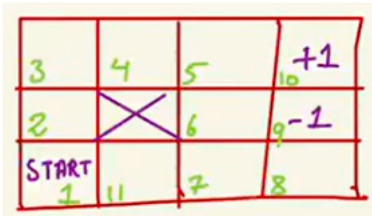
CS3243: Introduction to Artificial Intelligence
Lecture Notes 8: Markov Decision Process (MDP)

1. Probabilistic Transition Models

Let us revisit the case of a Mopbot. In Lecture 2, we learnt that we can model the behavior of the agent with a:

- Set of **states** S that the agent can take. The diagram below shows the layout of a room, and the position of the Mopbot S_i can represent the state that the Mopbot is in.



- Set of **actions** A that the agent can make.

- The **transition model** $T(s, a)$, if it was deterministic, takes in an initial state s and an action a, and returns the next state that the agent is in after taking action a from initial state s. For example,

$$T(S_1, Up) = S_2.$$

Now, we shall extend the definition of a transition model to make the agent's behavior **stochastic** instead of deterministic. $T(s, a)$ would become a probability distribution over the states that the agent will transition to upon taking an action a in state s.

Let's program the Mopbot to move to the intended state with probability 0.7, or move to other states with probability 0.1 each. When in S_1, if the agent takes the action Up, it can land in S_1, S_2 or S_{11}. The transition model would then instead be

$$T(S_1, Up) = \begin{cases} 0.7, & S_2 \\ 0.1, & S_{11} \\ 0.2, & S_1 \end{cases}$$

- The **reward function** $R: States \rightarrow \mathbb{R}$ maps every state to a real number. An equivalent model of the reward function is $R': States \times Action \rightarrow \mathbb{R}$. We define it as $R(s) = -0.4$ for $s \in S \backslash \{S_9, S_{10}\}$, and $R(S_{10}) = +1$, $R(S_{11}) = -1$.

- **Initial state** is defined as S_1.

- **Goal node** is going to be flexible.

- **Terminal states:** In such states, no action is taken after you reach these states. We define the set of terminal states to be $\{S_9, S_{10}\}$.

Scribe: Joshua Chew Jian Xiang

2. Plan, Policy and Utility

A **plan** refers to a sequence of actions. To travel from initial state S_1 to the state S_{10}, the plan is to take the sequence of actions $[Up, Up, Right, Right, Right]$ or to follow the sequence of states $[S_2, S_3, S_4, S_5, S_{10}]$. The probability of reaching S_{10} from S_1 with the above plan, using the probabilistic transition model given earlier, would be $(0.7)^5 < 0.2$.

A **policy** is a function $\pi: States \rightarrow Actions$ which tells us in whatever state we are in, what action we should take. Following a policy does not give you a deterministic sequence of states.

With a cost function, we would try to minimize cost. With a reward function, we would like to maximize rewards. With a given path, we would have to lift the notion of reward.

Utility is defined with respect to a *sequence of states*. The utility of a path,

$$\begin{aligned} &U_n([S_0, S_1, S_2, \cdots, S_n]) \\ &\qquad = R(S_0) + R(S_1) + R(S_2) + \cdots \text{ (additive notion)} \\ &\qquad = R(S_0) + \gamma R(S_1) + \gamma^2 R(S_2) + \cdots \text{ where } \gamma \in [0,1) \text{ (discounted notion)} \end{aligned}$$

Applying the geometric series, we can show that if there is a maximum reward R_{max} such that for all S_i, $R(S_i) \leq R_{max}$, then for any path,

$$U_n([S_{i_1}, S_{i_2}, \cdots]) \leq \frac{R_{max}}{1-\gamma}$$

Let S_i be a random variable that refers to the state reached at time i. Suppose that we follow the following policy:

$$\pi(S_1) = \pi(S_7) = \pi(S_6) = \pi(S_8) = \pi(S_2) = Up$$
$$\pi(S_3) = \pi(S_{11}) = \pi(S_4) = \pi(S_5) = Right$$

We might observe many different sequences of states. Some possible sequences include $[S_1, S_2, S_3, S_4, S_5, S_{10}]$ and $[S_1, S_{11}, S_7, S_8, S_7, S_6, S_5, S_4, S_5, S_6, S_5, S_{10}]$.

The utility for a state s for policy π,

$$U^{\pi}(s) = E[U_h(\tau)] = E[\sum_{t=0}^{\infty} \gamma^t R(S_t)]$$

where τ refers to the sequence that we observe, and S_t is a random variable.

The optimal policy, as a result, is independent of the start state, i.e.

$$\pi^*(s) = (\underset{\pi}{argmax}\, U^{\pi}(s))(s)$$

This is why we have defined the notion of policy as from $States \rightarrow Action$ instead of $Sequence\ of\ States \rightarrow Actions$.

3. Finding an Optimal Policy

If we find ourselves in state s and we want to find out which is the optimal action a to take, we look at all the available actions and compare them, such that

$$\pi^*(s) = \underset{a \in A(s)}{argmax}\, P(s' \mid s, a)\;\; U^{\pi^*}(s')$$

Moreover, since the optimal policy is independent of the start state, we get

$$U(s) = U^{\pi^*}(s')$$

$$\pi^*(s) = \underset{a \in A(s)}{argmax}\, P(s' \mid s, a)\, U(s)$$

The **Bellman Equation** is as follows:

$$U(s) = R(s) + \gamma \max_{a \in A(s)} (\sum_{s'} P(s' \mid s, a)\, U(s'))$$

where $A(s)$ is the set of possible actions that can be taken at state s.

Derivation:

$$U(s) = \mathrm{E}[\sum_{t=0}^{\infty} \gamma^t R(S_t)]$$

$$= E[R(S_0) + \sum_{t=1}^{\infty} \gamma^t R(S_t)]$$

$$= R(s) + E[\sum_{t=1}^{\infty} \gamma^t R(S_t \mid S_0 = s)]$$

$$= R(s) + \sum P(s' \mid s, \pi^*(s))\, \gamma(R(s') + E[\sum_{t=2}^{\infty} \gamma^{t-1}(R(s_t \mid s_1 = s'))])$$

$$= R(s) + \sum P(s' \mid s, \pi^*(s))\, \gamma(R(s') + E[\sum_{t'=1}^{\infty} \gamma^{t'}(R(s_{t'} \mid s_0 = s'))])$$

$$= R(s) + \gamma \sum P(s' \mid s, \pi^*(s))\, U(s')$$

$$= R(s) + \gamma \max_{a \in A(s)} (\sum_{s'} P(s' \mid s, a)\, U(s'))$$