

1 5 기 정 규 세 션

ToBig's 14기 장혜림

강화학습

contents

Unit 00 | Intro

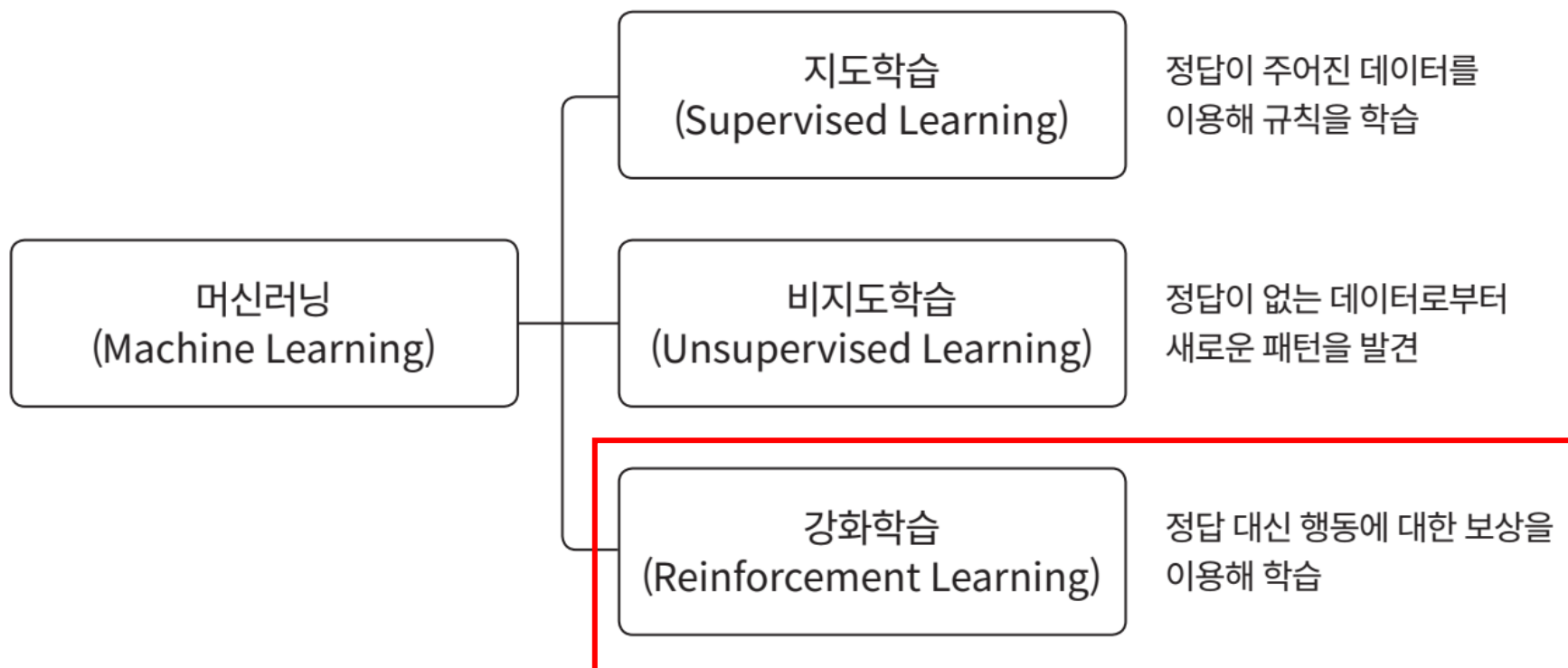
Unit 01 | 강화학습 기본 요소

Unit 02 | MDP / Bellman Equation

Unit 03 | 동적계획법

Unit 04 | Model-free algorithm(Q-learning)

Unit 00 | Intro



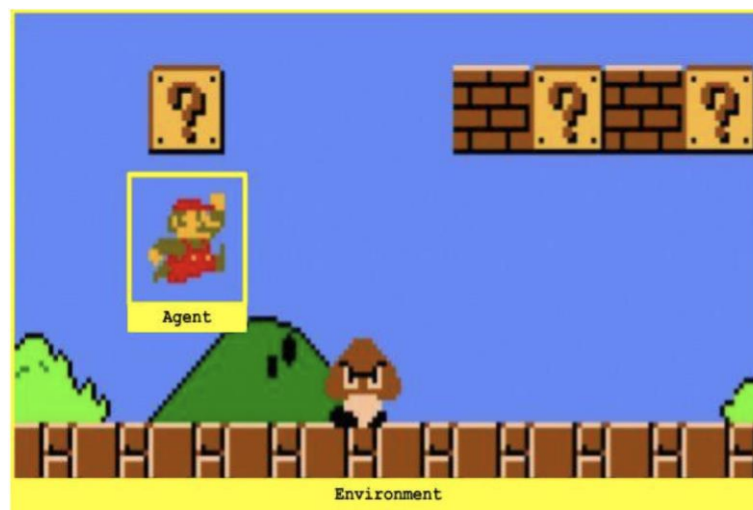
Unit 00 | Intro

Trial and Error를 통해 학습



Unit 00 | Intro

강화학습 활용 분야



Unit 01 | 강화학습 기본 요소

환경

상태

에이전트

행동

상태전이확률

보상

수익

정책

에피소드

Unit 01 | 강화학습 기본 요소

환경

- 환경(Environment)은 강화학습을 이용해 풀고자 하는 대상

ex. 미로 탐색 → 환경 = 미로

자동 주식 트레이딩 → 환경 = 주식시장

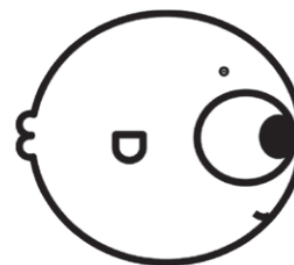
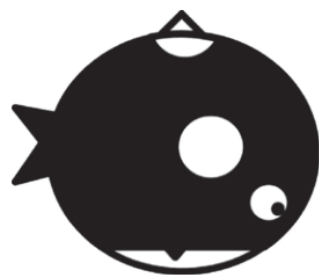
- 미로 탐색 문제를 예시로 사용 !

출발		
		도착

Unit 01 | 강화학습 기본 요소

에이전트

- 환경에 대해 특정 행동을 수행하고 학습하는 프로그램 또는 로봇
- 에이전트가 환경에 대해 여러 가지 행동을 반복하면서 최적의 행동을 학습함



에이전트

Unit 01 | 강화학습 기본 요소

상태

- 상태(State)는 에이전트가 위치하거나 감지하고 있는 정보
ex. 미로 : 로봇(에이전트)의 위치
- 상태 S는 모든 상태의 집합

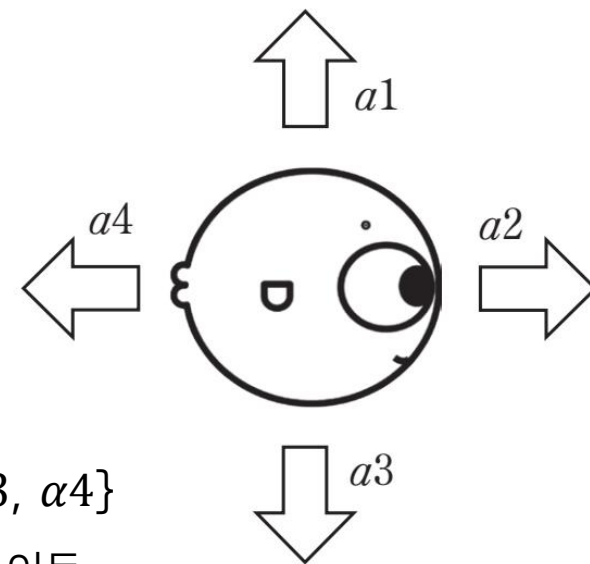
s_0	s_1	s_2
s_3	s_4	s_5
s_6	s_7	s_8

$$S = \{s_0, s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8\}$$

Unit 01 | 강화학습 기본 요소

행동

- 에이전트가 상태 S 에서 취할 수 있는 행동(Action)
ex. 미로: 상하좌우
- 에이전트가 어떤 상태에서 어떤 행동을 취하면 새로운 상태로 이동함



$$A = \{a1, a2, a3, a4\}$$

$a1$ = 위로 한 칸 이동

$a2$ = 오른쪽으로 한 칸 이동

$a3$ = 아래로 한 칸 이동

$a4$ = 왼쪽으로 한 칸 이동

Unit 01 | 강화학습 기본 요소

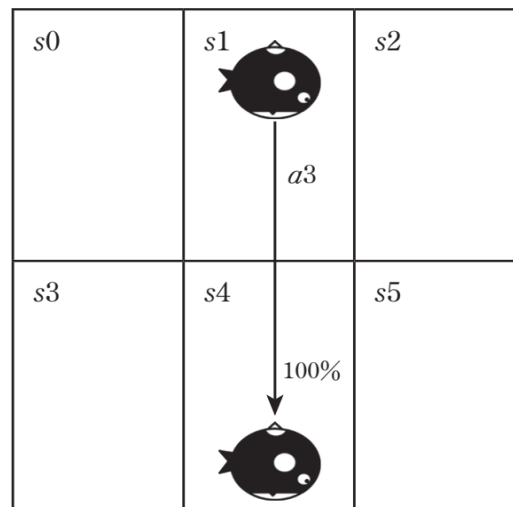
상태전이확률

- 상태전이확률(State transition probability)은 시간 t 일 때 에이전트가 상태 s 에서 행동 a 를 취했을 때 s' 로 이동할 확률

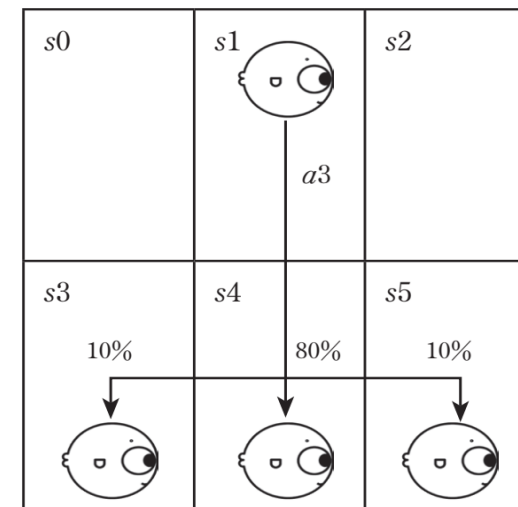
$$P(s'|s, a) = P_{ss'}^a = Pr[S_{t+1} = s' | S_t = s, A_t = a]$$

$$P(s_4 | s_1, a_3) = 1$$

: s_1 에서 a_3 행동을 취하면 100% 상태 s_4 로 이동



a) 결정론적 환경



b) 확률적 환경

ex. 빗길에
운전을 하는 경우

직진을 하려는
운전자의 의지와
상관없이 차가 좌우로
미끄러지는 경우

Unit 01 | 강화학습 기본 요소

보상

- 보상(Reward)은 에이전트가 취한 행동에 대한 평가
- 미로 탐색에서 보상 정의
 - 도착지점에 도착하는 행동은 +1의 보상을 받는다.
 - 미로 밖으로 나가는 행동은 -3의 보상을 받는다.
 - 미로 내에서 이동하는 행동은 -1의 보상을 받는다.

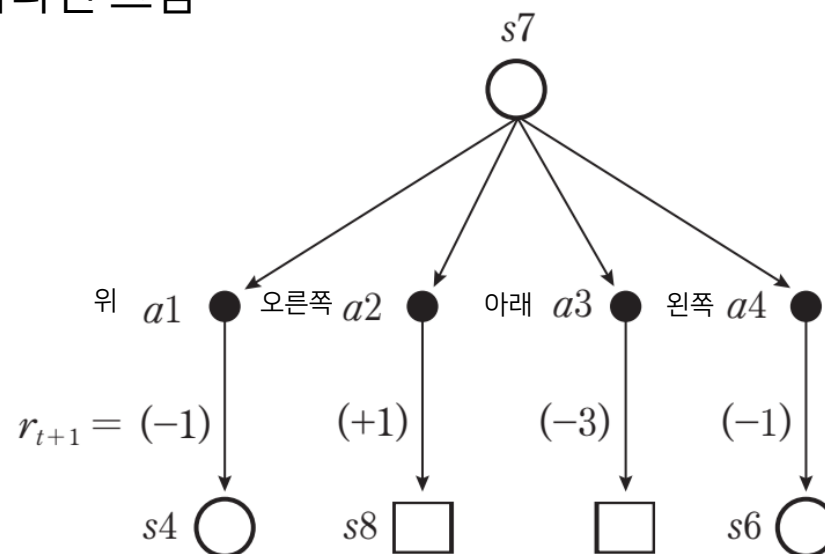
s_0 (-1)	s_1 (-1)	s_2 (-1)
s_3 (-1)	s_4 (-1)	s_5 (-1)
s_6 (-1)	s_7 (-1)	s_8 (+1)

각 상태 이동(도착)에 따른 보상 정의

Unit 01 | 강화학습 기본 요소

보상

- 백업 다이어그램으로 표현
: 상태와 행동, 다음 상태의 관계를 나타낸 그림



○ : 상태(state)

● : 행동(action)

□ : 마지막 상태(terminal state)

$s0$ (-1)	$s1$ (-1)	$s2$ (-1)
$s3$ (-1)	$s4$ (-1)	$s5$ (-1)
$s6$ (-1)	$s7$ (-1)	$s8$ (+1)

Unit 01 | 강화학습 기본 요소

수익

- 수익(Gain)은 시간 t 로부터 에이전트가 계속적으로 행동을 취한다고 가정했을 때, 얻게 되는 보상의 총합
- 감가율 γ 의 크기에 따라 계속 받는 보상의 가치를 정의할 수 있음
 - γ 이 0에 가까우면 현재의 보상을 중요시
 - γ 이 1에 가까우면 먼 미래의 보상까지 고려

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$
$$0 \leq \gamma \leq 1$$

Unit 01 | 강화학습 기본 요소

정책

- 정책(Policy)은 에이전트가 어떤 상태에 있을 때 어떤 행동을 선택할지 결정하는 기준
- 시간이 t 일 때 상태 s 에서 행동 a 를 선택할 확률

$$\pi(a|s) = Pr[A_t = a | S_t = s]$$

Unit 01 | 강화학습 기본 요소

정책

- 정책(Policy)은 에이전트가 어떤 상태에 있을 때 어떤 행동을 선택할지 결정하는 기준
- 시간이 t 일 때 상태 s 에서 행동 a 를 선택할 확률

$$\pi(a|s) = Pr[A_t = a | S_t = s]$$

강화학습의 목표 = 최적 정책(Optimal Policy)을 찾는 것

* Optimal Policy : 수익이 최대가 되는 행동을 선택하는 정책

Unit 01 | 강화학습 기본 요소

에피소드

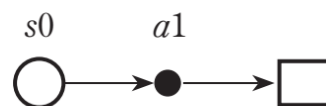
- 초기 상태에서부터 시작해서 목표 상태에 도착하거나 실패로 인한 상태 종료까지의 일련의 과정

s_0 (-1)	s_1 (-1)	s_2 (-1)
s_3 (-1)	s_4 (-1)	s_5 (-1)
s_6 (-1)	s_7 (-1)	s_8 (+1)

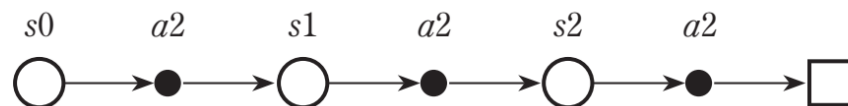
에피소드 1



에피소드 2



에피소드 3



미로 밖으로 나가게 되어 종료되는 과정

contents

Unit 00 | Intro

Unit 01 | 강화학습 기본 요소

Unit 02 | MDP / Bellman Equation

Unit 03 | 동적계획법

Unit 04 | Model-free algorithm(Q-learning)

Unit 02 | MDP / Bellman Equation

MDP (Markov Decision Process, 마르코프 의사결정 과정)

- 강화학습은 MDP로 정의된 문제를 푸는 것과 같음
- MDP는 앞서 배운 기본 요소로 구성되어 있음

 $\langle S, A, P, R, \gamma \rangle$ S : State A : Action P : state transition Probability R : Reward γ : discount factor목적 : 출발지점(s_0)에서 출발해서 도착지점(s_8)에 도착

s_0 (출발)	s_1	s_2
s_3	s_4	s_5
s_6	s_7	s_8 (도착)

 $S = \{s_0, s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8\}$ $A = \{ \text{위: } a_1, \text{오른쪽: } a_2, \text{아래: } a_3, \text{왼쪽: } a_4 \}$ $P(s'|s, a) = 1$ $R = \begin{cases} 1 : s_8 \text{에 도착} \\ -1 : \text{미로 안에서 이동} \\ -3 : \text{미로 밖으로 이동} \end{cases}$ $\gamma = 0.9$

Unit 02 | MDP / Bellman Equation

MDP (Markov Decision Process, 마르코프 의사결정 과정)

- 강화학습은 MDP로 정의된 문제를 푸는 것과 같음
- MDP는 앞서 배운 기본 요소로 구성되어 있음

⇒ 수익이 최대가 되는 행동을 선택하도록 학습

$\langle S, A, P, R, \gamma \rangle$

S : State

A : Action

P : state transition Probability

R : Reward

γ : discount factor



목적 : 출발지점(s_0)에서 출발해서 도착지점(s_8)에 도착

s_0 (출발)	s_1	s_2
s_3	s_4	s_5
s_6	s_7	s_8 (도착)

$S = \{s_0, s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8\}$

$A = \{ \text{위: } a_1, \text{오른쪽: } a_2, \text{아래: } a_3, \text{왼쪽: } a_4 \}$

$P(s'|s, a) = 1$

$R = \begin{cases} 1 : s_8 \text{에 도착} \\ -1 : \text{미로 안에서 이동} \\ -3 : \text{미로 밖으로 이동} \end{cases}$

$\gamma = 0.9$

Unit 02 | MDP / Bellman Equation

MDP (Markov Decision Process, 마르코프 의사결정 과정)

- 강화학습은 MDP로 정의된 문제를 푸는 것과 같음
- MDP는 앞서 배운 기본 요소로 구성되어 있음

⇒ 수익이 최대가 되는 행동을 선택하도록 학습
어떻게? 상태와 행동의 가치를 계산!

 $\langle S, A, P, R, \gamma \rangle$

S : State

A : Action

P : state transition Probability

R : Reward

γ : discount factor



목적 : 출발지점(s_0)에서 출발해서 도착지점(s_8)에 도착

s_0 (출발)	s_1	s_2
s_3	s_4	s_5
s_6	s_7	s_8 (도착)

$S = \{s_0, s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8\}$

$A = \{ \text{위: } a_1, \text{오른쪽: } a_2, \text{아래: } a_3, \text{왼쪽: } a_4 \}$

$P(s'|s, a) = 1$

$R = \begin{cases} 1 : s_8 \text{에 도착} \\ -1 : \text{미로 안에서 이동} \\ -3 : \text{미로 밖으로 이동} \end{cases}$

$\gamma = 0.9$

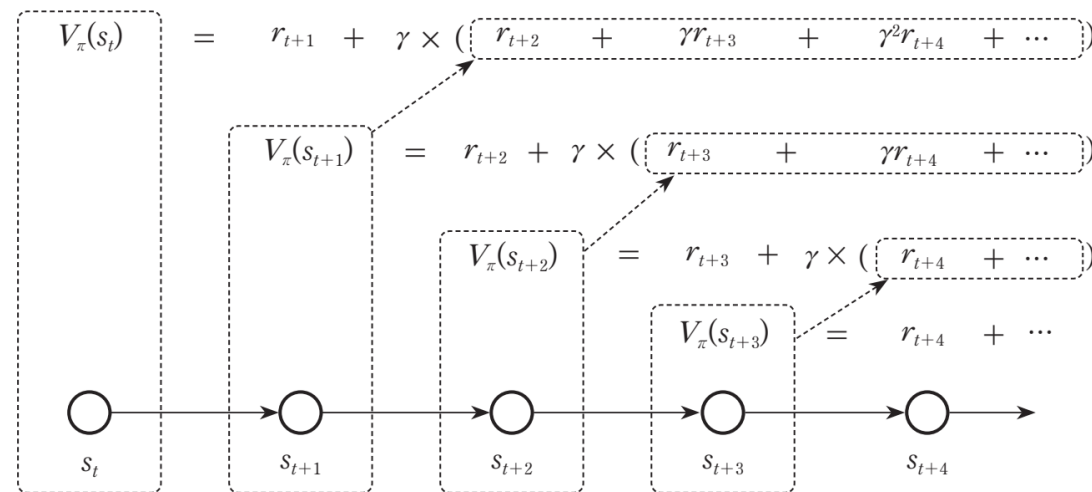
Unit 02 | MDP / Bellman Equation

1. 상태 가치 함수 (state-value function)

- 상태가치 $V(s)$ = 시간 t 일 때 상태 s 에서의 기대수익 G_t
- 벨만 방정식 (Bellman Equation)

: 현재 시간이 t 일 때 상태 s_t 의 상태가치 $V_\pi(s_t)$ 와 다음 상태 s_{t+1} 의 상태가치 $V_\pi(s_{t+1})$ 의 관계를 식으로 나타낸 것

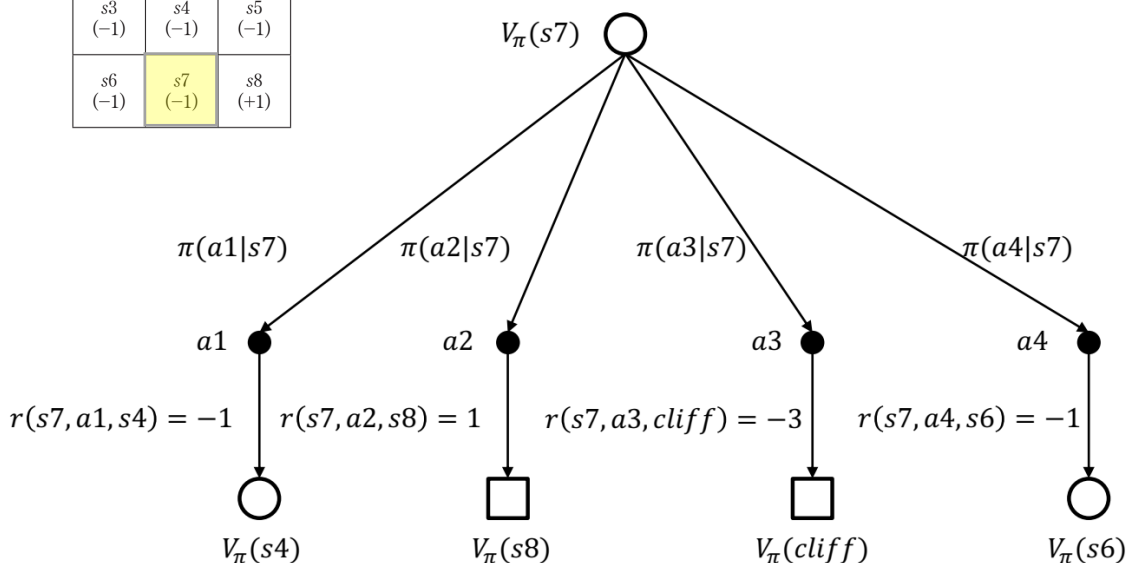
$$\begin{aligned}
 V_\pi(s) &= E[G_t | S_t = s] \\
 &= E[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} \dots | S_t = s] \\
 &= E[r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \gamma^2 r_{t+4} \dots) | S_t = s] \\
 &= E[r_{t+1} + \gamma G_{t+1} | S_t = s] \\
 &= E[r_{t+1} + \gamma V_\pi(s_{t+1}) | S_t = s] \\
 &= \sum_a \pi(a|s_t) \sum_{s_{t+1}} P(s_{t+1}|s_t, a) [r(s_t, a, s_{t+1}) + \gamma V_\pi(s_{t+1})]
 \end{aligned}$$



Unit 02 | MDP / Bellman Equation

1. 상태 가치 함수 (state-value function)

s_0 (-1)	s_1 (-1)	s_2 (-1)
s_3 (-1)	s_4 (-1)	s_5 (-1)
s_6 (-1)	s_7 (-1)	s_8 (+1)



$$\begin{aligned}
 V_\pi(s) &= E[G_t | S_t = s] \\
 &= E[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} \dots | S_t = s] \\
 &= E[r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \gamma^2 r_{t+4} \dots) | S_t = s] \\
 &= E[r_{t+1} + \gamma G_{t+1} | S_t = s] \\
 &= E[r_{t+1} + \gamma V_\pi(s_{t+1}) | S_t = s] \\
 &= \sum_a \pi(a|s_t) \sum_{s_{t+1}} P(s_{t+1}|s_t, a) [r(s_t, a, s_{t+1}) + \gamma V_\pi(s_{t+1})]
 \end{aligned}$$

* 기댓값이란?

각 사건이 발생했을 때의 수익과 그 사건이
벌어질 확률을 곱한 것을 전체 사건에 대해
합한 값

→ 어떤 확률적 사건에 대한 평균 의미

$$\begin{aligned}
 V_\pi(s_7) &= \pi(a1|s7) \times P(s4|s7, a1) \times \{r(s7, a1, s4) + \gamma V_\pi(s4)\} \\
 &\quad + \pi(a2|s7) \times P(s8|s7, a2) \times \{r(s7, a2, s8) + \gamma V_\pi(s8)\} \\
 &\quad + \pi(a3|s7) \times P(cliff|s7, a3) \times \{r(s7, a3, cliff) + \gamma V_\pi(cliff)\} \\
 &\quad + \pi(a4|s7) \times P(s6|s7, a4) \times \{r(s7, a4, s6) + \gamma V_\pi(s6)\}
 \end{aligned}$$

Unit 02 | MDP / Bellman Equation

2. 행동 가치 함수 (action-value function)

- '어떤 상태 s 에서 최적의 행동 a 는 무엇인지'를 구하고 싶음!

$$\begin{aligned}
 Q_{\pi}(s, a) &= E[G_t | S_t = s, A_t = a] \\
 &= E[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} \dots | S_t = s, A_t = a] \\
 &= E[r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \gamma^2 r_{t+4} \dots) | S_t = s, A_t = a] \\
 &= E[r_{t+1} + \gamma V_{\pi}(s') | S_t = s, A_t = a] \\
 &= \sum_{s' \in S} P(s' | s, a) [r(s, a, s') + \gamma V_{\pi}(s')]
 \end{aligned}$$

$$\begin{aligned}
 V_{\pi}(s) &= \sum_a \pi(a|s) \sum_{s'} P(s'|s, a) [r(s, a, s') + \gamma V_{\pi}(s')] \\
 Q_{\pi}(s, a) &= \sum_{s' \in S} P(s'|s, a) [r(s, a, s') + \gamma V_{\pi}(s')]
 \end{aligned}$$



$$V_{\pi}(s) = \sum_a \pi(a|s) Q_{\pi}(s, a)$$

$$Q_{\pi}(s, a) = \sum_{s' \in S} P(s'|s, a) \left[r(s, a, s') + \gamma \sum_a \pi(a'|s') Q_{\pi}(s', a') \right]$$

Unit 02 | MDP / Bellman Equation

2. 행동 가치 함수 (action-value function)

- '어떤 상태 s 에서 최적의 행동 a 는 무엇인지'를 구하고 싶음!

$$Q_{\pi}(s, a) = E[G_t | S_t = s, A_t = a]$$

$$= E[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \dots | S_t = s, A_t = a]$$

$$= E[r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \gamma^2 r_{t+4} \dots) | S_t = s, A_t = a]$$

$$= E[r_{t+1} + \gamma V_{\pi}(s') | S_t = s, A_t = a]$$

$$= \sum_{s' \in S} P(s' | s, a) [r(s, a, s') + \gamma V_{\pi}(s')] = \arg \max_a Q_{\pi}(s, a)$$

최적 정책: 행동가치가 가장 큰 행동을 선택하는 정책

$$V_{\pi}(s) = \sum_a \pi(a|s) \sum_{s' \in S} P(s'|s, a) [r(s, a, s') + \gamma V_{\pi}(s')]$$

$$= \sum_{a \in A} \pi(a|s) \sum_{s' \in S} P(s'|s, a) [r(s, a, s') + \gamma V_{\pi}(s')]$$



$$V_{\pi}(s) = \sum_a \pi(a|s) Q_{\pi}(s, a)$$

$$Q_{\pi}(s, a) = \sum_{s' \in S} P(s'|s, a) \left[r(s, a, s') + \gamma \sum_{a'} \pi(a'|s') Q_{\pi}(s', a') \right]$$

contents

Unit 00 | Intro

Unit 01 | 강화학습 기본 요소

Unit 02 | MDP / Bellman Equation

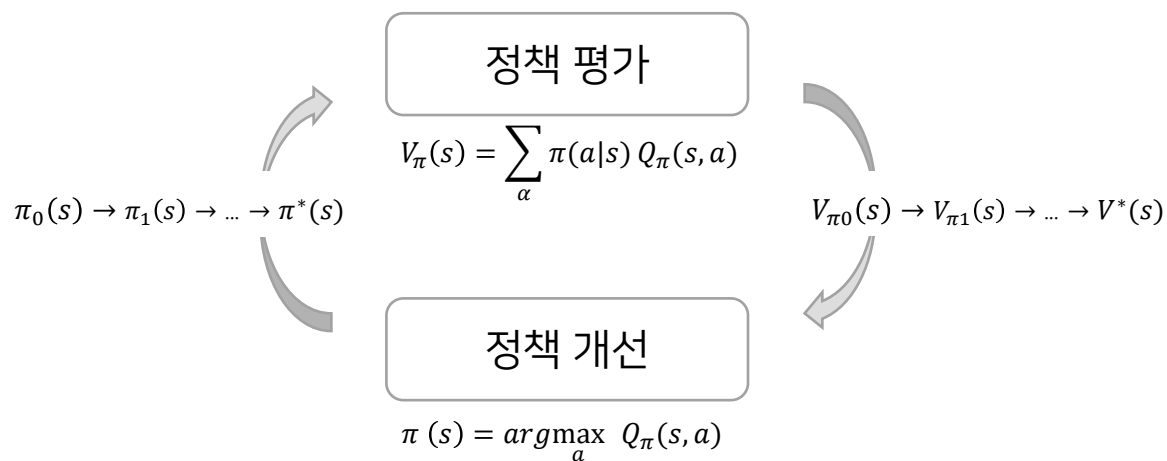
Unit 03 | 동적계획법

Unit 04 | Model-free algorithm(Q-learning)

Unit 03 | 동적계획법

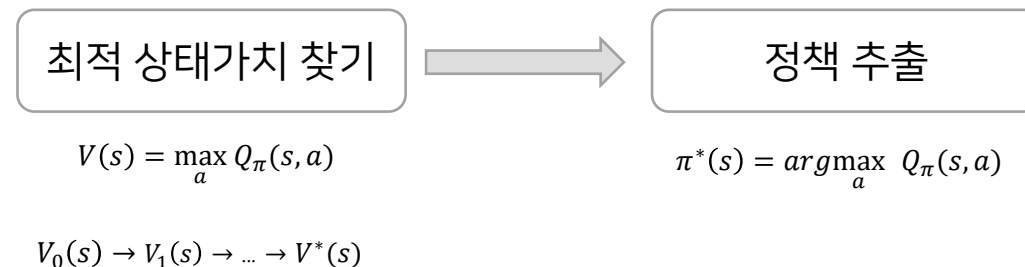
정책 반복

정책 π 를 이용해 상태가치함수 V_π 를 평가(evaluation)하고,
그 가치함수를 이용해 다시 정책 π 를 개선(improvement)하는 과정을
반복하면서 최적의 가치함수와 최적의 정책을 찾아가는 알고리즘



가치 반복

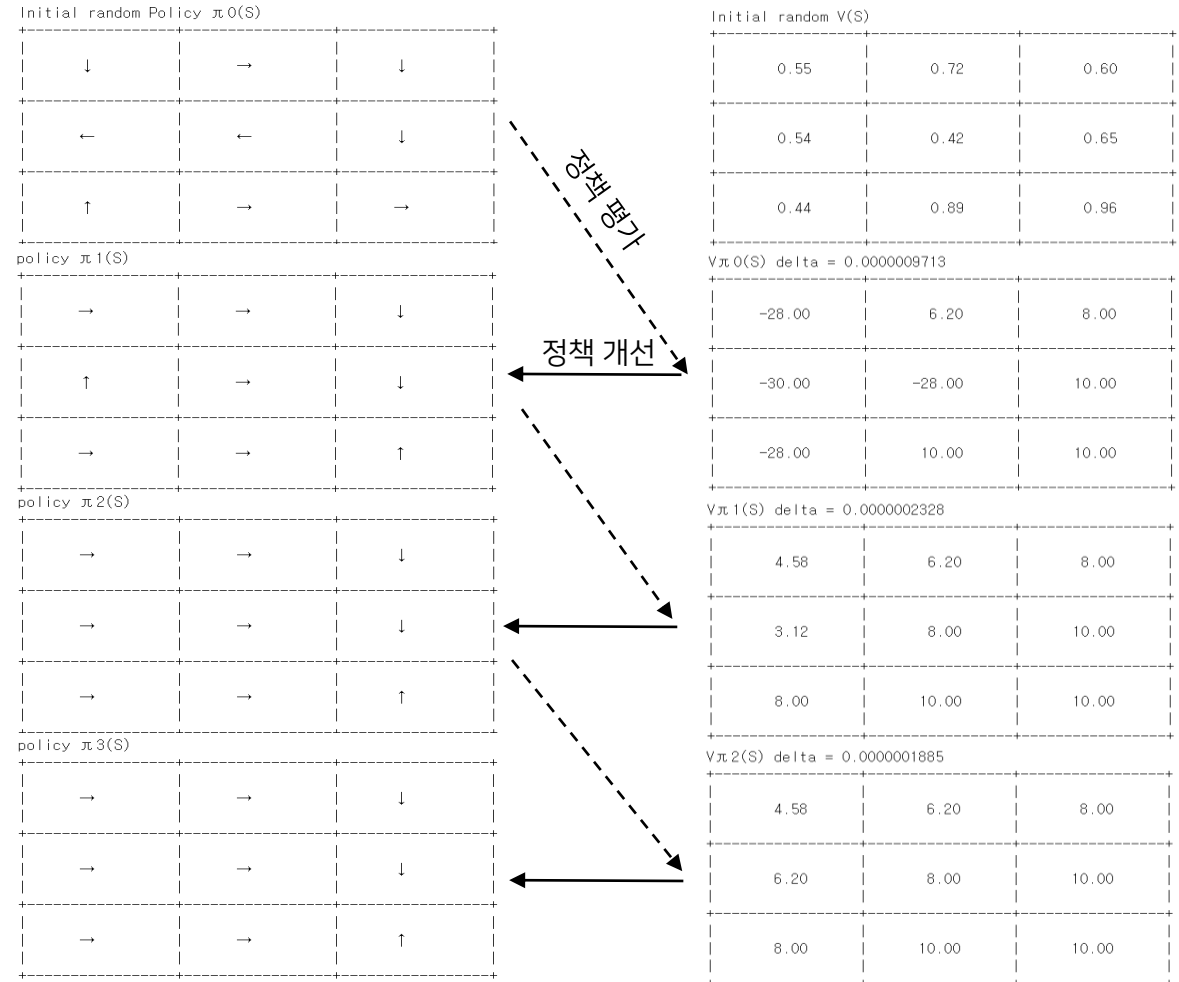
최적의 상태가치를 찾을 때까지 반복하고,
최적의 상태가치를 찾으면 그로부터 최적 정책을 추출하는 알고리즘



Unit 03 | 동적계획법

정책 반복 (Policy Iteration)

1. Initialization
 $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$
2. Policy Evaluation
 Repeat
 $\Delta \leftarrow 0$
 For each $s \in \mathcal{S}$:
 $v \leftarrow V(s)$
 $V(s) \leftarrow \sum_{s'} \mathcal{P}_{ss'}^{\pi(s)} [\mathcal{R}_{ss'}^{\pi(s)} + \gamma V(s')]$
 $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
 until $\Delta < \theta$ (a small positive number)
3. Policy Improvement
 $policy_stable \leftarrow true$
 For each $s \in \mathcal{S}$:
 $b \leftarrow \pi(s)$
 $\pi(s) \leftarrow \arg \max_a \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V(s')]$
 If $b \neq \pi(s)$, then $policy_stable \leftarrow false$
 If $policy_stable$, then stop; else go to 2



Unit 03 | 동적계획법

가치 반복 (Value Iteration)

Initialize V arbitrarily, e.g., $V(s) = 0$, for all $s \in \mathcal{S}^+$

Repeat

$\Delta \leftarrow 0$

For each $s \in \mathcal{S}$:

$v \leftarrow V(s)$

$V(s) \leftarrow \max_a \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

until $\Delta < \theta$ (a small positive number)

Output a deterministic policy, π , such that

$\pi(s) = \arg \max_a \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V(s')]$

Initial random $V_0(S)$

0.55	0.72	0.60
0.54	0.42	0.65
0.44	0.89	0.96

 $V_1(S)$: k = 1 delta = 1.221402

-0.36	-0.46	-0.42
-0.61	-0.20	1.87
-0.20	1.87	1.87

 $V_{153}(S)$: k = 153 delta = 0.000000

4.58	6.20	8.00
6.20	8.00	10.00
8.00	10.00	10.00

정책 추출

Optimal policy

→	→	↓
→	→	↓
→	→	↑

Unit 03 | 동적계획법

가치 반복 (Value Iteration)

Initialize V arbitrarily, e.g., $V(s) = 0$, for all $s \in \mathcal{S}^+$

Repeat

$\Delta \leftarrow 0$

For each $s \in \mathcal{S}$:

$v \leftarrow V(s)$

$V(s) \leftarrow \max_a \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

until $\Delta < \theta$ (a small positive number)

Output a deterministic policy, π , such that

$\pi(s) = \arg \max_a \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V(s')]$

Initial random $V_0(S)$

0.55	0.72	0.60
0.54	0.42	0.65
0.44	0.89	0.96

 $V_1(S) : k = 1 \quad \text{delta} = 1.221402$

-0.36	-0.46	-0.42
-0.61	-0.20	1.87
-0.20	1.87	1.87

 $V_{153}(S) : k = 153 \quad \text{delta} = 0.000000$

4.58	6.20	8.00
6.20	8.00	10.00
8.00	10.00	10.00

정책 추출

Optimal policy

→	→	↓
→	→	↓
→	→	↑

Up $Q_\pi(s, a1) = -3 + 0.9 \times 0.55 = -2.51$
 Right $Q_\pi(s, a2) = -1 + 0.9 \times 0.72 = -0.36$
 Down $Q_\pi(s, a3) = -1 + 0.9 \times 0.54 = -0.51$
 Left $Q_\pi(s, a4) = -3 + 0.9 \times 0.55 = -2.51$

contents

Unit 00 | Intro

Unit 01 | 강화학습 기본 요소

Unit 02 | MDP / Bellman Equation

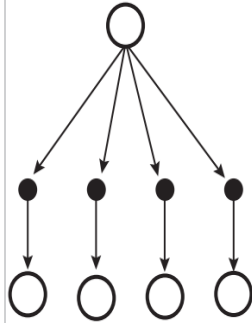

Unit 03 | 동적계획법

Unit 04 | Model-free algorithm(Q-learning)

Unit 04 | Q-learning

Model-free algorithm


- 환경에 대한 정보를 정확히 알지 못한 채,
직접 행동을 취해보고 보상을 받으며 학습하는 알고리즘
- 몬테카를로 방법(Monte Carlo Method, MC)
- 시간차학습 (Temporal difference, TD) : SARSA, Q-learning

	동적계획법	시간차 학습
환경 정보	필요(model-based)	불필요(model-free)
가치함수 계산	상태전이확률	샘플링
학습 단위	Time Step	Time Step
백업 다이어그램		

Unit 04 | Q-learning

Q-learning

$$Q(s, a) \leftarrow \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma \max_{a'} Q(s', a') \right]$$


$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha \left[R(s, a, s') + \gamma \max_{a'} Q(s', a') \right]$$

- 기존 값과 새로운 추정 값을 α 비율로 섞은 후에 업데이트 (지수이동평균)
- 더이상 상태전이확률이 필요없음!

Unit 04 | Q-learning

Q-learning

```

Initialize  $Q(s, a)$  arbitrarily
Repeat (for each episode):
  Initialize  $s$ 
  Repeat (for each step of episode):
    Choose  $a$  from  $s$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)
    Take action  $a$ , observe  $r, s'$ 
     $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$ 
     $s \leftarrow s'$ 
  until  $s$  is terminal

```

Q-learning : $Q(s, a)$

1.10	2.56	4.17
1.10 4.56	3.10 6.17	4.55 4.17
4.56	6.18	7.97
3.10	4.56	6.17
2.56 6.18	4.56 7.98	6.18 5.97
6.18	7.97	9.97
4.56	6.18	9.97
4.17 7.97	6.17 9.97	9.97 9.97
4.17	5.97	9.97

Q-learning : optimal policy

↓	↓	↓
→	→	↓
→	→	←

과제

논문 리뷰 : 'Playing Atari with Deep Reinforcement Learning'

강의안과 함께 첨부된 논문을 읽고 리뷰해주세요.

형식은 자유롭게 해주시면 됩니다. (주피터 노트북의 마크다운, 워드, 한글 등)

다른 리뷰나 분석글 등을 참고해주셔도 좋습니다.

단, '참고'만 해서 본인이 고민한 흔적을 남겨주시고 참고한 글 출처는 꼭 남겨주세요.

참고문헌

- 투빅스 12기 윤기오님 강의자료
- [기초부터 시작하는 강화학습/신경망 알고리즘] 책
- T academy [토크ON세미나] 강화학습 입문하기 유튜브 강의



Q & A

들어주셔서 감사합니다.