Markov Decision Processes

Alice Gao

Lecture 18

Readings: RN 17.1. PM 9.5.

Outline

Learning Goals

By the end of the lecture, you should be able to

- Describe motivations for modeling a decision problem as a Markov decision process.
- Describe components of a fully-observable Markov decision process.
- Describe reasons for using a discounted reward function.
- Define the policy of a Markov decision process.
- Give examples of how the reward function affects the optimal policy of a Markov decision process.

Learning Goals

Introduction to Markov Decision Processes

A Grid World

Policies

The optimal policies of the grid world

Determine the Optimal Policy Given $V^*(s)$

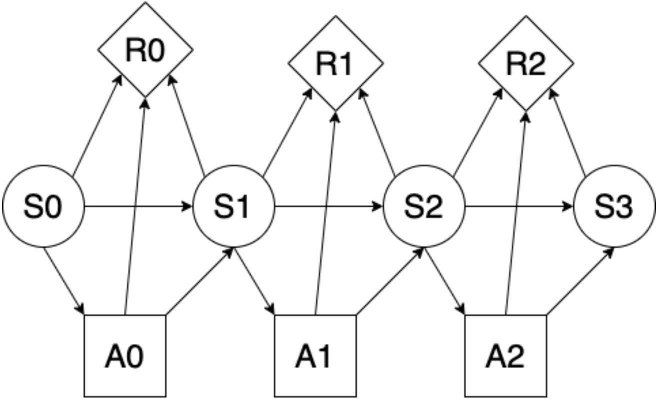
Revisiting the Learning goals

Modeling an Ongoing Decision Process

- Finite-stage v.s. ongoing problems

- Utility at the end v.s. a sequence of rewards

A Markov Decision Process



Rewards

- Total reward
- Average reward
- Discounted reward

Variations of MDP

- A fully-observable MDP

- A partially observable MDP (POMDP) combines a MDP and a hidden Markov model.

A 3×4 Grid World Problem

What should the robot do to maximize its rewards?

	1	2	3	4
1	Start			
2		X		-1
3				+1

- Let s_{ij} be the position in row i and column j.
- s_{11} is the initial state.
- There is a wall at s_{22}.
- s_{24} and s_{34} are goal states.
  The robot escapes the world at either goal state.

An MDP for the 3×4 Grid World

- There are four actions: up, down, left, and right.
  Every action is possible in every state.

- The transition model $P(s'|s, a)$.
  An action achieves its intended effect with probability 0.8.
  An action leads to a 90-degree left turn with probability 0.1.
  An action leads to a 90-degree right turn with probability 0.1.
  If the robot bumps into a wall, it stays in the same square.

- The reward function $R(s)$ is the reward of entering state s.
  $R(s_{24}) = -1$.
  $R(s_{34}) = 1$.
  Otherwise, $R(s) = -0.04$.

CQ: Understanding the transition model

CQ: The robot is in s_{14} and tries to move to our right, what is the probability that the robot stays in s_{14}?

(A) 0.1

(B) 0.2

(C) 0.8

(D) 0.9

(E) 1.0

CQ: A fixed sequence of actions

CQ: If the environment is deterministic, an optimal solution to the grid world problem is the fixed action sequence:
down, down, right, right, and right.

(A) True

(B) False

(C) I don't know

CQ: A fixed sequence of actions

CQ: Consider the action sequence "down, down, right, right, and right". This action sequence could take the robot to more than one square with positive probability.

(A) True

(B) False

(C) I don't know

Policies

A policy specifies what the agent should do
as a function of the current state.

A policy is

- non-stationary if it is a function of the state and the time.
- stationary if it is a function of the state.

The optimal policies of the grid world

The optimal policy of the grid world changes based on $R(s)$ for any non-goal state s. It shows a careful balancing of risk and reward.

	1	2	3	4
1	Start			
2		X		-1
3				$+1$

The optimal policy when life is ...

When the reward function is

- $R(s) < -1.6284$
- $-0.4278 < R(s) < -0.0850$
- $R(s) = -0.04$
- $-0.0221 < R(s) \leq 0$
- $0 < R(s)$

	1	2	3	4
1	Start			
2		X		-1
3				$+1$

The optimal policy when life is quite unpleasant

When $-0.4278 < R(s) < -0.0850$,
what does the optimal policy look like?

	1	2	3	4
1	Start			
2		X		-1
3				$+1$

The optimal policy when life is painful

When $R(s) < -1.6284$,
what does the optimal policy look like?

	1	2	3	4
1	Start			
2		X		-1
3				$+1$

The optimal policy when life is unpleasant

When $R(s) = -0.04$,
what does the optimal policy look like?

	1	2	3	4
1	Start			
2		X		-1
3				$+1$

The optimal policy when life is only slightly dreary

When $-0.0221 < R(s) \leq 0$,
what does the optimal policy look like?

	1	2	3	4
1	Start			
2		X		-1
3				$+1$

The optimal policy when life is GOOD =D

When $R(s) > 0$,
what does the optimal policy look like?

	1	2	3	4
1	Start			
2		X		-1
3				$+1$

Learning Goals

Introduction to Markov Decision Processes

A Grid World

Policies

The optimal policies of the grid world

Determine the Optimal Policy Given $V^*(s)$

Revisiting the Learning goals

The Expected Utility of a Policy

$V^{\pi}(s)$: expected utility of entering state s and following the policy π thereafter.

$V^{*}(s)$: expected utility of entering state s and following the optimal policy π^{*} thereafter.

The Values of $V^*(s)$

	1	2	3	4
1	0.705	0.655	0.611	0.388
2	0.762	X	0.660	-1
3	0.812	0.868	0.918	$+1$

Figure: $V^*(s)$ for $\gamma = 1$ and $R(s) = -0.04, \forall s \neq s_{24}, s \neq s_{34}$.

Calculate the Optimal Policy Given $V^*(s)$

Calculate my expected utility if I am in state s and take action a.

$$Q^*(s,a) = \sum_{s'} P(s'|s,a) V^*(s') \tag{1}$$

In state s, choose an action that maximizes my expected utility.

$$\pi^*(s) = \arg\max_a Q^*(s,a) \tag{2}$$

CQ: Determine optimal action given $V^*(s)$

CQ: What is the optimal action for state s_{13}?
(A) Up (B) Down (C) Left (D) Right

$$Q^*(s,a) = \sum_{s'} P(s'|s,a)V^*(s')$$
$$\pi(s) = \arg\max_a Q^*(s,a).$$

The values of $V^*(s)$ are given below.

	1	2	3	4
1	0.705	0.655	0.611	0.388
2	0.762	X	0.660	-1
3	0.812	0.868	0.918	+1

Revisiting the Learning Goals

By the end of the lecture, you should be able to

- Describe motivations for modeling a decision problem as a Markov decision process.
- Describe components of a fully-observable Markov decision process.
- Describe reasons for using a discounted reward function.
- Define the policy of a Markov decision process.
- Give examples of how the reward function affects the optimal policy of a Markov decision process.