^{\pi^*}(s) = v^*(s)$ for all $s \in \mathcal{S}$*
- *every optimal policy π^* achieves the optimal action-value function, i.e. $q^{\pi^*}(s,a) = q^*(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$*

2. Derive the **Bellman optimality equations** for the optimal value functions:

(a) $\boxed{v^*(s) = \max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s,a,s') \cdot v^*(s') \right)}$

(b) $\boxed{q^*(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s,a,s') \cdot \left(\max_{a' \in \mathcal{A}} q^*(s',a') \right)}$

Hint: use Definition 6 and $\boxed{v^*(s) = \max_{a \in \mathcal{A}} q^*(s,a)}$