로봇인지및지능제어 Midterm

1970030 박 근 아

1. 예제 9-5 에서 가치함수 식은 다음과 같다.

$$v_\pi(s_t) = v_\pi(s_t) + \rho((r_{t+1} + \gamma v_\pi(s_{t+1})) - v_\pi(s_t))$$

그리고 학습률 $\rho = 0.1$, 할인율 $\gamma = 1.0$, 1칸마 16칸을 제외하고는 -1의 보상을 받는다.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

(a) 예제 격자 보드

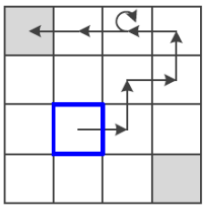
초기 가치함수

	0	0	0
0	0	0	0
0	0	0	0
0	0	0	

(b) 가치함수 초기화

1차 에피소드에 대한 상태가치 함수를 구해보면 다음과 같다.

1차 에피소드



$$v_\pi(10) = \underset{0}{v_\pi(10)} + 0.1((-1 + 1 \cdot \underset{0}{v_\pi(11)}) - \underset{0}{v_{tt}(10)}) = -0.1$$

$$v_\pi(11) = \underset{0}{v_\pi(11)} + 0.1((-1 + 1 \cdot \underset{0}{v_\pi(7)}) - \underset{0}{v_{tt}(11)}) = -0.1$$

$$v_\pi(7) = \underset{0}{v_\pi(7)} + 0.1((-1 + 1 \cdot \underset{0}{v_\pi(8)}) - \underset{0}{v_{tt}(7)}) = -0.1$$

$$v_\pi(8) = \underset{0}{v_\pi(8)} + 0.1((-1 + 1 \cdot \underset{0}{v_\pi(4)}) - \underset{0}{v_{tt}(8)}) = -0.1$$

$$v_\pi(4) = \underset{0}{v_\pi(4)} + 0.1((-1 + 1 \cdot \underset{0}{v_\pi(3)}) - \underset{0}{v_{tt}(4)}) = -0.1$$

$$v_\pi(3) = \underset{0}{v_\pi(3)} + 0.1((-1 + 1 \cdot \underset{0}{v_\pi(3)}) - \underset{0}{v_{tt}(3)}) = -0.1$$

$$v_\pi(3) = \underset{-0.1}{v_\pi(3)} + 0.1((-1 + 1 \cdot \underset{0}{v_\pi(2)}) - \underset{-0.1}{v_{tt}(3)}) = -0.19$$

$$v_\pi(2) = \underset{0}{v_\pi(2)} + 0.1((5 + 1 \cdot \underset{0}{v_\pi(1)}) - \underset{0}{v_{tt}(2)}) = 0.5$$

결과를 격자에 표시하면 다음과 같다.

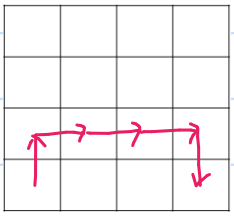
	0.5	-0.19	-0.1
0	0	-0.1	-0.1
0	-0.1	-0.1	0
0	0	0	

이어서 임의로 생성한 2nd 에피소드에 대한 상태가치 함수를 구해보면 다음과 같다.

가치함수

	0.5	-0.19	-0.1
0	0	-0.1	-0.1
0	-0.1	-0.1	0
0	0	0	

2nd 에피소드



$v_\pi(13) = \underset{0}{v_\pi(13)} + 0.1\left(\left(-1 + 1 \cdot \underset{0}{v_\pi(9)}\right) - \underset{0}{v_{tt}(13)}\right) = -0.1$

$v_\pi(9) = \underset{0}{v_\pi(9)} + 0.1\left(\left(-1 + 1 \cdot \underset{-0.1}{v_\pi(10)}\right) - \underset{0}{v_{tt}(9)}\right) = -0.11$

$v_\pi(10) = \underset{-0.1}{v_\pi(10)} + 0.1\left(\left(-1 + 1 \cdot \underset{-0.1}{v_\pi(11)}\right) - \underset{-0.1}{v_{tt}(10)}\right)$

$= -0.1 + 0.1(-1.1 + 0.1) = -0.2$

$v_\pi(11) = \underset{-0.1}{v_\pi(11)} + 0.1\left(\left(-1 + 1 \cdot \underset{0}{v_\pi(12)}\right) - \underset{-0.1}{v_{tt}(11)}\right)$

$= -0.1 + 0.1(-1 + 0.1) = -0.19$

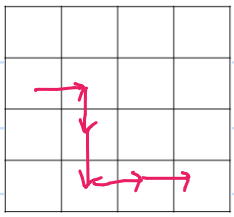
$v_\pi(12) = \underset{0}{v_\pi(12)} + 0.1\left(\left(5 + 1 \cdot \underset{0}{v_\pi(16)}\right) - \underset{0}{v_{tt}(12)}\right) = 0.5$

가치함수

	0.5	-0.19	-0.1
0	0	-0.1	-0.1
-0.11	-0.2	-0.19	0.5
-0.1	0	0	

업데이트 해준 가치함수는 왼쪽격자와 같다.

마지막으로 3rd 에피소드를 생성하고 상태가치함수를 구하면 다음과 같다.



$v_\pi(5) = \underset{0}{v_\pi(5)} + 0.1\left(\left(-1 + 1 \cdot \underset{0}{v_\pi(6)}\right) - \underset{0}{v_{tt}(5)}\right) = -0.1$

$v_\pi(6) = \underset{0}{v_\pi(6)} + 0.1\left(\left(-1 + 1 \cdot \underset{-0.2}{v_\pi(10)}\right) - \underset{0}{v_{tt}(6)}\right) = 0.1 \times (-1.2) = -0.12$

$v_\pi(10) = \underset{-0.2}{v_\pi(10)} + 0.1\left(\left(-1 + 1 \cdot \underset{0}{v_\pi(14)}\right) - \underset{-0.2}{v_{tt}(10)}\right) = -0.2 + 0.1(-1 + 0.2)$

$= -0.28$

$v_\pi(14) = \underset{0}{v_\pi(14)} + 0.1\left(\left(-1 + 1 \cdot \underset{0}{v_\pi(15)}\right) - \underset{0}{v_{tt}(14)}\right) = -0.1$

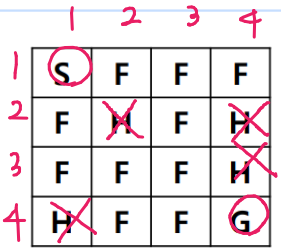
$v_\pi(15) = \underset{0}{v_\pi(15)} + 0.1\left(\left(5 + 1 \cdot \underset{0}{v_\pi(16)}\right) - \underset{0}{v_{tt}(15)}\right) = 0.5$

가치함수

	0.5	-0.19	-0.1
-0.1	-0.12	-0.1	-0.1
-0.11	-0.28	-0.19	0.5
-0.1	-0.1	0.5	

업데이트 해준 가치함수는 왼쪽격자와 같다.

2.



문제에 대해 정한 기준들은 다음과 같다.

- Down, Right, Up, Left의 순서로 Q테이블을 작성해준다.
- 환경의 바깥으로 나가는 행동은 포함하지 않는다.
- Hole로 이동하는 경우 가치를 0.1, Goal에 도착하는 경우 가치를 0.8, 나머지는 모두 0.5로 부여한다.
- 입실론-그리디 정책에서 액션을 선택할 때 동일한 최대값이 여러 개인 경우, Q테이블에서 가장 먼저 나오는 액션을 취한다.
- 멈춤조건은 update 될 수 있는 행동들의 가치가 모두 바뀌고 더 이상 가치함수의 갱신으로 인해 선택하는 행동이 바뀌지 않는 경우로 한다.

앞서 정한 기준으로 Q테이블 값을 채워보면 다음과 같이 정리된다.

상태	행동	가치
(1,1)	Down	0.5
(1,1)	Right	0.5
(1,2)	Down	0.1
(1,2)	Right	0.5
(1,2)	Left	0.5
(1,3)	Down	0.5
(1,3)	Right	0.5
(1,3)	Left	0.5
(1,4)	Down	0.1
(1,4)	Left	0.5

상태	행동	가치
(2,1)	Down	0.5
(2,1)	Right	0.1
(2,1)	Up	0.5
(2,3)	Down	0.5
(2,3)	Right	0.1
(2,3)	Up	0.5
(2,3)	Left	0.1

상태	행동	가치
(3,1)	Down	0.1
(3,1)	Right	0.5
(3,1)	Up	0.5
(3,2)	Down	0.5
(3,2)	Right	0.5
(3,2)	Up	0.1
(3,2)	Left	0.5
(3,3)	Down	0.5
(3,3)	Right	0.1
(3,3)	Up	0.5
(3,3)	Left	0.5

상태	행동	가치
(4,2)	Right	0.5
(4,2)	Left	0.1
(4,2)	Up	0.5
(4,3)	Right	0.8
(4,3)	Left	0.5
(4,3)	Up	0.5
(4,4)	Left	0.5

1st loop 부터 실행해보면 시작위치를 (1,1)로 정한다.

상태 (1,1)에서는 Right와 Down 행동 모두 같은 가치를 가져 먼저나온 Down 행동을 취해준다.

$$q_f[(1,1), \text{Down}] = \underbrace{q_f[(1,1), \text{Down}]}_{0.5} + 0.1(0.3 + 1 \cdot \underbrace{\max[(2,1), \text{action}]}_{0.5,\ \text{Down}} - \underbrace{q_f[(1,1), \text{Down}]}_{0.5})$$

$$= 0.5 + 0.1(0.3 + 0.5 - 0.5) = 0.53$$

$$q_f[(2,1), \text{Down}] = \underbrace{q_f[(2,1), \text{Down}]}_{0.5} + 0.1(0.3 + 1 \cdot \underbrace{\max[(3,1), \text{action}]}_{0.5,\ \text{Right}} - \underbrace{q_f[(2,1), \text{Down}]}_{0.5})$$

$$= 0.5 + 0.1(0.3 + 0.5 - 0.5) = 0.53$$

$$q_f[(3,1), \text{Right}] = \underbrace{q_f[(3,1), \text{Right}]}_{0.5} + 0.1(0.3 + 1 \cdot \underbrace{\max[(3,2), \text{action}]}_{0.5,\ \text{Down}} - \underbrace{q_f[(3,1), \text{Right}]}_{0.5})$$

$$= 0.5 + 0.1(0.3 + 0.5 - 0.5) = 0.53$$

$$q_f[(3,2), \text{Down}] = \underbrace{q_f[(3,2), \text{Down}]}_{0.5} + 0.1(0.3 + 1 \cdot \underbrace{\max[(4,2), \text{action}]}_{0.5,\ \text{Right}} - \underbrace{q_f[(3,2), \text{Down}]}_{0.5})$$

$$= 0.5 + 0.1(0.3 + 0.5 - 0.5) = 0.53$$

$q[(4,2), \text{Right}] = q[(4,2), \text{Right}] + 0.1(0.3 + 1 \cdot \max[(4,3), \text{action}] - q[(4,2), \text{Right}])$

$q[(4,2), \text{Right}]$: 0.5 / $\max[(4,3), \text{action}]$: 0.8, Right / $q[(4,2), \text{Right}]$: 0.5

$= 0.5 + 0.1(0.3 + 0.8 - 0.5) = 0.56$

$q[(4,3), \text{Right}] = q[(4,3), \text{Right}] + 0.1(0.3 + 1 \cdot \max[(4,4), \text{action}] - q[(4,3), \text{Right}]) = 0.8$

$q[(4,3), \text{Right}]$: 0.8 / $\max[(4,4), \text{action}]$: 0.5 / $q[(4,3), \text{Right}]$: 0.8

(4,4)는 목표상태이기 때문에 계산을 멈춘다.

계산 결과를 바탕으로 Q-table을 업데이트 해주면 다음과 같다.

상태	행동	가치
(1,1)	Down	0.53
(1,1)	Right	0.5
(1,2)	Down	0.1
(1,2)	Right	0.5
(1,2)	Left	0.5
(1,3)	Down	0.5
(1,3)	Right	0.5
(1,3)	Left	0.5
(1,4)	Down	0.1
(1,4)	Left	0.5

상태	행동	가치
(2,1)	Down	0.53
(2,1)	Right	0.1
(2,1)	Up	0.5
(2,3)	Down	0.5
(2,3)	Right	0.1
(2,3)	Up	0.5
(2,3)	Left	0.1

상태	행동	가치
(3,1)	Down	0.1
(3,1)	Right	0.53
(3,1)	Up	0.5
(3,2)	Down	0.53
(3,2)	Right	0.5
(3,2)	Up	0.1
(3,2)	Left	0.5
(3,3)	Down	0.5
(3,3)	Right	0.1
(3,3)	Up	0.5
(3,3)	Left	0.5

상태	행동	가치
(4,2)	Right	0.56
(4,2)	Left	0.1
(4,2)	Up	0.5
(4,3)	Right	0.8
(4,3)	Left	0.5
(4,3)	Up	0.5

(4,4) Left 0.5

2nd loop를 실행해보면 이번에 시작위치를 랜덤하게 (1,2)로 정해본다.

$q[(1,2), \text{Right}] = q[(1,2), \text{Right}] + 0.1(0.3 + 1 \cdot \max[(1,3), \text{action}] - q[(1,2), \text{Right}])$

$q[(1,2), \text{Right}]$: 0.5 / $\max[(1,3), \text{action}]$: 0.5, Down / $q[(1,2), \text{Right}]$: 0.5

$= 0.5 + 0.1(0.3 + 0.5 - 0.5) = 0.53$

$q[(1,3), \text{Down}] = q[(1,3), \text{Down}] + 0.1(0.3 + 1 \cdot \max[(2,3), \text{action}] - q[(1,3), \text{Down}])$

$q[(1,3), \text{Down}]$: 0.5 / $\max[(2,3), \text{action}]$: 0.5, Down / $q[(1,3), \text{Down}]$: 0.5

$= 0.5 + 0.1(0.3 + 0.5 - 0.5) = 0.53$

$q[(2,3), \text{Down}] = q[(2,3), \text{Down}] + 0.1(0.3 + 1 \cdot \max[(3,3), \text{action}] - q[(2,3), \text{Down}])$

$q[(2,3), \text{Down}]$: 0.5 / $\max[(3,3), \text{action}]$: 0.5, Down / $q[(2,3), \text{Down}]$: 0.5

$= 0.5 + 0.1(0.3 + 0.5 - 0.5) = 0.53$

$q[(3,3), \text{Down}] = q[(3,3), \text{Down}] + 0.1(0.3 + 1 \cdot \max[(4,3), \text{action}] - q[(3,3), \text{Down}])$

$q[(3,3), \text{Down}]$: 0.5 / $\max[(4,3), \text{action}]$: 0.8, Right / $q[(3,3), \text{Down}]$: 0.5

$= 0.5 + 0.1(0.3 + 0.8 - 0.5) = 0.56$

$q[(4,3), \text{Right}] = \underbrace{q[(4,3), \text{Right}]}_{0.8} + 0.1(0.3 + 1 \cdot \underbrace{\max[(4,4), \text{action}]}_{0.5} - \underbrace{q[(4,3), \text{Right}]}_{0.8}) = 0.8$

(4,4)는 목표상태이기 때문에 계산을 멈춘다.

계산 결과를 바탕으로 Q-table을 업데이트 해주면 다음과 같다.

상태	행동	가치
(1,1)	Down	0.53
(1,1)	Right	0.5
(1,2)	Down	0.1
(1,2)	Right	0.53
(1,2)	Left	0.5
(1,3)	Down	0.53
(1,3)	Right	0.5
(1,3)	Left	0.5
(1,4)	Down	0.1
(1,4)	Left	0.5

상태	행동	가치
(2,1)	Down	0.53
(2,1)	Right	0.1
(2,1)	Up	0.5
(2,3)	Down	0.53
(2,3)	Right	0.1
(2,3)	Up	0.5
(2,3)	Left	0.1

상태	행동	가치
(3,1)	Down	0.1
(3,1)	Right	0.53
(3,1)	Up	0.5
(3,2)	Down	0.53
(3,2)	Right	0.5
(3,2)	Up	0.1
(3,2)	Left	0.5
(3,3)	Down	0.56
(3,3)	Right	0.1
(3,3)	Up	0.5
(3,3)	Left	0.5

상태	행동	가치
(4,2)	Right	0.56
(4,2)	Left	0.1
(4,2)	Up	0.5
(4,3)	Right	0.8
(4,3)	Left	0.5
(4,3)	Up	0.5

(4,4) Left 0.5

3rd loop를 실행해보면 이번에 시작위치를 랜덤하게 (1,4)로 정해본다.

$q[(1,4), \text{Left}] = \underbrace{q[(1,4), \text{Left}]}_{0.5} + 0.1(0.3 + 1 \cdot \underbrace{\max[(1,3), \text{action}]}_{0.53\ \text{Down}} - \underbrace{q[(1,4), \text{Left}]}_{0.5})$

$= 0.5 + 0.1(0.3 + 0.03) = 0.533$

$q[(1,3), \text{Down}] = \underbrace{q[(1,3), \text{Down}]}_{0.53} + 0.1(0.3 + 1 \cdot \underbrace{\max[(2,3), \text{action}]}_{0.53\ \text{Down}} - \underbrace{q[(1,3), \text{Down}]}_{0.53})$

$= 0.53 + 0.1(0.3 + 0.53 - 0.53) = 0.56$

$q[(2,3), \text{Down}] = \underbrace{q[(2,3), \text{Down}]}_{0.53} + 0.1(0.3 + 1 \cdot \underbrace{\max[(3,3), \text{action}]}_{0.56\ \text{Down}} - \underbrace{q[(2,3), \text{Down}]}_{0.53})$

$= 0.53 + \underbrace{0.1(0.3 + 0.03)}_{0.033} = 0.563$

$q[(3,3), \text{Down}] = \underbrace{q[(3,3), \text{Down}]}_{0.56} + 0.1(0.3 + 1 \cdot \underbrace{\max[(4,3), \text{action}]}_{0.8,\ \text{Right}} - \underbrace{q[(3,3), \text{Down}]}_{0.56})$

$= 0.56 + 0.1(1.1 - 0.56) = 0.614$

$q[(4,3), \text{Right}] = \underbrace{q[(4,3), \text{Right}]}_{0.8} + 0.1(0.3 + 1 \cdot \underbrace{\max[(4,4), \text{action}]}_{0.5} - \underbrace{q[(4,3), \text{Right}]}_{0.8}) = 0.8$

(4,4)는 목표상태에기 때문에 계산을 멈춘다.

계산 결과를 바탕으로 Q-table을 업데이트 해주면 다음과 같다.

상태	행동	가치
(1,1)	Down	0.53
(1,1)	Right	0.5
(1,2)	Down	0.1
(1,2)	Right	0.53
(1,2)	Left	0.5
(1,3)	Down	0.56
(1,3)	Right	0.5
(1,3)	Left	0.5
(1,4)	Down	0.1
(1,4)	Left	0.533

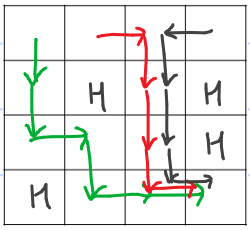
상태	행동	가치
(2,1)	Down	0.53
(2,1)	Right	0.1
(2,1)	Up	0.5
(2,3)	Down	0.563
(2,3)	Right	0.1
(2,3)	Up	0.5
(2,3)	Left	0.1

상태	행동	가치
(3,1)	Down	0.1
(3,1)	Right	0.53
(3,1)	Up	0.5
(3,2)	Down	0.53
(3,2)	Right	0.5
(3,2)	Up	0.1
(3,2)	Left	0.5
(3,3)	Down	~~0.5~~ 0.614
(3,3)	Right	0.1
(3,3)	Up	0.5
(3,3)	Left	0.5

상태	행동	가치
(4,2)	Right	0.56
(4,2)	Left	0.1
(4,2)	Up	0.5
(4,3)	Right	0.8
(4,3)	Left	0.5
(4,3)	Up	0.5

(4,4) Left 0.5

지금까지 loop를 통해 나온 경로는 다음과 같다.



- 1st loop
- 2nd loop
- 3rd loop

처음 결과를 보면 어떤 시작위치를 랜덤하게 고르더라도 1, 2, 3 번째 loop의 경로를 따라가게 된다. 그리고 그 경로를 거의 따라가게 되는 3번째 loop의 Q-table update을 확인해보면 미세하게 가치들이 update 되어 그 값이 커지는 것을 알수있다. 따라서 loop를 반복하면 상황들의 가치는 update를 반복하며 커지지만 그로 인해서 랜덤한 시작위치에서 행동의 선택에는 더 이상 영향이 없을 것을 확인할 수 있다. 따라서 3번째 loop에서 멈춤 조건을 만족했다고 보고 이때의 상태-행동 가치함수를 저장해준다.