CS3243: Introduction to Artificial Intelligence **Fall 2020**

Lecture 8: October 14

Lecturer: Prof. Kuldeep S. Meel *Scribe: Ang Zheng Yong*

8.1 Markov Decision Processes

Back in Lecture 2, we defined a problem with:

1. States: S
2. Actions: A
3. Transition model: $T(s, a) : State \times Action \rightarrow State$
4. Performance measure: $h : State \times Action \rightarrow \mathbb{R}$
5. Initial state: s_0
6. Goal state: s_g

Today, we will consider a stochastic model, and we will redefine some of the aforementioned parameters to fit into our stochastic model.

8.1.1 Transition model

Recap: the transition model reflects the state that the agent will transition to from state s after executing the action a. For instance, given the following state space:

s_3	s_4	s_5	s_{10}
s_2	$\times$	s_6	s_9
s_1	s_{11}	s_7	s_8

If the transition model is $T(s_1, UP)$, the agent will get to the state s_2.

Now, let us consider a model where the agent moves in the direction that we want most of the time, and moves in another direction with a small probability. For example, if we request the agent to go "UP", it will move:

- up with a probability of 0.7,
- down, left, or right, each with a probability of 0.1.

The previous transition model is unable to encapsulate the idea of stochasticity, so we will need to extend the definition of the transition model to include the stochasticity of the model. We revise $T(s, a)$ (also written as $P(s'|s,a))$ to represent the probability distribution over the states that the agent will transition to, upon performing action a in state s. Thus, for the above example:

$$T(s_1, \mathit{UP}) = P(s'|s_1, \mathit{UP}) = \begin{cases} 0.7 & \text{transitioning to } s_2 \\ 0.2 & \text{transitioning to } s_1 \\ 0.1 & \text{transitioning to } s_{11} \end{cases}$$

How effective is this model? If nondeterministic elements exist in a problem, this transition model would be able to accurately model the problem.

8.1.2 Reward function

The performance measure which we have previously introduced will be replaced by a reward function, which serves a similar purpose.

The reward function R can be expressed as $R : \mathit{States} \to \mathbb{R}$ or $R : \mathit{States} \times \mathit{Actions} \to \mathbb{R}$ (both expressions are equivalent to each other and can be used interchangeably). However, we will be continue our discussion with the model $R : \mathit{States} \to \mathbb{R}$ for the sake of being consistent with AIMA.

8.1.3 Goal state

Similar to previous lectures, we also aim to find a path from the initial state to the goal state. However, in today's discussion, we will not consider the goal to be a permanent entity; instead, we will consider the goal to be flexible.

8.1.4 Terminal states

In addition to the six parameters in our orignal problem definition, we will also introduce the idea of a terminal state. Terminal states are states where no further actions are taken after we reach them.

8.1.5 Example

Let us consider the example which we have previously mentioned.

s_3	s_4	s_5	s_{10}
s_2	×	s_6	s_9
s_1	s_{11}	s_7	s_8

We will consider the set of states to be s_i, where an × indicates an unreachable state (i.e. a wall), and $\{s_9, s_{10}\}$ to be the set of terminal states. We will also assign rewards to each state:

- $R(s_9) = -1$,
- $R(s_{10}) = +1$,
- $\forall s_i \in \mathit{States} \backslash \{s_9, s_{10}\}, R(s_i) = -0.4$.

If the model was deterministic, a possible plan (i.e. sequence of actions) might have been:

$$\text{UP}, \text{UP}, \text{RIGHT}, \text{RIGHT}, \text{RIGHT}$$

and the states that we would've visited would be $s_1 \to s_2 \to s_3 \to s_4 \to s_5 \to s_{10}$.

Unfortunately, if the model was probabilistic/stochastic, our plan may not always work. Assuming a probabilistic model (with the probabilities listed in Section 8.1.1), the probability of our plan successfully bringing the agent to s_{10} is merely $0.7^5 = 0.16807 < 0.2$. Let us take a closer look at one possible circumstance under this probabilistic model.

- Initially, when we take the action "UP" in s_1, we could have landed in either s_2, s_1, or s_{11}, with probabilities 0.7, 0.2, and 0.1 respectively.
- If our agent (is extremely unlucky and) lands in s_{11} after executing the first action, it will continue to execute the action "UP" on the next iteration if it followed our plan, despite the cell above s_{11} being a wall.

Thus, our analysis shows that our plan may not always work for a stochastic model. Instead, we will require a more general idea, which is more adaptive than a plan. For instance, we could use a **policy**, i.e. a function which returns the action to be taken in every single state which we could be in. A policy is modelled as

$$\pi : \textit{States} \to \textit{Actions}$$

8.2 Utility

How do we put the reward function into our calculations? We would like our agent to maximize the amount of rewards gained, but how do we show this?

Assuming that our agent performs a sequence of actions, and the actions gives rise to a path (i.e. sequence of states) $[s_0, s_1, \ldots, s_n]$. We can define the utility of the sequence of states as:

$$U_h([s_0, s_1, \ldots, s_n]) = R(s_0) + R(s_1) + R(s_2) + \ldots + R(s_n) \tag{8.1}$$

$$\text{or } U_h([s_0, s_1, \ldots, s_n]) = R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \ldots + \gamma^n R(s_n) \qquad \text{where } \gamma \in [0, 1) \tag{8.2}$$

Equation 8.1 is known as the **additive model**, whereas equation 8.2 is known as the **discounted model**, which reduces (or discounts) the magnitude of future rewards. These are just 2 ways of representing the utility, and there are many other ways to model the utility. Note that the utility function is defined over a sequence of states instead of just a single state ($U_h([s_i]) = R(s_i)$, and it disregards the probability distribution of the model).

Why might we need to discount future rewards? We may want our agent to favor immediate rewards over future rewards, and it would also help us to navigate infinite state spaces. The state space would be infinitely large if there are no terminal states.

- If we use the discounted model, then $\forall s_i, R(s_i) \leq R_{max}, U_h([s_i, s_{i+1}, \ldots]) = \sum_i \gamma^i R(s_i) \leq \frac{R_{max}}{1-\gamma}$. This means that the maximum utility is upper bounded by a value $\frac{R_{max}}{1-\gamma}$ in an infinite state space.

8.3 Optimal Policies

Before discussing the idea of an optimal policy, we will need to first define the following:

- s_i represents the i^{th} state in the state space, whereas
- S_i represents the state reached at time i; i.e. it refers to a *random variable.*

For instance, given the following state space:

s_3	s_4	s_5	s_{10}
s_2	$\times$	s_6	s_9
s_1	s_{11}	s_7	s_8

we might have the following policy:

$$\pi(s_i) = \begin{cases} \text{UP} & \text{if } i \in \{1, 2, 6, 7, 8\} \\ \text{RIGHT} & \text{if } i \in \{3, 4, 5, 11\} \end{cases}$$

What are some possible sequences of states which we might observe? In fact, there are (infinitely) many possible combinations, such as:

$$\begin{gathered} [s_1, s_2, s_3, s_4, s_5, s_{10}] \\ [s_1, s_{11}, s_7, s_6, s_5, s_{10}] \\ [s_1, s_2, s_4, s_5, s_6, s_7, s_{11}, s_1, \ldots, s_{10}] \\ \vdots \end{gathered}$$

As such, we might want to define the utility of a state s for a policy π, instead of defining the utility over a sequence of states. The utility of a state s for a policy π is given by:

$$\begin{aligned} U^\pi(s) &= E[U_h(\tau)] && \text{where } \tau \text{ is the seq. of states observed by policy } \pi \\ &= E\left[\sum_{t=0}^{\infty} \gamma^t R(S_t)\right] && \text{summation to } \infty \text{ indicates an infinite horizon} \end{aligned}$$

With this, we can define the **optimal policy** as:

$$\pi^* = \arg\max_\pi \{U^\pi(\mathit{State})\}$$

and thus the action executed by an agent adopting the optimal policy at state s is

$$\pi^*(s) = \left(\arg\max_\pi \{U^\pi(s)\}\right)(s)$$

Note that the optimal policy is *independent of the start state.* The above equation does not care about where the agent originally started; it only looks at the utilities of all the policies from the *current state s*, finds out the policy that maximizes the utility, and then uses the corresponding action that should be taken.

This is why we could afford to define policies mapping from states to actions, instead of a *sequence of* states to actions. When we are looking at the discounted notion of utility under an infinite horizon, a policy would only depend on the current state which we are in.

8.3.1 Finding the optimal policy

How do we decide what action to take at state s following the optimal policy?

- We first look at all the states s' that are reachable from state s, i.e. $P(s'|s,a)$.
- Then, for all states s' that are reachable, we would want to compute the utility that we would get if we had followed the optimal policy π^* from there onwards, i.e. $U^{\pi^*}(s')$.
- Thus, the action that we should that is the action that maximizes the expected utility, i.e.

$$\arg\max_{a \in A(s)} \left\{ P(s'|s,a) U^{\pi^*}(s') \right\}$$

Since the optimal policy does not depend on the initial state, we can also define $U(s) = U^{\pi^*}(s)$. Then, we could rewrite our optimal policy as:

$$\pi^*(s) = \arg\max_{a \in A(s)} \{ P(s'|s,a) U(s') \}$$

8.3.2 Bellman equation

Remember that previously, we have discussed that

$$\begin{aligned} U(s) &= E[U_h(\tau)] \\ &= E\left[\sum_{t=0}^{\infty} \gamma^t R(S_t)\right] \end{aligned}$$

For $t = 0$, we know that $\gamma^t R(S_t) = R(S_0)$, so

$$\begin{aligned} U(s) &= E\left[\sum_{t=0}^{\infty} \gamma^t R(S_t)\right] \\ &= E\left[R(S_0) + \sum_{t=1}^{\infty} \gamma^t R(S_t)\right] \\ &= R(s) + E\left[\sum_{t=1}^{\infty} \gamma^t R(S_t | S_0 = s)\right] && \text{when } t = 0 \text{, we were in state } s \end{aligned} \tag{8.3}$$

If we try to expand $E\left[\sum_{t=1}^{\infty} \gamma^t R(S_t|S_0 = s)\right]$ by looking at all the states that we might end up in, we would get

$$\begin{aligned}
E\left[\sum_{t=1}^{\infty} \gamma^t R(S_t|S_0 = s)\right] &= \sum_{s'} P(s'|s, \pi^*(s)) \left(\gamma R(s') + E\left[\sum_{t=2}^{\infty} \gamma^t R(S_t|S_1 = s')\right]\right) \\
&= \sum_{s'} P(s'|s, \pi^*(s))\gamma \left(R(s') + E\left[\sum_{t=2}^{\infty} \gamma^{t-1} R(S_t|S_1 = s')\right]\right) \\
&= \sum_{s'} P(s'|s, \pi^*(s))\gamma \left(R(s') + E\left[\sum_{t'=1}^{\infty} \gamma^{t'} R(S'_t|S_0 = s')\right]\right) && \text{by letting } t' = t - 1 \\
&= \sum_{s'} P(s'|s, \pi^*(s))\gamma U(s') && \text{from equation 8.3}
\end{aligned}$$

Therefore,

$$\begin{aligned}
U(s) &= R(s) + E\left[\sum_{t=1}^{\infty} \gamma^t R(S_t|S_0 = s)\right] \\
&= R(s) + \gamma \sum_{s'} P(s'|s, \pi^*(s)) U(s') \\
&= R(s) + \gamma \max_{a \in A(s)} \left\{\sum_{s'} P(s'|s, a) U(s')\right\} && \text{from how } a = \pi^*(s) \text{ is selected} \qquad (8.4)
\end{aligned}$$

Equation 8.4 is also known as the **Bellman equation**.