COMP3411/9814 Artificial Intelligence 20T0, 2020

Tutorial Solutions - Week 5 tutorial 10

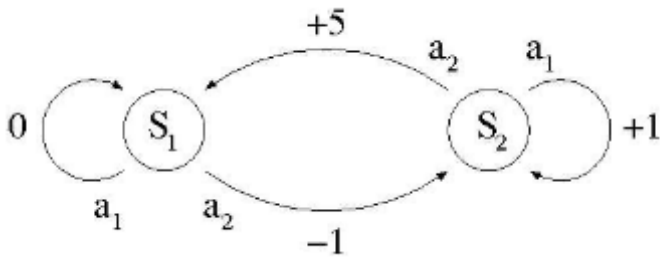
Tutorial 10: Reinforcement Learning

10.1 (Activity 9.2: Q-Learning - Open learning)

Consider a world with two states S = {S_1, S_2} and two actions A = {a_1, a_2}, where the transitions δ and reward *r* for each state and action are as follows:

$$\begin{array}{ll} \delta(S_1, a_1) = S_1 & r(S_1, a_1) = 0 \\ \delta(S_1, a_2) = S_2 & r(S_1, a_2) = -1 \\ \delta(S_2, a_1) = S_2 & r(S_2, a_1) = +1 \\ \delta(S_2, a_2) = S_1 & r(S_2, a_2) = +5 \end{array}$$

(i) Draw a picture of this world, using circles for the states and arrows for the transitions.



(ii) Assuming a discount factor of γ = 0.9, determine:

(a) The optimal policy is:

$$\pi^*(S_1) = a_2$$
$$\pi^*(S_2) = a_2$$

(b) The optimal value function V^* is calculated as follows.

$$V^*(S_1) = -1 + \gamma V^*(S_2)$$
$$V^*(S_2) = 5 + \gamma V^*(S_1)$$

So $V^*(S_1) = -1 + 5\gamma + \gamma^2 V^*(S_1)$
i.e. $V^*(S_1) = (-1 + 5\gamma)/(1 - \gamma^2) = 3.5/0.19 = 18.42$

$$V^*(S_2) = 5 + \gamma V^*(S_1) = 5 + 0.9 * 3.5/0.19 = 21.58$$

(c) The Q function for the optimal policy is calculated as follows.

$$Q(S_1, a_1) = \gamma V^*(S_1) = 16.58$$
$$Q(S_1, a_2) = V^*(S_1) = 18.42$$
$$Q(S_2, a_1) = 1 + \gamma V^*(S_2) = 20.42$$
$$Q(S_2, a_2) = V^*(S_2) = 21.58$$

(iii)Write the Q values in a table.

Q	a_1	a_2
S_1	16.58	18.42
S_2	20.42	21.58

(iv) Trace through the first few steps of the Q-learning algorithm, with all Q values initially set to zero. Explain why it is necessary to force exploration through probabilistic choice of actions in order to ensure convergence to the true Q values.

current state	chosen action	new Q value
S_1	a_1	$0 + \gamma * 0 = 0$
S_1	a_2	$-1 + \gamma * 0 = -1$
S_2	a_1	$1 + \gamma * 0 = 1$

At this point, the table looks like this:

Q	a_1	a_2
S_1	0	-1
S_2	1	0

If the agent always chooses the current best action, it can have a policy where it always prefers a suboptimal action, e.g. a_1 in state S_2, so will never sufficiently explore action a_2. This means that $Q(S_2,a_2)$ will remain zero forever, instead of converging to the true value of 21.58. With exploration, the next few steps might look like this:

current state	chosen action	new Q value
S_2	a_2	$5+\gamma*0=5$
S_1	a_1	$0+\gamma*0=0$
S_1	a_2	$-1+\gamma*5=3.5$
S_2	a_1	$1+\gamma*5=5.5$
S_2	a_2	$5+\gamma*3.5=8.15$

Now we have this table:

Q	a_1	a_2
S_1	0	3.5
S_2	5.5	8.15

From this point on, the agent will prefer action a_2 both in state S_1 and in state S_2. Further steps refine the Q value estimates, and, in the limit, they will converge to their true values.

current state	chosen action	new Q value
S_1	a_1	$0+\gamma*3.5=3.15$
S_1	a_2	$-1+\gamma*8.15=6.335$
S_2	a_1	$1+\gamma*8.15=8.335$
S_2	a_2	$5+\gamma*6.34=10.70$
...	...	...