Prof. Dr. techn. Wolfgang Nejdl
Franziska Schoger

November 15th 2023

Advanced Topics in AI

Exercise 4

Question 1: MDPs: Micro-Blackjack

In micro-blackjack, you repeatedly draw a card (with replacement) that is equally likely to be a 2, 3, or 4. You can either Draw or Stop if the total score of the cards you have drawn is less than 6. If your total score is 6 or higher, the game ends, and you receive a utility of 0. When you Stop, your utility is equal to your total score (up to 5), and the game ends. When you Draw, you receive no utility. There is no discount ($\gamma = 1$). Let's formulate this problem as an MDP with the following states: 0, 2, 3, 4, 5 and a Done state, for when the game ends.

a. What is the transition function and the reward function for this MDP?

b. Fill in the following table of value iteration values for the first 4 iterations.

States	0	2	3	4	5
V_0					
V_1					
V_2					
V_3					
V_4					

c. You should have noticed that value iteration converged above. What is the optimal policy for the MDP?

States	0	2	3	4	5
π^*					

d. Perform one iteration of policy iteration for one step of this MDP, starting from the fixed policy below:

States	0	2	3	4	5
π_i	Draw	Stop	Draw	Stop	Draw
V^{π_i}					
π_{i+1}					

e. Is the policy π_{i+1} optimal?

Question 2: Golf as an MDP

In this exercise we will formulate golf as an MDP as follows:

- State Space : {Tee, Fairway, Sand, Green}
- Actions : {Conservative shot, Power shot}
- Initial State : Tee
- Terminal State : Green
- Transition model : (note that any successor state not on this list has a transition probability 0, and “Conservative” stands for “Conservative shot”)

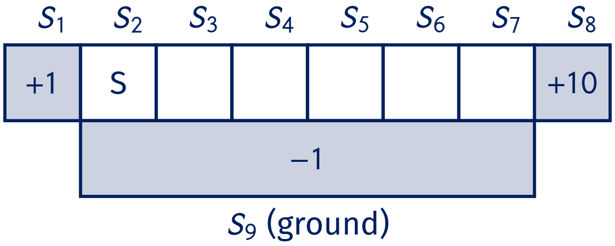
s	a	s'	$T(s,a,s')$
Tee	Conservative	Fairway	0.9
Tee	Conservative	Sand	0.1
Tee	Power shot	Green	0.5
Tee	Power shot	Sand	0.5
Fairway	Conservative	Green	0.8
Fairway	Conservative	Sand	0.2
Sand	Conservative	Green	1.0

- Rewards: (note: $R(\cdot,\cdot,s)$ means that the reward is received for transitioning to state s, regardless of the action taken or previous state)

s'	$R(\cdot,\cdot,s')$
Fairway	-1
Sand	-2
Green	3

Draw a state graph defining this MDP problem. A state graph shows the states as nodes and has the actions as arcs labeled with T and R values. Remember, in a state graph no states repeat.

Question 3: Robot Balancing



Consider the above MDP, representing a robot on a balance beam. Each grid square is a state and the available actions are right and left. The agent starts in state s_2 , and all states have reward 0 aside from the ends of the grid s_1 and s_8 and the ground state, which have the rewards shown. Moving left or right results in a move left or right (respectively) with probability p. With probability $1-p$, the robot falls off the beam (transitions to ground, and receives a reward of -1). Falling off, or reaching either endpoint, result in the end of the episode (i.e., they are terminal states). Note that terminal states receive no future reward.

a. For what values of p is the optimal action from s_2 to move right if the discount γ is 1?

b. For what values of γ is the optimal action from s_2 to move right if $p = 1$?

c. Given initial value estimates of zero, show the results of one, then two rounds of value iteration (assume $\gamma = 1$).

Question 4: Dice Bonanza

A casino is considering adding a new game to their collection, but need to analyze it before releasing it on their floor. They have hired you to execute the analysis. On each round of the game, the player has the option of rolling a fair 6-sided die. That is, the die lands on values 1 through 6 with equal probability. Each roll costs 1 dollar, and the player must roll the very first round. Each time the player rolls the die, the player has two possible actions:

1. $Stop$: Stop playing by collecting the dollar value that the die lands on, or
2. $Roll$: Roll again, paying another 1 dollar.

Having taken "Introduction to Artificial Intelligence", you decide to model this problem using an infinite horizon Markov Decision Process (MDP). The player initially starts in state $Start$, where the player only has one possible action: $Roll$. State s_i denotes the state where the die lands on i. Once a player decides to $Stop$, the game is over, transitioning the player to the End state.

(a) In solving this problem, you consider using policy iteration. Your initial policy π is in the table below. Evaluate the policy at each state, with γ = 1. Build a system of linear equations and solve directly for $v^\pi(s_i),\ i \in \{1, \dots, 6\}$.

State	s_1	s_2	s_3	s_4	s_5	s_6
$\pi(s)$	$Roll$	$Roll$	$Stop$	$Stop$	$Stop$	$Stop$
$v^\pi(s)$						

(b) Having determined the values, perform a policy update to find the new policy π'. The table below shows the old policy π and has filled in parts of the updated policy π' for you. If both Roll and Stop are viable new actions for a state, write down both $Roll/Stop$. In this part as well, we have $\gamma = 1$.

State	s_1	s_2	s_3	s_4	s_5	s_6
$\pi(s)$	$Roll$	$Roll$	$Stop$	$Stop$	$Stop$	$Stop$
$\pi'(s)$	$Roll$					$Stop$

(c) Is $\pi(s)$ from part (a) optimal? Explain why or why not.

(d) Suppose that we were now working with some $\gamma \in [0, 1]$ and wanted to run **value iteration**. Select the **one** statement that would hold true at convergence, or write the correct answer next to Other if none of the options are correct.

○ $V^*(s_i) = \max\left\{-1 + \frac{i}{6}, \sum_j \gamma V^*(s_j)\right\}$

○ $V^*(s_i) = \max\left\{i, \frac{1}{6}\left[-1 + \sum_j \gamma V^*(s_j)\right]\right\}$

○ $V^*(s_i) = \max\left\{-\frac{1}{6} + i, \sum_j \gamma V^*(s_j)\right\}$

○ $V^*(s_i) = \max\left\{i, -\frac{1}{6} + \sum_j \gamma V^*(s_j)\right\}$

○ $V^*(s_i) = \frac{1}{6} \cdot \sum_j \max\left\{i, -1 + \gamma V^*(s_j)\right\}$

○ $V^*(s_i) = \frac{1}{6} \cdot \sum_j \max\left\{-1 + i, \sum_k V^*(s_j)\right\}$

○ $V^*(s_i) = \sum_j \max\left\{-1 + i, \frac{1}{6}\gamma V^*(s_j)\right\}$

○ $V^*(s_i) = \sum_j \max\left\{\frac{i}{6}, -1 + \gamma V^*(s_j)\right\}$

○ $V^*(s_i) = \max\left\{i, -1 + \frac{\gamma}{6} \sum_j V^*(s_j)\right\}$

○ $V^*(s_i) = \sum_j \max\left\{i, -\frac{1}{6} + \gamma V^*(s_j)\right\}$

○ $V^*(s_i) = \sum_j \max\left\{-\frac{i}{6}, -1 + \gamma V^*(s_j)\right\}$

Question 5: Policy Evaluation with Q-Values

In this question, you will be working in an MDP with states S, actions A, discount factor γ, transition function T, and reward function R.

We have some fixed policy $\pi : S \longrightarrow A$, which returns an action $a = \pi(s)$ for each state $s \in S$. We want to learn the Q function $Q^\pi(s, a)$ for this policy: the expected reward from taking action a in state s and then continuing to act according to π : $Q^\pi(s, a) = \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma Q^\pi(s', \pi(s'))]$. The policy π will not change.
Can we guarantee anything about how the values Q^π compared to the values Q^* for an optimal policy π^* ?

○ $Q^\pi(s, a) \leq Q^*(s, a)$ for all s, a

○ $Q^\pi(s, a) = Q^*(s, a)$ for all s, a

○ $Q^\pi(s, a) \geq Q^*(s, a)$ for all s, a

○ None of the above guaranteed