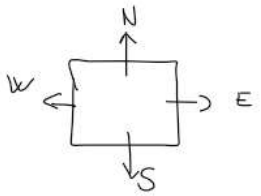
Quiz_9

mandag 11. april 2022 17.33

3	(1,3)	-90	80
2			
1	-90	-90	(3,1) 60
	1	2	3



$\gamma = 1$

iscord

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5
(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0
(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0
(2,2), S, (2,1), -90	(2,2), S, (2,1), -90	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0	(2,2), N, (2,3), -90
		(3,2), S, (3,1), 60	(3,2), S, (3,1), 60	

state, action, next state, reward

Q-learning update:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(\text{trial}_{t,t+1} - Q(s_t, a_t))$$

$$\text{trial}_{t,t+1} = r_{t+1} + \gamma \max_a Q(s_{t+1}, a)$$

and γ is discount factor and α is learning rate. For following values Q and upper episodes find out *first episode* and *iteration* (t), when value Q will be non zero. Write it in form E:2, t:3 - in the 2nd episode and 3rd iteration.

$Q((2,2), \text{E}) =$ _____
$Q((3,2), \text{S}) =$ _____
$Q((1,2), \text{E}) =$ _____

States: {(1,2), (1,3), (2,1), (2,2), (3,1), (3,2), (2,3)}
Terminal states (that we know): {(2,1), (2,3), (3,1)}
Non-terminal states: {(1,2), (1,3), (2,2), (3,2)}

Actions: {N, S, E, W}

rewards: 0 if not terminal state
r(2,1) = -90
r(2,3) = -90
r(3,1) = 60

iscord

Q-tables without terminal states:

Q-table (EP 0)

	N	S	E	W
(1,2)	0	0	0	0
(1,3)	0	0	0	0
(2,2)	0	0	0	0
(3,2)	0	0	0	0

Q-table (EP 1)

	N	S	E	W
(1,2)	0	0	0	0
(1,3)	0	0	0	0
(2,2)	0	<0	0	0
(3,2)	0	0	0	0

Q-table (EP2)

	N	S	E	W
(1,2)	0	0	0	0
(1,3)	0	0	0	0
(2,2)	0	<0	0	0
(3,2)	0	0	0	0

Q-table (EP3)

	N	S	E	W
(1,2)	0	0	>0	0
(1,3)	0	>0	0	0
(2,2)	0	<0	>0	0
(3,2)	0	>0	0	0

Q-table (EP4)

	N	S	E	W
(1,2)	0	0	>0	0
(1,3)	0	>0	0	0
(2,2)	0	<0	>0	0
(3,2)	0	>0	0	0

Q-table (EP5)

	N	S	E	W
(1,2)	0	0	>0	0
(1,3)	0	>0	0	0
(2,2)	<0	<0	>0	0
(3,2)	0	>0	0	0

Based on formula for q-update, I see that for instance in episode 1 only the q-value for (2,2) will change, and as it is negative it will not be the max q-value and the values of (1,3) and (1,2) therefore won't change.

The same thing happens in episode 2.

In episode 3 we get a positive q-value of (3,2), S, so the previous states will also get positive q-values.

Episode 4 is the same as episode 3.

In episode 5 we see that the q-value for (2,2),N will become negative, however the max q-value for (2,2) is still positive, so this will not affect the previous state's q-values to become negative.

So both q((2,2),E), q((3,2),S) and q((1,2),E) become positive in episode 3, but we need the change of value in (q(3,2),S) before the change of value in q((2,2),E) and the change of value in q((2,2),E) before the change of value in q((1,2),E). Therefore the iterations are different.

$Q((2,2), \mathrm{E}) =$ E:3, t:2

$Q((3,2), \mathrm{S}) =$ E:3, t:1

$Q((1,2), \mathrm{E}) =$ E:3, t:3