BASICS OF REINFORCEMENT LEARNING

REPORT

Ameya Deshmukh

ABSTRACT

Reinforcement learning is the process of an agent defining a policy of actions that maximizes the cumulative reward it receives from its environment in the future. Unlike supervised learning, the agent does not have access to certain states with the corresponding optimal actions, rather it has to explore the environment as part of the learning process.

1 Definitions

- *Reward*- A scalar signal from the environment, which indicates the quality of the action on the agent given the current state.

  $$\text{Reward at the } t^{\text{th}} \text{ time step as a result of the action } A_t \text{ in the state } S_t = R_t$$

- *State*- The current information about the environment that is available to the learning agent. We assume a discrete set of states.
- *Policy*- A generally stochastic mapping of actions to be taken given the perceived state of the agent.

  $$\pi(a|s) = \mathbb{P}(\text{Next action} = a|\text{ Current state} = s)$$

- *Value*- This is a mapping from perceived states to a measure of the expected **total** reward in the future. An ideal agent thus seeks actions which lead to states with good value and not just those which maximize the next reward.

  $$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \ldots |S_t = s]$$

- *Model*- This is the stochastic process which the environment uses to reward the agent and to evolve. Here S_t is the environment's state and not the perceived state of the agent.

2 Multi-arm Bandits

3 Markov Decision Processes

These are processes in which we have a completely observable environment model in which the current state renders the history of the environment redundant. Hence we can drop the distinction between the perceived state of the agent and the state of the environment, and say the following about the model:

$$\mathbb{P}(S_{t+1}|S_t, S_{t-1}, \ldots, S_1) = \mathbb{P}(S_{t+1}|S_t)$$

We first deal with a simpler variant of MDP's, in which we remove the 'decision' part by scrapping actions.

Markov Reward Processes

Here we can characterize an MDP by the $\mathcal{P}_{ss'} = \mathbb{P}(S_{t+1} = s'|S_t = s)$ values. We also have the $\mathcal{R}_s = \mathbb{E}(R_{t+1}|S_t = s)$ values, which can be thought of as the average reward of leaving a particular state. With these we can also observe the following about the values:

$$\begin{aligned} v(s) &= \mathbb{E}(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \ldots | S_t = s) \\ &= \mathbb{E}(R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} \ldots) | S_t = s) \\ &= \mathbb{E}(R_{t+1} + \gamma (v(S_{t+1})) | S_t = s) \\ &= \mathcal{R}_s + \gamma \sum_{s'} \mathcal{P}_{ss'} v(s') \end{aligned}$$

Introducing actions

Now we have the following quantities defined given a policy π:

$$\begin{aligned} \mathcal{P}^a_{ss'} &= \mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a) \\ \mathcal{R}^a_s &= \mathbb{E}(R_{t+1} | S_t = s, A_t = a) \\ \mathcal{P}^\pi_{ss'} &= \sum_a \pi(a|s) \mathcal{P}^a_{ss'} = \text{ The probability to get to } s' \text{ through any action} \\ \mathcal{R}^\pi_s &= \sum_a \pi(a|s) \mathcal{R}^a_s = \text{ The expectation of the reward on leaving } s \text{ through any action} \\ v_\pi(s) &= \mathbb{E}_\pi(G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+1} \ldots | S_t = s) = \text{The state value function} \\ q_\pi(s,a) &= \mathbb{E}_\pi(G_t | S_t = s, A_t = a) = \text{The action value function} \end{aligned}$$

Applying the same recursive idea as in MRP's

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi(R_{t+1} | S_t = s) + \gamma \mathbb{E}_\pi(R_{t+2} + \gamma R_{t+3} \ldots | S_t = s) \\ &= \sum_a \pi(a|s) \mathbb{E}(R_{t+1} | S_t = s, A_t = a) + \gamma \sum_{s'} \mathcal{P}^\pi_{ss'} \mathbb{E}_\pi(G_{t+1} | S_{t+1} = s') \\ &= \mathcal{R}^\pi_s + \gamma \sum_{s'} \mathcal{P}^\pi_{ss'} v_\pi(s') \end{aligned}$$

Another manipulation for the expression is simply:

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a) \longrightarrow \text{Bellman Equation 1}$$

Similarly for the action value function:

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}(R_{t+1} | S_t = s, A_t = a) + \gamma \mathbb{E}_\pi(G_{t+1} | S_t = s, A_t = a) \\ &= \mathcal{R}^a_s + \gamma \sum_{s'} \mathcal{P}^a_{ss'} \mathbb{E}_\pi(G_{t+1} | S_{t+1} = s') \\ &= \mathcal{R}^a_s + \gamma \sum_{s'} \mathcal{P}^a_{ss'} v_\pi(s') \longrightarrow \text{Bellman Equation 2} \end{aligned}$$

These 2 interdependent equations can easily be combined to get equations involving only one of the value functions.