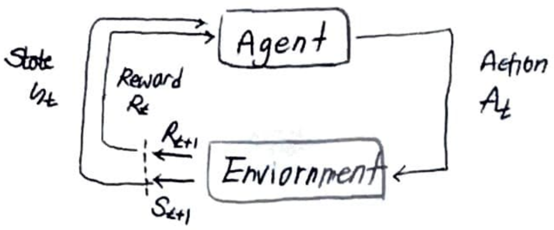
Reinforcement Learning

강화학습의 모든 문제는 다음과 같은 틀로 귀결된다.



Environment 이 Agent에게 <u>특정상황 State</u>를 주면 / Agent는 그에 대해 <u>반응action</u>을 하고 / env는 agent에게 <u>보상reward</u>를 준다.

Value Function

policy

- Model : Mathematic models of dynamics and rewards.

π · Policy : Function mapping agent's states to action　action을 취하는 방법론 $s \xrightarrow{a_t} s'$

q, V · Value Function : future rewards from being in a state and / or action when following a particular policy

$V(S)$

ex State-value function, state-action value function

R · Reward R_s^a , $R(S_t = s, a_t = a)$

G · Return G_t discounted sum of rewards from time step t

- State transition Matrix

$$P_{ss'} = P[S_{t+1} = s' \mid S_t = s]$$

$$P = \begin{pmatrix} P(S_1|S_1) & P(S_2|S_1) & \cdots & P(S_N|S_1) \\ P(S_1|S_2) & P(S_2|S_2) & \cdots & P(S_N|S_2) \\ \vdots & & & \vdots \\ P(S_1|S_N) & P(S_2|S_N) & \cdots & P(S_N|S_N) \end{pmatrix} = \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix}$$

1 · Markov Property

State S_t is Markov $\Leftrightarrow$ $P(S_{t+1} \mid S_t, a_t) = P(S_{t+1} \mid h_t, a_t)$

현재의 state가 이전의 state에만 영향을 준다.

공식정리 Bellman Expectation Equation

return $G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \quad = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

reward $R_s = E[r_t \mid S_t = s]$

$R_s^a = E[r_t \mid S_t = s, a_t = a]$

policy $\pi(a|s) = P(a_t = a, S_t = s)$

Value Function

state-value $V(s) = E[G_t \mid S_t = s]$

$= E[R_t + \gamma V(S_{t+1}) \mid S_t = s]$

$= R(s) + \gamma \sum_{s' \in S} P(s'|s) V(s')$ ✓

$= R + \gamma P V$

$V_\pi(s) = E_\pi[r + \gamma E_\pi[G_{t+1} \mid S_{t+1} = s'] \mid S_t = s]$

$= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s,a) [r + \gamma E_\pi[G_{t+1} \mid S_{t+1} = s']]$

$= \sum_a \pi(a|s) [R_s^a + \gamma \sum_{s'} P_{ss'}^a V_\pi(s')]$ ✓

$= \sum_a \pi(a|s) q_\pi(s,a) \rightarrow V_\pi = R^\pi + \gamma P^\pi V_\pi$

State-action value $q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a]$

$= R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_\pi(s')$

Bellman Optimal Equation (max로 차이)

$V_*(s) = \max_\pi V_\pi(s)$

$q_*(s,a) = \max_\pi q_\pi(s,a)$

결정시 $\pi_*(a|s) = \begin{cases} 1 & \arg\max_a q_*(s,a) \\ 0 & \end{cases}$

$V_*(s) = \max_a q_*(s,a)$

$= \max_a R_s^a + \gamma \sum_{s'} P_{ss'}^a V_*(s')$

- there is always a deterministic policy of MDP. (unique 하지 않아도 가능).
- non-linear

Markov Process

$< S, P >$

No reward. No action
only state

Markov Reward Process

$< S, P, R, \gamma >$

$< S, P^{\pi}, R^{\pi}, \gamma >$

No action

Markov Decision Process

$< S, A, P, R, \gamma >$

$$P_{ss'} = P[S_{t+1} = s' \mid S_t = s]$$

$$P_{ss'}^{a} = P[S_{t+1} = s' \mid S_t = s, a_t = a]$$

$$P_{ss'}^{\pi} = \sum_{a} \pi(a \mid s) P_{ss'}^{a}$$

$$V(s) = E[G_t \mid S_t = s]$$

$$= R(s) + \gamma \sum_{s' \in S} P(s' \mid s) V(s')$$

$$\left(\begin{array}{l} V_{\pi} = R^{\pi} + \gamma P^{\pi} V_{\pi}' \\ q_{\pi} = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^{a} V_{\pi}(s') \end{array} \right) \longrightarrow$$

$$V_k^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s' \mid s, \pi(s)) \cdot V_{k-1}^{\pi}(s')$$

$$q_{\pi}(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^{a} V_{\pi}(s')$$