hw5_rl_q2_model_based_rl_cycle

Question 2: Model-Based RL: Cycle

0 points possible (ungraded)

We recommend you work out the solutions to the following questions on a sheet of scratch paper, and then enter your results into the answer boxes.

Consider an MDP with 3 states, A, B and C; and 2 actions Clockwise and Counterclockwise. We do not know the transition function or the reward function for the MDP, but instead, we are given samples of what an agent experiences when it interacts with the environment (although, we do know that we do not remain in the same state after taking an action). In this problem, we will first estimate the model (the transition function and the reward function), and then use the estimated model to find the optimal actions.

To find the optimal actions, model-based RL proceeds by computing the optimal V or Q value function with respect to the estimated T and R. This could be done with any of value iteration, policy iteration, or Q-value iteration. Last week you already solved some exercises that involved value iteration and policy iteration, so we will go with Q value iteration in this exercise.

Consider the following samples that the agent encountered. (Note that nan stands for not-a-number and indicates that this entry cannot be estimated from the samples.)

s	a	s'	r
A	Clockwise	B	0.0
A	Clockwise	C	1.0
A	Clockwise	C	1.0
A	Clockwise	B	0.0
A	Clockwise	B	0.0
A	Counterclockwise	C	1.0
A	Counterclockwise	C	1.0
A	Counterclockwise	C	1.0
A	Counterclockwise	B	0.0
A	Counterclockwise	C	1.0

s	a	s'	r
B	Clockwise	A	3.0
B	Clockwise	C	0.0
B	Clockwise	C	0.0
B	Clockwise	C	0.0
B	Clockwise	C	0.0
B	Counterclockwise	A	-7.0
B	Counterclockwise	A	-7.0
B	Counterclockwise	C	0.0
B	Counterclockwise	A	-7.0
B	Counterclockwise	A	-7.0

s	a	s'	r
C	Clockwise	A	0.0
C	Clockwise	B	-10.0
C	Clockwise	B	-10.0
C	Clockwise	A	0.0
C	Clockwise	B	-10.0
C	Counterclockwise	A	0.0
C	Counterclockwise	B	-9.0
C	Counterclockwise	B	-9.0
C	Counterclockwise	A	0.0
C	Counterclockwise	A	0.0

Part 1

We start by estimating the transition function, T(s,a,s') and reward function R(s,a,s') for this MDP. Fill in the missing values in the following table for T(s,a,s') and R(s,a,s').

Discount Factor, γ = 0.5

s	a	s'	T(s,a,s')	R(s,a,s')
A	Clockwise	B	M	N
A	Clockwise	C	O	P
A	Counterclockwise	B	0.200	0.000
A	Counterclockwise	C	0.800	1.000
B	Clockwise	A	0.200	3.000
B	Clockwise	C	0.800	0.000
B	Counterclockwise	A	0.800	-7.000
B	Counterclockwise	C	0.200	0.000
C	Clockwise	A	0.400	0.000
C	Clockwise	B	0.600	-10.000
C	Counterclockwise	A	0.600	0.000
C	Counterclockwise	B	0.400	-9.000

M	0.6 Answer: **0.600**
N	0 Answer: **0.000**
O	0.4 Answer: **0.400**
P	1 Answer: **1.000**

Part 2

Now we will run Q-iteration using the estimated T and R functions. The values of $Q_k\ (s, a)$, are given in the table below.

	A	B	C
Clockwise	-0.14	-0.76	-5.66

Counterclockwise	-0.58	-5.64	-3.24

Fill in the values for $Q_{k+1}\ (s, a)$.

	A	B	C
Clockwise	-0.476 Answer: -0.476	-0.71 Answer: -0.710	-6.256 Answer: -6.256
Counterclockwise	-0.572 Answer: -0.572	-5.98 Answer: -5.980	-3.794 Answer: -3.794

Part 3

Suppose Q-iteration converges to the following Q^* function, $Q^*\ (s, a)$.

	A	B	C
Clockwise	-0.734	-1.089	-6.473
Counterclockwise	-0.924	-6.297	-4.038

What is the optimal action, either Clockwise or Counterclockwise, for each of the states?

A	B	C
Clockwise Answer: *Clockwise*	Clockwise Answer: *Clockwise*	Counterclockwise Answer: *Counterclockwise*

For this problem, you may press "Check" as many times as you want without resetting the problem, so that you don't have to reset the problem for trivial math mistakes.

Part 1:

For the transition function, find the expectation for the specified initial state and action:

M: $\frac{\textbf{num samples ending at B}}{\textbf{total num samples of clockwise from A}}$

O: $\frac{\textbf{num samples ending at C}}{\textbf{total num samples of clockwise from A}}$

For the reward function, simply find a sample with s, a, s' matching the s, a, s' from the table

Part 2:

To find the Q values, plug in the estimated T and R functions, found in the previous part, into: $Q_{k+1}(s, a) = sum_{s'} T(s, a, s') [R(s, a, s') + \gamma * max_{a'} Q_k(s', a')]$

Part 3:

For each state, take the action, either clockwise or counterclockwise, that has the highest Q^* value according to the table.

Submit