6.034 MDP Homework Due 4/20

Name:

Changes (4/18)

1.1. We've further specified that gamma (the discount factor) is between 0 and 1 (exclusive) and provided a hint

1.2. We have changed this problem to fix a mistake in the states we gave, and we also provided some constants that should make your math easier

1.6 We fixed a typo in a column header

1. Value Iteration for Survivalists

You live in the words and made due by hunting and resting. But you don't know when to hunt and when to rest. Your survivalist manual says to use MDPs in this situation, so you complete problem 1 below.

Actions and Transitions

From State	Action	Probability	Result
Hungry, Tired	Rest	1.0	Hungry, Rested
	Hunt	0.8	Hungry, Tired
		0.2	Full, Tired
Hungry, Rested	Rest	1.0	Hungry, Rested
	Hunt	0.8	Full, Rested
		0.2	Hungry, Tired
Full, Tired	Rest	1.0	Hungry, Rested
	Hunt	0.8	Hungry, Tired
		0.2	Full, Tired
Full, Rested	Rest	1.0	Hungry Rested
	Hunt	0.8	Full, Tired
		0.2	Hungry, Tired

Rewards

The value of transitioning *into* a state is:

R(*,*,State)	Has Value
Hungry, Tired	0
Hungry, Rested	1
Full, Tired	1
Full, Rested	2

1.1 What is V^{π} (Hungry-Tired) for the policy that always rests? Assume gamma (the discount factor) is some number between 0 and 1 exclusive. Give your answer in terms of gamma. **(Hint: think about geometric series)**

1.2 What is Q(Hungry-Rested, Hunt) assuming the following

V*(Full-Rested) = k
Q(Hungry-Tired, Rest) = A
Q(Hungry-Tired,Hunt) = B
B > A
Discount Factor = Gamma (between 0 and 1)

Answer in terms of the variables above

1.3 Value Iteration. Perform value iteration, for three iterations, filling in the tables below. (To make computation easier, assume the discount factor is 1.)

Itr	V*(Hungry-Tired)	V*(Hungry,Rested)	V*(Full,Tired)	V*(Full,Rested)
0	**0**	**0**	**0**	**0**
1				
2				
3				

1.4 What is the policy at iteration 3?

1.5 Temporal Difference Learning. You have a fixed policy that always hunts at state **HungryTired** and you'd like to do Temporal Difference learning. Assuming:

- V*(FullTired) = 2
- Learning rate = 0.5
- Discount factor = 1

Fill out the table after each sample is observed:

Observed	**V^{π} (HungryTired)**
(initial value)	**0**
(HungryTired, Hunt, HungryTired) Reward = 0	
(HungryTired, Hunt, FullTired) Reward = 1	
(HungryTired, Hunt, FullTired) Reward = 1	
(HungryTired, Hunt, FullTired) Reward = 1	

1.6 Q Learning. Suppose we experience the following sequence of actions and results

	S	A	S'	Reward
1)	Hungry,Tired	Rest	Hungry,Rested	1
2)	Hungry,Tired	Hunt	Hungry,Tired	0
3)	Hungry,Rested	Rest	Hungry,Rested	1
4)	Hungry,Rested	Hunt	Full,Rested	2
5)	Full,Rested	Rest	Hungry,Rested	1
6)	Hungry,Rested	Hunt	Hungry,Tired	0

What are the resulting Q(s,a) values below if the learning ratio is 0.5, the discount is 1, and we start with Q(s,a)=0 for all (s,a)? Fill in the tabe below; each row should hold the q-values after the transition specified from the corresponding transition in the list above. You may leave unchanged values blank.

Act	Q Hung-Tired, Hunt	Q Hung-Tired, Rest	Q Hung-Rested, Hunt	Q Full-Rested Rest	Q Hung-Rested, Rest
0	**0**	**0**	**0**	**0**	**0**
1					
2					
3					
4					
5					
6					

1.7 Given the sequence of actions above, what is the value of the transitions:

T(Hungry-Tired,Hunt,Hungry-Tired) =

T(Hungry-Rested,Hunt,Hungry-Tired) =

2. Problem 17-10 b and c from the Textbook

Consider an undiscounted MDP having three states (1,2,3) with rewards -1, -2, and 0 respectively. State 3 is a terminal state. In states 1 and 2, there are two possible actions: *a* and *b*. The transition model is as follows:

- In state 1, action *a* moves the agent to state 2 with probability 0.8 and makes the agent stay put with probability 0.2
- In State 2, action *a* moves the agent to state 1 with probability 0.8 and makes the agent stay put with probability 0.2
- In either state 1 or 2, action *b* moves the agent to state 3 with probability 0.1 and makes the agent stay put with probability 0.9

b. Apply policy iteration, showing each step in full, to determine the optimal policy and the values of states 1 and 2. Assume that the initial policy has action *b* in both states.

c. What happens to policy iteration if the initial policy has action *a* in both states Does discounting help? Does the optimal policy depend on the discount factor?

Final Note

If you are looking for more practice with MDPs, we suggest problems 17-8 and 17-9 in the textbook.