




Reinforcement Learning
: Q-learning


Intelligent Information System Lab, Hanyang University



INDEX


1. Introduction to Machine Learning
2. Classification of Machine Learning
3. What is Reinforcement learning?
4. Q-Learning
5. Application of Q-Learning

Intelligent Information System Lab, Hanyang University




1. Introduction to Machine Learning

Intelligent Information System Lab, Hanyang University




Introduction to Machine Learning

- **인공지능(Artificial Intelligence)**
사람이 하는 **지능적 행동**(behavior)을 컴퓨터(기계)가 수행하도록 하는 학문
- **기계학습(Machine Learning)**
인공지능의 한 방법론으로서 다량의 데이터를 컴퓨터가 학습하는 알고리즘과 기술을 개발하는 분야
데이터를 이용해서 모델을 만들어내는 것.
예측.



Intelligent Information System Lab, Hanyang University




Introduction to Machine Learning

벡터로 표현~

- 기계학습의 핵심은 표현(representation)과 일반화(generalization)
- **표현(representation)**
학습을 위해 주어진 데이터에 대한 표현
- **일반화(generalization)**
아직 주어지지 않은 데이터에 대한 처리

• • • • • • • • • •


Intelligent Information System Lab, Hanyang University



2. Classification of Machine Learning

• • • • • • • • • •

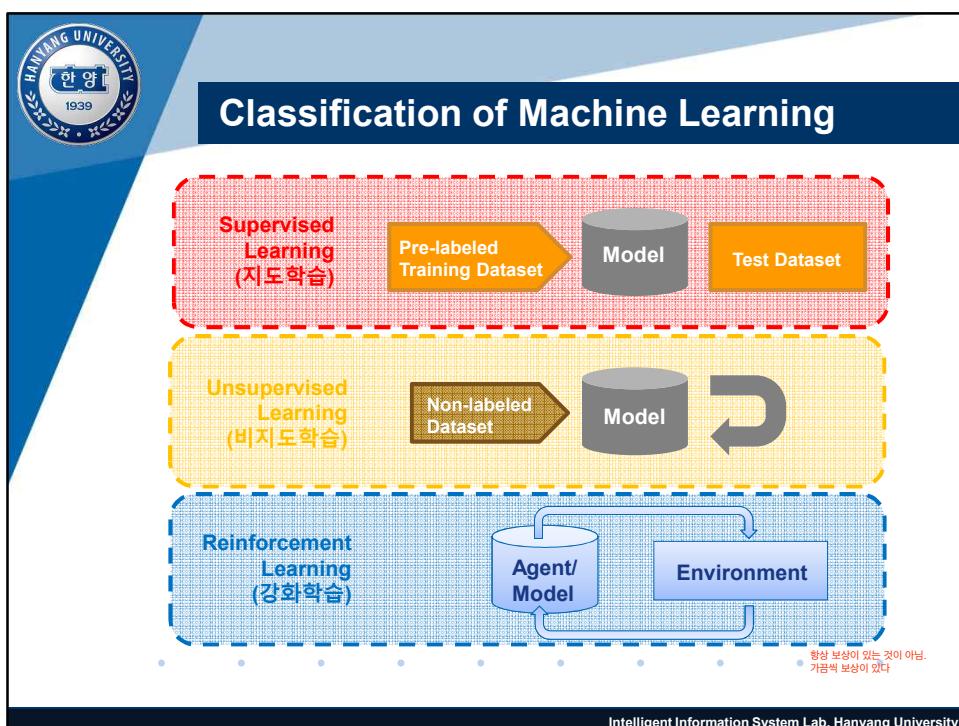
Intelligent Information System Lab, Hanyang University




Classification of Machine Learning

- 기계학습(Machine Learning)은 학습 방법에 따라 크게 아래와 같이 세 가지로 분류 할 수 있다
 - ✓ Supervised Learning (지도 학습) 일반적인 classification
 - 목표값(label)이 제시된 데이터가 학습데이터로 주어짐
 - ✓ Unsupervised Learning (비지도 학습)
 - 목표값(label)이 제시되지 않은 데이터가 학습데이터로 주어짐
 - ✓ Reinforcement Learning (강화 학습)
 - Unsupervised Learning의 일종으로, 에이전트가 자신이 수행한 행동에 대하여 보상값(reward)을 받아 점차적으로 효율적인 방식으로 행동을 강화시키는 학습 방법


Intelligent Information System Lab, Hanyang University





3. What is Reinforcement learning?


Intelligent Information System Lab, Hanyang University



What is Reinforcement learning?

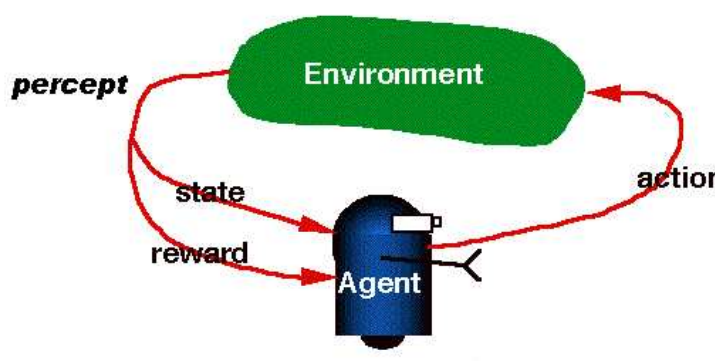
- Reinforcement learning
 - 반복적인 시도를 통해 시행착오를 겪으며 주어지는 외부 환경으로 부터의 Reward를 통해 Goal에 도달하는 방법으로 학습해 나가는 형태의 기계 학습
 - ✓ Learn via experiences!
- Reinforcement learning의 특징
 - ✓ Delayed reward – 시행착오 중에 선택한 Action이 적합한 선택인지 당장 알 수 없다
 - ✓ 현재 선택한 Action의 결과를 정확히 알 수 없어도 무관하다
 - ✓ Life-long learning에 활용 가능

Intelligent Information System Lab, Hanyang University


 HANYANG UNIVERSITY
한양
1939

What is Reinforcement learning?

- Reinforcement learning




Intelligent Information System Lab, Hanyang University

 HANYANG UNIVERSITY
한양
1939

4. Q-Learning

Intelligent Information System Lab, Hanyang University




Q-Learning

- Reinforcement Learning 가운데 가장 널리 사용되는 기계학습 알고리즘
- 현재 상태에서 선택 가능한 Action중에 임의의 Action을 선택하고 실행 한 뒤, 외부환경으로부터 Reward를 받음 (시행착오)
- 학습 시점에서는 Action에 대한 평가가 완료되지 않았으므로, 해당 시점에서 최적으로 평가 된 Action이라도, 실제로는 최적 Action이 아닐 수 있음
- Action에 대한 Reward를 전파하는 형태로 학습

.

Intelligent Information System Lab, Hanyang University




Q-Learning

- $Q(s, a)$ 는 estimated utility function으로서, State s 에서 Action a 를 선택하는 것이 유리한 정도를 나타냄
- $Q(s, a)$ 는 Action a 를 선택함으로써 얻을 수 있는 즉각적인 reward와 Action a 로 인해 변화된 State s' 에서 얻을 수 있는 잠재적 reward의 최대값의 합으로 정의
- $Q(s, a)$ 의 학습이 완료되면 각 Step마다 현재 State s 에 대하여 평가함수 $Q(s, a)$ 를 최대화하는 Action a 를 선택하는 방식으로 이동

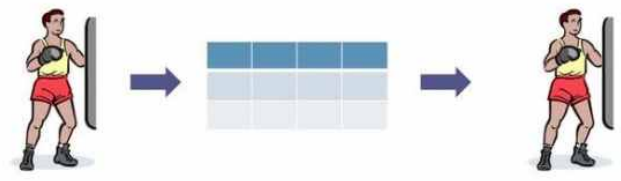
.

Intelligent Information System Lab, Hanyang University



Q-Learning

➔ 강화학습 (Q-러닝 알고리즘)




특정한 상태(state)에 놓여있을 때
취할 수 있는 각각의 행동(action)에
대한 보상 값을
미리 계산 및 학습

학습 한 것을
테이블에 저장

비슷한 상태가 나왔을 때
최적의 행동을 테이블에서 검색

Intelligent Information System Lab, Hanyang University



Q-Learning

• The algorithm

=> 현재 상황에서 어떤 액션을 취했을때 리워드의 값은
즉시 들어오는 리워드 값 + 그 후에 들어올 수 있는 최고의 리워드 값(여기는 100%다 들어오진 않는다)

Immediate reward

$r(s, a)$

상태 s에서 액션 a를 했을때
보상 값

Delayed reward

$\gamma \max_{a'} Q(s', a')$

a' : 그 다음에 취할 수 있는 액션
s'는 s에서 a를 취했을 때, 일이 성공하고나서
현재 상황을 다시 체크. (1로 이동하자! : s,a => 1로 이동했을 때 : s)

$$Q(s, a) = r(s, a) + \gamma \max_{a'} Q(s', a')$$

바로 들어오는 reward라면 수
reward 없다면 0이 된다

계속 Q계산하면 어떤 값으로 수렴할 것
이 때 성능이 얼마나 좋으냐가 Q

- $r(s, a)$ = Action a를 선택했을 시 주어지는 즉각적인 reward (Immediate reward)
- γ = reward가 전파되는 정도를 나타내는 계수 [0,1]
- s' = Action a를 선택했을 시 new state 예측
- a, a' = 각각 state s, s'에서 취할 수 있는 Action


Q(s,a) 학습완료 시, Optimal move를 위한 action 선택:

$$\pi(s) = \arg \max_a Q(s, a)$$

수렴된 Q값을 최고로 하는 것 : 파이

수렴된 Q

Intelligent Information System Lab, Hanyang University



Q-Learning


- The algorithm
 - 각 state-action pair(s,a)에 대하여, Q(s,a)값을 0으로 초기화 시킴
 - 1~5 과정을 Q(s,a)가 수렴할 때까지 반복

1. 현재 State s에서 선택 가능한 임의의 Action a를 선택하고 실행
2. 외부환경으로 부터 immediate reward를 받음
3. 새로운 State s'을 감지
4. Q(s,a)값을 아래와 같은 수식으로 갱신

$$Q(s,a) = r + \gamma \max_a Q(s',a')$$

수식: reward에 따른 Q(s,a) 갱신
5. s = s'

Intelligent Information System Lab, Hanyang University



Q-Learning (Example)

- Example Problem

화살표 : 액션
다발 : s


Q-러닝을 적용시키기에 좋은 예제는 아닌
결과적으로 S6으로 가는 문제

각 s마다 최대 3방향으로 갈 수 있고 액션은 두가지만

S1	$\xrightarrow{a12}$ $\xleftarrow{a21}$	S2	$\xrightarrow{a23}$ $\xleftarrow{a32}$	S3
$\xleftarrow{a41}$	S4	$\xleftarrow{a52}$	S5	$\xleftarrow{a63}$
$\xrightarrow{a14}$	$\xrightarrow{a45}$	$\xrightarrow{a56}$	S6	END
$\xrightarrow{a41}$	$\xleftarrow{a54}$	$\xleftarrow{a65}$		

- 위 와 같은 형태의 공간 상 임의의 위치에서 시작해서 S6으로 이동하도록 학습시키는 문제

Intelligent Information System Lab, Hanyang University




Q-Learning (Example)

- Example Problem

s,a	Q(s,a)
S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	0
S2,a25	0
S3,a32	0
S3,a36	0
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

- Q(s,a)는 State s에서 선택 가능한 Action a를 선택 했을 때 유리한 정도를 나타내는 Function. 학습 전 초기값은 모두 0으로 초기화

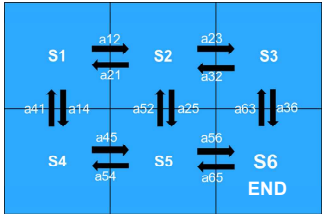
Intelligent Information System Lab, Hanyang University



Q-Learning (Example)


- Initial State

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	0
S2,a25	0
S3,a32	0
S3,a36	0
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0



• • • • • • • • •

Intelligent Information System Lab, Hanyang University



Q-Learning (Example)

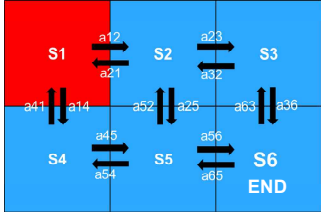
- The algorithm (Iteration 1)

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	0
S2,a25	0
S3,a32	0
S3,a36	0
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0


Current Position: Red

Available actions: a12, a14

→ Chose a12 (Move to S2)



Intelligent Information System Lab, Hanyang University



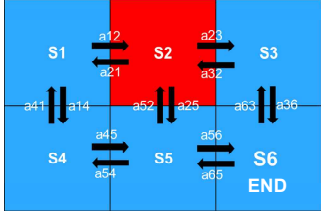
Q-Learning (Example)

- Update Q(S1, a12)

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	0
S2,a25	0
S3,a32	0
S3,a36	0
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

Current Position: Red

Available actions: a21, a25, a23




Update Q(S1, a12) :

- $$Q(S1, a12) = r + 0.5 * \max(Q(S2, a21), Q(S2, a25), Q(S2, a23)) = 0$$

여기서 0이 나오기 때문이다
 그러므로 r = 0

Intelligent Information System Lab, Hanyang University



Q-Learning (Example)

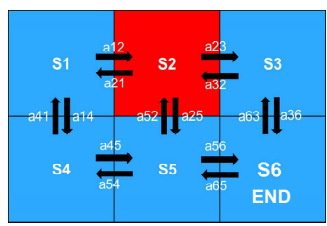
- Next Move

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	0
S2,a25	0
S3,a32	0
S3,a36	0
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0


Current Position: Red

Available actions: a21, a25, a23

→ Chose a23 (Move to S3)



Intelligent Information System Lab, Hanyang University



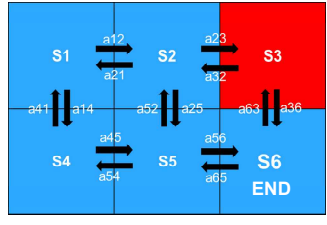
Q-Learning (Example)

- Update Q(S2, a23)

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	0
S2,a25	0
S3,a32	0
S3,a36	0
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

Current Position: Red

Available actions: a32, a36




Update Q(S2, a23):

$$Q(S2, a23) = r + 0.5 * \max(Q(S3, a32), Q(S3, a36)) = 0$$

Q(S3,a36)의 값은 현재 0

Intelligent Information System Lab, Hanyang University



Q-Learning (Example)

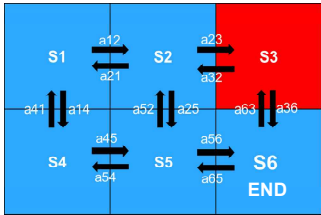
- Next Move

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	0
S2,a25	0
S3,a32	0
S3,a36	0
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

Current Position: Red


Available actions: a32, a36

→ Chose a36 (Move to S6)



• • • • • • • • •

Intelligent Information System Lab, Hanyang University



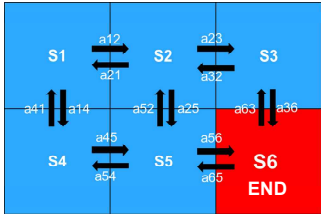
Q-Learning (Example)

- Update Q(S3,a36)

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	0
S2,a25	0
S3,a32	0
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

Current Position: Red


FINAL STATE!



• • • • • • • • •

Update Q(S3,a36) : $Q(S3,a36) = r = 100$

Intelligent Information System Lab, Hanyang University



Q-Learning (Example)

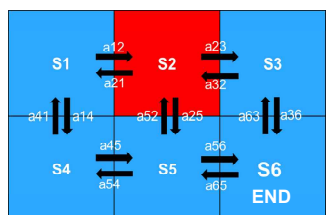
- **New Game (Iteration 2)**

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	0
S2,a25	0
S3,a32	0
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0


Current Position: Red

Available actions: a21, a25, a23

→ Chose a23 (Move to 3)



Intelligent Information System Lab, Hanyang University



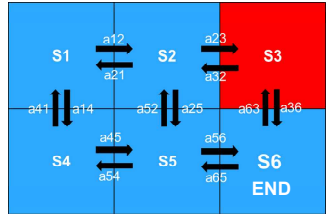
Q-Learning (Example)

- **Update Q(S2, a23)**

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	0
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

Current Position: Red


Available actions: a32, a36



Update Q(S2, a23) :

$$Q(S2, a23) = r + 0.5 * \max(Q(S3, a32), Q(S3, a36)) = 0 + 0.5 * 100 = 50$$

Intelligent Information System Lab, Hanyang University



Q-Learning (Example)

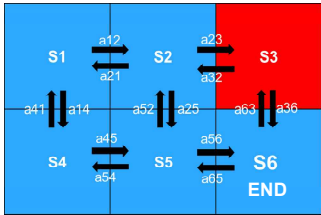
- Next Move

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	0
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

Current Position: Red


Available actions: a32, a36

→ Chose a36 (Move to S6)



• • • • • • • • •

Intelligent Information System Lab, Hanyang University



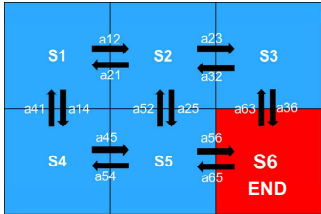
Q-Learning (Example)

- Update Q(S3,a36)

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	0
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

Current Position: Red


FINAL STATE!



• • • • • • • • •

Update Q(S3,a36) : $Q(S3,a36) = r = 100$

Intelligent Information System Lab, Hanyang University



Q-Learning (Example)

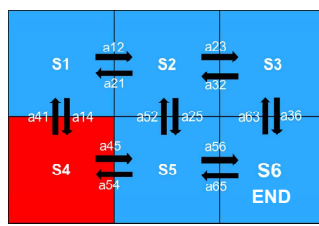
- **New Game (Iteration 3)**

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	0
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0


Current Position: Red

Available actions: a41, a45

→ Chose a41 (Move to 1)



Intelligent Information System Lab, Hanyang University



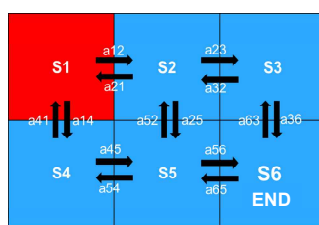
Q-Learning (Example)

- **Update Q(S4,a41)**

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	0
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

Current Position: Red


Available actions: a12, a14



Update Q(S4,a41) :

$Q(S4, a41) = r + 0.5 * \max(Q(S1,a12), Q(S1,a14)) = 0$

Intelligent Information System Lab, Hanyang University



Q-Learning (Example)

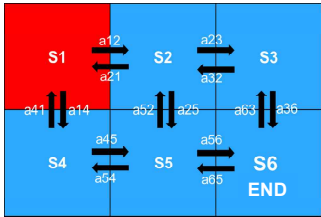
- Next Move

S1,a12	0
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	0
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0


Current Position: Red

Available actions: a12, a14

→ Chose a12 (Move to 2)



Intelligent Information System Lab, Hanyang University



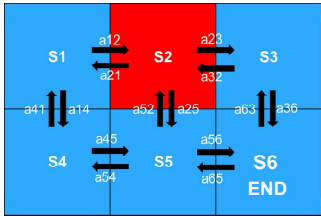
Q-Learning (Example)

- Update Q(S1,a12)

S1,a12	25
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	0
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

Current Position: Red


Available actions: a21, a25, a23



Update Q(S1,a12) :

$$Q(S1, a12) = r + 0.5 * \max(Q(S2,a21), Q(S2,a25), Q(S2,a23)) = 0 + 0.5 * 50 = 25$$

Intelligent Information System Lab, Hanyang University



Q-Learning (Example)

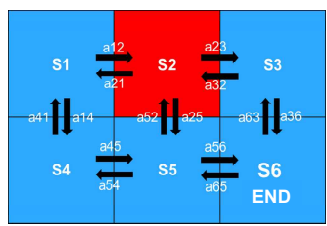
- Next Move

S1,a12	25
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	0
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0


Current Position: Red

Available actions: a21, a25, a23

→ Chose a23 (Move to 3)



Intelligent Information System Lab, Hanyang University



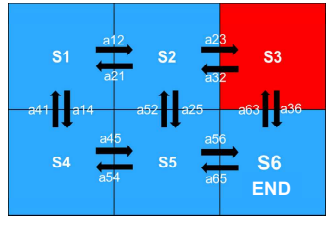
Q-Learning (Example)

- Update Q(S2,a23)

S1,a12	25
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	0
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

Current Position: Red


Available actions: a32, a36



Update Q(S2,a23) :

$Q(S2, a23) = r + 0.5 * \max(Q(S3, a32), Q(S3, a36)) = 0 + 0.5 * 100 = 50$

Intelligent Information System Lab, Hanyang University



Q-Learning (Example)

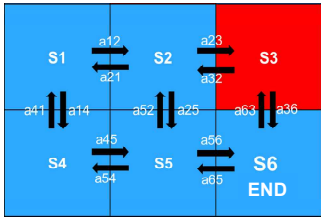
- Next Move

S1,a12	25
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	0
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0


Current Position: Red

Available actions: a32, a36

→ Chose a32 (Move to 2)



Intelligent Information System Lab, Hanyang University



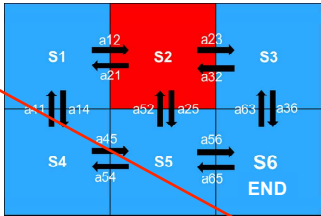
Q-Learning (Example)

- Update Q(S3,a32)

S1,a12	25
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	25
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

Current Position: Red


Available actions: a21, a25, a23



Update Q(S3,a32) :

$$Q(S3, a32) = r + 0.5 * \max(Q(S2,a21), Q(S2,a25), Q(S2,a23)) = 0 + 0.5 * 50 = 25$$

Intelligent Information System Lab, Hanyang University



Q-Learning (Example)

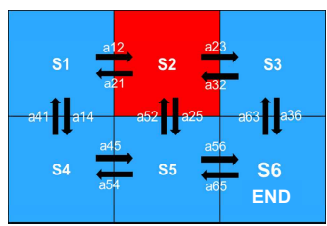
- Next Move

S1,a12	25
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	25
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0


Current Position: Red

Available actions: a21, a25, a23

→ Chose a25 (Move to 5)



Intelligent Information System Lab, Hanyang University



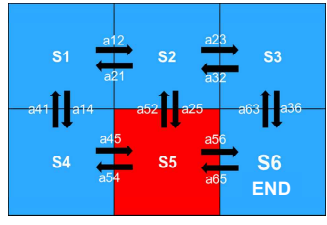
Q-Learning (Example)

- Update Q(S2,a25)

S1,a12	25
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	25
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0

Current Position: Red


Available actions: a54, a52, a56



Update Q(S2,a25) :

$Q(S2, a25) = r + 0.5 * \max(Q(S5,a54), Q(S5,a52), Q(S5,a56)) = 0$

Intelligent Information System Lab, Hanyang University



Q-Learning (Example)

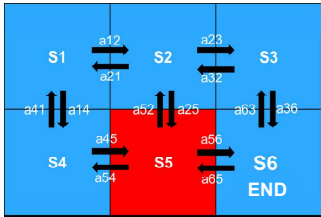
- Next Move

S1,a12	25
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	25
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	0


Current Position: Red

Available actions: a54, a52, a56

→ Chose a56 (Move to 6)



Intelligent Information System Lab, Hanyang University



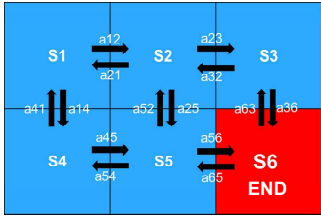
Q-Learning (Example)

- Update Q(S5,a56)

S1,a12	25
S1,a14	0
S2,a21	0
S2,a23	50
S2,a25	0
S3,a32	25
S3,a36	100
S4,a41	0
S4,a45	0
S5,a54	0
S5,a52	0
S5,a56	100


Current Position: Red

FINAL STATE!



Update Q(S5,a56) : $Q(S5,a56) = r = 100$

Intelligent Information System Lab, Hanyang University

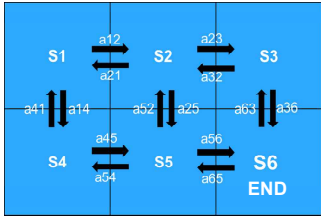


Q-Learning (Example)


- Final State (after many iterations)

S1,a12	25
S1,a14	25
S2,a21	12.5
S2,a23	50
S2,a25	50
S3,a32	25
S3,a36	100
S4,a41	12.5
S4,a45	50
S5,a54	25
S5,a52	25
S5,a56	100

← 학습이 완료된 $Q(s,a)$ function




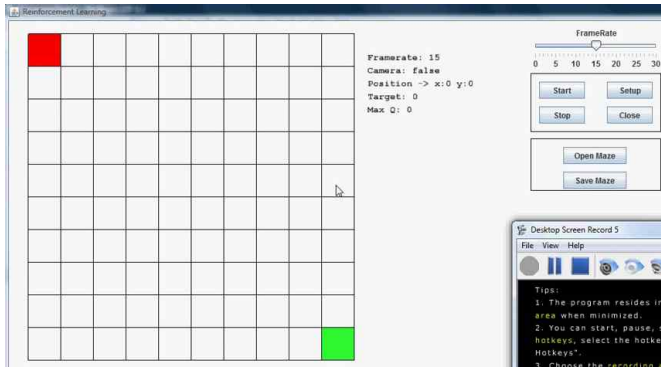
Intelligent Information System Lab, Hanyang University



5. Application of Q-Learning


Intelligent Information System Lab, Hanyang University


 **Application of Q-Learning**



<https://www.youtube.com/watch?v=tovrpoUkzYU>

Intelligent Information System Lab, Hanyang University

 **Application of Q-Learning**



<https://www.youtube.com/watch?v=zOgSC---rgM>

Intelligent Information System Lab, Hanyang University

