Week 2

Tuesday, August 27, 2019 9:14 AM

Markov Decision Processes [MDP's]

Markov decision processes [MDP's] is a theoretical formalization of sequential decision making problems. It provides a framework by which we can theoretically describe sequential decision making problems.

The need of this framework arises from the limitations of the one used to describe bandit problems. The bandit problem idea can't be used if :-

1. The action to be taken varies across time steps.
2. The agent needs to adjust its behavior as the state changes.

The above characteristics are very often present in real world problems and so the need of a more robust and flexible framework arises.

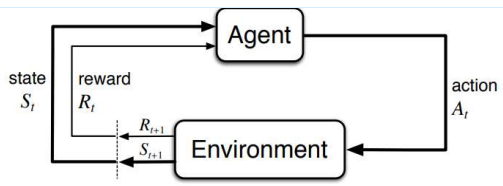
Mathematically an MDP problem can be described using a function of the form $p: \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{R} \rightarrow [0, 1]$

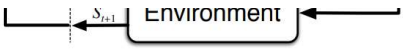
or $p(s', r \mid s, a) = Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$

This can be described as follows :-

→ time $t-1$

An agent initially in state s takes an action a. It then enters the state s' and receives a reward r.





The eqn above describes the probability that agent will enter a new state s' and receive a reward r given initially it's in state s and takes an action a. Since it's a probability

$$\sum_{s' \in S} \sum_{r \in R} p(s', r \mid s, a) = 1 \quad \forall s \in S, a \in A(s)$$

In a finite MDP process, there are finite states, rewards and actions.

In a Markov process, the probabilities of each possible value of S_t and R_t depend only on the values of the immediately preceding values of states and actions. If a state or action prior to the immediately preceding state or action affects future values of states or actions, then this information must be captured in the immediately preceding state and action values.

The MDP framework is abstract and flexible and can be applied to a wide array of problems.

The goal in reinforcement Learning

The main goal in RL is to maximize the reward over the lifetime of an agent.

We can think about this reward in many ways. We can think about it in terms of maximizing the immediate reward but that doesn't work because in the case where reward is high immediately, over the lifetime, it may fall.

Formally by total reward, we are referring to the return after time step t. We denote return as :-

$$G_t = R_{t+1} + R_{t+2} + \cdots\cdots$$

We often refer to the expected return value due to the presence of randomness in the environment. It is denoted as :-

$$E(G_t) = E[R_{t+1} + R_{t+2} + \cdots\cdots R_t]$$

(R_t : Final time step)

Note in the above given expected return value, we have a final time step. This will create tasks which end after some time. Such tasks are called episodic tasks. The various executions of these tasks break up into different episodes. Each episode is independent of the previous execution of an episode.

Episodic Tasks	Continuing Tasks
• Interaction breaks naturally into episodes	• Interaction goes on continually
• Each episode ends in a terminal state	• No terminal state
• Episodes are independent	
$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \ldots + R_T$	$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \ldots = \infty ?$

As, we can see above, the expected return for continuous tasks may consist of an infinite number of terms. So how to calculate it ?

To do so, we use the concept of discounting. After discounting we get the foll. :- for the expected return value :-

$$E(G_t) \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots\cdots$$

$$E(G_t) \doteq E \sum_{k}^{\infty} \gamma^k R_{t+k+1} \quad \text{with} \quad 0 \le \gamma \le 1$$

$$E(G_t) \doteq E \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad \text{with} \quad 0 \leq \gamma \leq 1$$

This is always going to be a finite sequence as shown below :-

$$E(G_t) = E \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Assume R_{max} is the maximum expected reward, we get

$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \leq \sum_{k=0}^{\infty} \gamma^k R_{max}$$

Since R_{max} is a constant, we get

$$\leq R_{max} \sum_{k=0}^{\infty} \gamma^k$$

→ A geometric series $= \frac{1}{1-\gamma}$

$$\leq R_{max} \times \frac{1}{1-\gamma}$$

This gives a method to find the expected reward for continuous tasks.

Note ① If $\gamma = 0$, we get

$G_t = R_{t+1}$

ie the agent becomes a greedy agent ← Short sighted

② If $\gamma = 1$

the agent is more far-sighted.

the agent is more far sighted.

One of the most important properties of G_t is that it can be written recursively.

$$G_t = R_{t+1} + \gamma \underbrace{\left(R_{t+2} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \right)}_{G_{t+1}}$$

$$G_t = R_{t+1} + \gamma G_{t+1}$$