COSC1125/1127 Artificial Intelligence

School of Science (Computer Science)
RMIT University
Semester 1, 2018
A/Prof. Sebastian Sardina

Tutorial Sheet 7
Markov Decision Processes (MDPs)

1. Recall a Markov Decision Process (MDP) represents a scenario where actions are non-deterministic. Consider the following minor gridworld example:

	1	2	3
1	X		X
2			
3	X		

Cells are indexed by (row, column). Cells $(1,1)$, $(3,1)$ and $(1,3)$ are terminal cells (marked with X), and have rewards of 20, -20 and 5, respectively. For all other cells, there is an reward of -1. The agent starts at cell $(3,3)$. The agent can move in north, east, south and west directions or stay where it is. The actions are stochastic and have the following outcomes:

- If go north, 70% of the time the agent will go north, 30% will go west.
- If go east, 60% of the time the agent will go east, 40% of the time will go south.
- If go south, 80% of the time the agent will go south, 20% of the time will go east.
- If go west, 70% of the time the agent will go west, 30% of the time will go north.
- If stay in current cell, 100% of the time will stay.

If an action causes the agent to travel into a wall, the agent will stay in their current cell.

Construct a MDP for this instance of gridworld. Recall a MDP consists of S (set of states), s_0 (starting state), possible terminal states, A (set of actions), T (transition model/function) and R (reward function), so all of these have to be specified. To help, consider constructing the MDP in steps:

a) What are the states in this gridworld?

b) What are the starting and terminal states?

c) What are the set of actions?

d) With cell $(3,3)$ as the current state s, construct the transition model/function for all actions and subsequent states, i.e., $T(s,a,s')$, where a are all possible actions for $(3,3)$ and s' are all possible successor states from $(3,3)$.

e) What is the reward function?

Answer

- Each grid corresponds to a state.
- Starting state is $(3, 3)$, terminal states are $(1, 1)$, $(3, 1)$ and $(1, 3)$.
- For each state, actions are $\{north, east, south, west, stay\}$:
  - $T((3, 3), north, (2, 3)) = 0.7$ (70%, go north)
  - $T((3, 3), north, (3, 2)) = 0.3$ (30%, go west)
  - $T((3, 3), east, (3, 3)) = 0.6$ (60%, go east)
  - $T((3, 3), east, (3, 3)) = 0.4$ (40%, go south)
  - $T((3, 3), south, (3, 3)) = 0.8$ (80%, go south)
  - $T((3, 3), south, (3, 3)) = 0.20$ (20%, go east)
  - $T((3, 3), west, (3, 2)) = 0.7$ (70%, go west)
  - $T((3, 3), west, (2, 3)) = 0.3$ (30%, go north)
  - $T((3, 3), stay, (3, 3)) = 1.0$ (100%, stay)
- Reward:
  - $R((1, 1)) = +20$
  - $R((3, 1)) = -20$
  - $R((1, 3)) = +5$
  - $R(\text{other states}) = -1$

2. Using the MDP built for question 1, consider the following sequences of states that our agent progressed through. For each sequence, compute the *additive reward* and *discounted reward* (with a decay factor γ of 0.5). Take note of the differences between the two approaches.

   a) $(3, 3)$, $(2, 3)$, $(2, 2)$, $(2, 1)$, $(1, 1)$

   b) $(3, 3)$, $(3, 2)$, $(2, 2)$, $(2, 3)$, ... (repeat)

Answer

(a) Additive: $(-1) + (-1) + (-1) + (-1) + (+20) = 16$

Discount: $(-1) + 0.5(-1) + 0.5^2(-1) + 0.5^3(-1) + 0.5^4(+20) = -0.625$ $(10/16)$

(b) Additive: $(-1) + (-1) + (-1) + (-1) + \ldots = -\infty$

Discount: $(-1)+0.5(-1)+0.5^2(-1)+0.5^3(-1)+0.5^4(-1)+\ldots = \sum_{t=0}^{\infty} 0.5^t * (-1) = \frac{-1}{1-0.5} = -2$

3. Consider the MDP from question 1, with a decay factor γ of 1.

a) Using value iteration (check algo here), compute the utilities for each state after two iterations. For this question take (observe we are using the simplest reward scheme $R(s)$ rather than $R(s, a, s')$):

$$V_{k+1}(s) = R(s) + \max_a \{\Sigma_{s' \in S} T(s, a, s') V_k(s')\}.$$

Initially, we initialize $V_0(s) = 0$ for all states $s \in S$. Remember terminal states do not get more rewards once reached.

b) Repeat and determine the values after three iterations.

Answer

Check this link for the detailed workings on cell $(1, 2)$. Note that, in the case of $(1, 2)$, it is clear that "moving west" is the best action (i.e., maximizes the second term in formula above), but in general one would need to calculate the term for every action and then keep the the max one (that is the best action to do!).

a)

	1	2	3
1	20	12.7	5
2	12.7	−2	2.2
3	−20	−2	−2

b)

	1	2	3
1	20	16.81	5
2	16.81	11.7	1.9
3	−20	−3	−0.059

4. Consider the utilities after two iterations of value iteration from question 3 (we typically only perform this when values have converged, but for this question, do this after 2 iterations). Using the Maximum Expected Utility principle, find the best action (i.e., policy) for each state.

Answer

For the cell $(1, 2)$ we can derive (as explained here) that the best action to do there after 2 iteration is west, that is, $\pi((1, 2)) = \text{west}$.

5. Consider the following policy for the MDP of question 1, with a decay factor γ of 1:

- $\pi((2, 1)) = \text{west}$.
- $\pi((2, 2)) = \text{west}$.
- $\pi((2, 3)) = \text{north}$.
- $\pi((3, 2)) = \text{north}$.
- $\pi((3, 3)) = \text{east}$.
- $\pi((1, 2)) = \text{south}$.

Evaluate this policy, i.e., compute the utility for each state, for one and two iterations (we typically can solve it as a system of simultaneous equations but for this question we use an iterative approach).

Answer

Here we are given the policy (it is fixed), so the task is to calculate its value at every state, that is, $V_\pi(s)$ or $U_\pi(s)$. We can do that as follows:

$$V^\pi_{k+1}(s) = R(s) + \Sigma_{s' \in S} T(s, \pi(s), s') V^\pi_k(s').$$

Observe we do not take the $\max$ of actions because we know, at every state s, what we should do, namely, $\pi(s)$!

BE CAREFUL, THIS HAS NOT BEEN RE-CHECKED CAREFULLY!

	1	2	3
1	20	-0.8	5
2	4.3	-2	-2.2
3	-20	-2	-2

6. Will the stay action be used for an optimal policy? Hint: consider the reward function for non-terminal cells.

Answer

If the reward function is positive valued and large in magnitude for non terminating states, and γ is small, then it is possible for an agent to choose to stay in a state.