Markov Decision Processes (I)

University of BRISTOL

EMAT31530/Nov 2020/Xiaoyang Wang

Outline

- Machine learning
- Search

Binary classification:

$x \to \boxed{?} \to y \in \{-1, +1\}$, single action

Search problem:

$x \to \boxed{?} \to$ action sequence $(a_1, a_2, a_3, a_4, \ldots)$

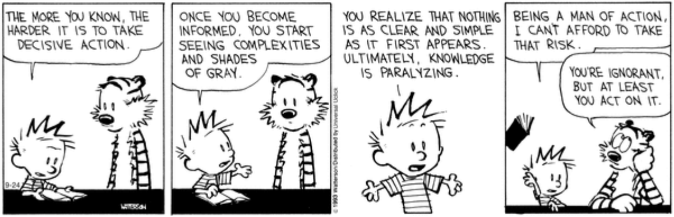
This lecture discusses complex decision making. The objective is to present the foundations of Markov Decision Processes:

- Sequential decision problems
- Rewards, Utiliy and Policies

Have a look at ...

... Russell and Norvig (Ch. 17 and Ch. 21)

... Sutton and Barto. Reinforcement Learning: An Introduction. MIT press



Example: Frozen Lake

S	F	F	F
F	H	F	H
F	F	F	H
H	F	F	G

S: Start

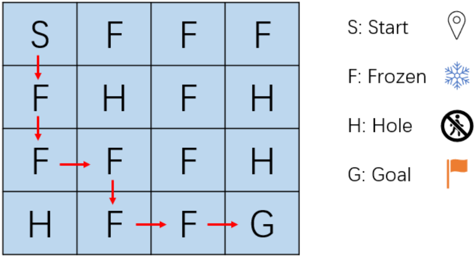
F: Frozen

H: Hole

G: Goal

Search problems

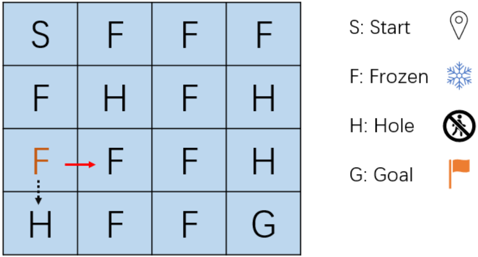
Example: Frozen Lake



action sequence: [down, down, right, down, right, right] $\rightarrow$ Goal

Environment Uncertainty

Example: Frozen Lake - slippery!



[‘slippery’ F, ‘right’] $\xrightarrow{p=?}$ F

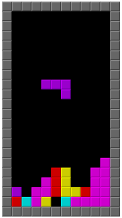
[‘slippery’ F, ‘right’] $\xrightarrow{p=?}$ H

Randomness

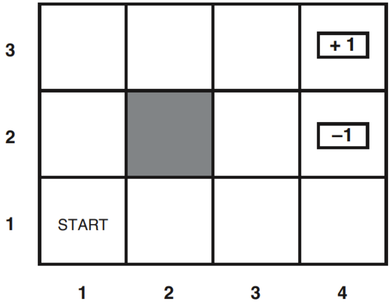
Motivation

Many important problems are MDPs ...

- Cleaning robot: hit obstacles; actuators fail
- Autonomous aircraft navigation
- Games
- Travel route planning



Example: Grid World

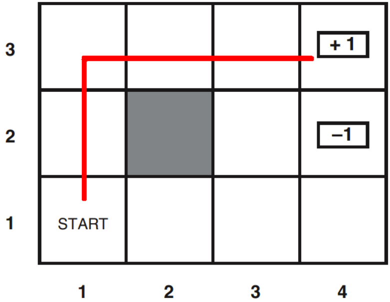


Start: (1,1)

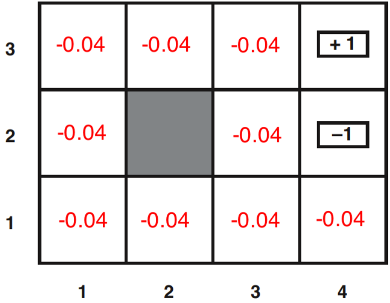
Goal: (4, 3) and (4, 2)

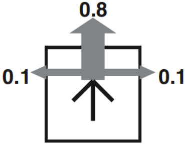
If the robot bumps the wall, it stays in the same square.

Example: Grid World



Example: Stochastic Grid World





The probability of reaching (4,3) following the previous solution:

$$0.8^5 = 0.32768$$

MDP formulation

MDP requires a structure to keep track of the decision sequences:

MDP

- s: state
- Actions(s): possible actions
- $P(s'|s, a)$ (or $T(s, a, s')$): probability of s' if take action a in state s
- Reward(s): reward for the state s
- Goal(s): whether at the end of the process
- $0 \leq \gamma \leq 1$: discount factor

Definitions

[Markov assumption] The probability of reaching s' from s depends only on s and not on the history of earlier states.

[Transition model] describes the outcome of each *action* in each *state*.

$P(s'|s,a) \rightarrow$ probability of reaching state s' if action a is done in state s.

$\sum_{s' \in \text{States}} P(s'|s,a) = 1$

$P(s'|s,a) > 0$

[Rewards] In each state s, we receive a reward $R(s)$ (positive or negative but bounded).

Solution - Policies

A solution should describe what the robot does in every state: this is called a policy, π.

- $\pi(s)$ for an individual state describes which action should be taken in s.

Q: why can't we use paths to define solutions?

Each time a given policy is executed starting from the initial state, the stochastic nature of the environment may lead to a different environment history.

The best thing to do? – Optimal policy

Policy evaluation

Utility and Value

Discounted rewards

The utility of a state sequence is

$$U_h([s_0, s_1, s_2, \ldots]) = R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \ldots,$$

where the discount factor γ is a number between 0 and 1.

if $\gamma = 0$: focus on the present
if $\gamma = 1$: the future is equally important with the present

Discount factor makes more distant future rewards less significant!

Value

Expected Utility $E[U]$

Optimal policy is one that yields the highest *expected utility*, denoted by π^*

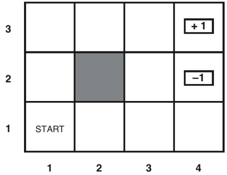
Utilities over Time

Finite horizon or infinite horizon?

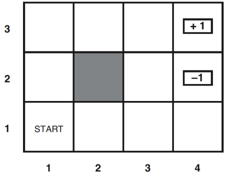
Finite horizon

There is a fixed time N after which nothing matters:

- $\forall k \quad U_h([s_0, s_1, ..., s_{N+k}]) = U_h([s_0, s_1, \ldots, s_N])$
- Leads to non-stationary optimal policies (N matters)



If $N = 2$, what is the optimal policy?



If $N = 5$, what is the optimal policy?

Utilities over Time

Finite horizon or infinite horizon?

Infinite horizon

Stationary optimal policies (time at state doesn't matter):

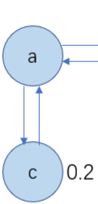
- Does **not** mean that all state sequences are infinite; it just means that there is no fixed deadline.

Choosing infinite horizon rewards creates a problem: some sequences will be infinite with infinite reward, **how do we compare them?**

[Solution 1] With discounted rewards, the utility of an infinite sequence is finite. In fact, if $\gamma < 1$ and rewards are bounded by $\pm R_{max}$, we have

$$U_h([s_0, s_1, s_2, \dots]) = \sum_{t=0}^{\infty} \gamma^t R(s_t) \leq \sum_{t=0}^{\infty} \gamma^t R_{max} = \frac{R_{max}}{(1-\gamma)}$$

[Solution 2] Compare average reward per time step.



Summary

- Markov decision processes (MDPs)
- MDP solutions - Policies
- Utility and Value