Markov Decision Processes

24. dubna 2019

B4M36PUI/BE4M36PUI — Planning for Artificial Intelligence

Outline

- MDP definition and examples
- MDP solution
- Value function calculation

Definitions

Markov Decision Process

Tuple $\langle S, A, D, T, R \rangle$:

- S: finite set of states agent can find itself in
- A: finite set of action agent can perform
- D: finite set of timesteps
- T: transition function - transitions between states
- R: reward function - rewards obtained from transitions

Markov Decision Process

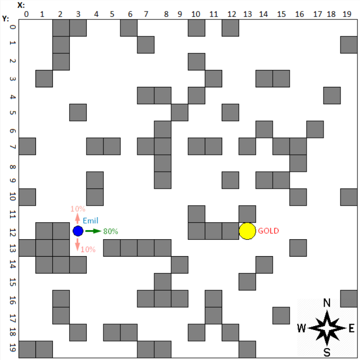
Tuple $\langle S, A, D, T, R \rangle$:

- S: finite set of states agent can find itself in
- A: finite set of action agent can perform
- D: finite set of timesteps
- T: transition function - transitions between states
- R: reward function - rewards obtained from transitions

⚠Only one of many possible definitions!

Example: Emil in the gridworld

- S: Possible Emils positions
- A: Move directions
- D: Emil has e.g. 200 steps to find gold
- T: stochastic movement, e.g. 10% to move to the side of selected action
- R: e.g. +100 for finding gold, -1 for each move



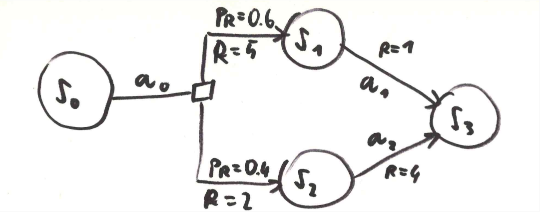
MDP example - blackjack

Blackjack

- S: Possible player hands and played cards
- A: Hit, Stand, ...
- T: Possible drawn cards,
- R: Win/loose at the end

Example: Abstract example

- S: S_0, S_1, S_2, S_3
- A: a_0, a_1, a_2
- $T:$
$$\begin{aligned} T(S_0, a_0, S_1) &= 0.6 \\ T(S_0, a_0, S_2) &= 0.4 \\ T(S_1, a_1, S_3) &= 1 \\ T(S_2, a_2, S_3) &= 1 \end{aligned}$$
- $R:$
$$\begin{aligned} R(S_0, a_0, S_1) &= 5 \\ R(S_0, a_0, S_2) &= 2 \\ R(S_1, a_1, S_3) &= 1 \\ R(S_2, a_2, S_3) &= 4 \end{aligned}$$



When MDP might be a good model?

- *Domain with uncertainty* - uncertain utoucomes of actions
- *Sequential decision making* - for *sequences* of decisions
- *Fair Nature* - no one is actively playing against us
- *Full observability, perfect sensors* - we know where agent is
- *Cyclic domain structures* - when states can be revisited

MDP Solution

Policy

Def: Policy

Assignment of action to state, $\pi : S \rightarrow A$

- *Partial policy* - e.g. output of robust replanning
- *Complete policy* - domain of π is whole state space S.
- *Stationary policy* - independent of timestep (e.g. emil)
- *Markovian policy* - dependent only on last state

⚠In general, policy can be history dependent and stochastic!

Value function (of a policy)

Def: Value function

Assignment of value to state, $V : S \to < -\infty, \infty >$

Value function (of a policy)

Def: Value function

Assignment of value to state, $V : S \rightarrow < -\infty, \infty >$

Def: Value function of a policy

Assignment of value to state based on utility of rewards obtained by following policy π from a state, $V^\pi : S \rightarrow < -\infty, \infty >$, $V^\pi(s) = u(R_1^{\pi_s}, R_2^{\pi_s}, \ldots)$

Value function (of a policy)

Def: Value function

Assignment of value to state, $V : S \rightarrow < -\infty, \infty >$

Def: Value function of a policy

Assignment of value to state based on utility of rewards obtained by following policy π from a state, $V^{\pi} : S \rightarrow < -\infty, \infty >$, $V^{\pi}(s) = u(R_1^{\pi_s}, R_2^{\pi_s}, \ldots)$

Def: Optimal MDP solution

Optimal MDP solution is a policy π^* such that value function V^{π^*} called optimal value function dominates all other value functions in all states, $\forall s V^{\pi^*}(s) \geq V^{\pi}(s)$.

Question: can we choose $u(R_1, R_2, \ldots) = \sum_i R_i$?

Expected linear aditive utility

Def: Expected linear aditive utility

Function $u(R_t, R_{t+1}, \ldots) = \mathbb{E}\left[\sum_{t'=t}^{|D|} \gamma^{t'} R_{t'}\right]$ is expected linear additive utility

- $\gamma \in (0, 1]$ is a discount factor, makes agent prefer earlier rewards.
- Risk-neutral
- For infinite D and bounded rewards, $\gamma < 1$ gives convergence (why?)
- Implies existence of optimal solution

Optimality principle

When using expected linear additive utility, "MDP" has an optimal deterministic Markovian policy π^*.

Thm: The optimality principle for infinite-horizon MDPs

Infinite horizon MDP with $V^\pi(s_t) = \mathbb{E}\left[\sum_{t'=0}^{\infty} \gamma^{t'} R^\pi_{t+t'}\right]$ and $\gamma \in [0,1)$. Then there exists optimal value function V^*, is stationary, Markovian, and satisfies for all s:

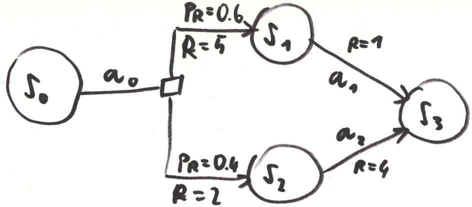
$$V^*(s) = \max_\pi V^\pi(s)$$

$$V^*(s) = \max_{a \in A} \left[\sum_{s' \in S} T(s,a,s')[R(s,a,s') + \gamma V*(s')] \right]$$

$$\pi^*(s) = \arg\max_{a \in A} \left[\sum_{s' \in S} T(s,a,s')[R(s,a,s') + \gamma V*(s')] \right]$$

Finding MDP solutions

- S: S_0, S_1, S_2, S_3
- A: a_0, a_1, a_2
- T:
$$T(S_0, a_0, S_1) = 0.6$$
$$T(S_0, a_0, S_2) = 0.4$$
$$T(S_1, a_1, S_3) = 1$$
$$T(S_2, a_2, S_3) = 1$$
- R:
$$R(S_0, a_0, S_1) = 5$$
$$R(S_0, a_0, S_2) = 2$$
$$R(S_1, a_1, S_3) = 1$$
$$R(S_2, a_2, S_3) = 4$$



Calculate value function for cyclic MDP

- S: S_0, S_1, S_2, S_3
- A: a_0, a_1, a_2

  $T(S_0, a_0, S_1) = 0.6$

  $T(S_0, a_0, S_2) = 0.4$
- T : $T(S_1, a_1, S_3) = 1$

  $T(S_2, a_2, S_3) = 0.7$

  $T(S_2, a_2, S_0) = 0.3$

  $R(S_0, a_0, S_1) = 5$

  $R(S_0, a_0, S_2) = 2$
- R : $R(S_1, a_1, S_3) = 1$

  $R(S_2, a_2, S_3) = 4$

  $R(S_2, a_2, S_0) = 3$