Lecture 2: RL and Markov decision processes

ZHANG Mofan

February 2021

0 Resources

UCL course on RL: `https://www.davidsilver.uk/teaching/`
Chinese blogs: `https://www.cnblogs.com/pinard/category/1254674.html`

1 Notation

$\mathcal{P}_{ss'}^{a} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$	state transition probability
$\mathcal{P}$	state transition probability matrix
$\mathcal{S}$	finite set of states
$\mathcal{A}$	finite set of actions
$\mathcal{R}$	reward function, $\mathcal{R}_s^a = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$
γ	discount factor $\in [0,1]$
$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$	total discounted reward from time step t
$v(s) = \mathbb{E}[G_t \mid S_t = s]$	state value function

2 Bellman Equation for Markov reward processes

2.1 Decomposition of state value function

- immediate reward R_{t+1}
- discounted value of successor state $\gamma v(S_{t+1})$

$$\begin{aligned} v(s) &= \mathbb{E}[G_t \mid S_t = s] \\ &= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... \mid S_t = s] \\ &= \mathbb{E}[R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + ...) \mid S_t = s] \\ &= \mathbb{E}[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \mathbb{E}[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s] \end{aligned} \tag{1}$$

Written in another form:

$$v(s) = \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} v(s') \tag{2}$$

2.2 Matrix formulation

$$v = \mathcal{R} + \gamma \mathcal{P} v \tag{3}$$

Solution:

$$v = (\mathcal{I} - \gamma \mathcal{P})^{-1} \mathcal{R} \tag{4}$$

3 Policies in Markov decision processes

A policy π is a distribution over actions given states,

3.1 From MRP to MDP

$$\pi(a|s) = \mathbb{P}[A_t = a | S_t = s] \tag{5}$$

Definition:

$$\mathcal{P}^{\pi}_{s,s'} = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}^{a}_{ss'} \tag{6}$$

$$\mathcal{R}^{\pi}_{s} = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{R}^{a}_{s} \tag{7}$$

The state value function under policy π:

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi[G_t | S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s] \end{aligned} \tag{8}$$

The action value function under policy π:

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}_\pi[G_t | S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a] \end{aligned} \tag{9}$$

The relation between $v_\pi(s)$ and $q_\pi(s,a)$ is:

$$\begin{aligned} v_\pi(s) &= \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s,a) \\ &= \sum_{a \in \mathcal{A}} \pi(a|s) (\mathcal{R}^{a}_{s} + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^{a}_{ss'} v_\pi(s')) \end{aligned} \tag{10}$$

$$\begin{aligned} q_\pi(s,a) &= \mathcal{R}^{a}_{s} + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^{a}_{ss'} v_\pi(s') \\ &= \mathcal{R}^{a}_{s} + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^{a}_{ss'} \sum_{a' \in \mathcal{A}} \pi(a'|s') q_\pi(s', a') \end{aligned} \tag{11}$$

3.2 Matrix form

$$v_\pi = \mathcal{R}^{\pi} + \gamma \mathcal{P}^{\pi} v_\pi \tag{12}$$

Direct solution:

$$v_\pi = (\mathcal{I} - \gamma \mathcal{P}^{\pi})^{-1} \mathcal{R}^{\pi} \tag{13}$$

4 Optimal value

4.1 Optimal value functions

Optimal state value function:

$$v_*(s) = \max_{\pi} v_\pi(s) \tag{14}$$

Optimal action value function:

$$q_*(s,a) = \max_{\pi} q_\pi(s,a) \tag{15}$$

4.2 Optimal policy

Define a partial ordering over policies:

$$\pi \geq \pi' \qquad \text{if} \quad v_\pi(s) \geq v_{\pi'}(s) \quad \forall s \tag{16}$$

Theorem: For any MDP,

1. There exists an optimal policy π_* that is better than or equal to all other policies, $\pi_* \geq \pi, \forall \pi$;
2. All optimal policies achieve the optimal state value function, $v_{\pi_*}(s) = v_*(s)$;
3. All optimal policies achieve the optimal action value function, $q_{\pi_*}(s,a) = q_*(s,a)$.

In practice, an optimal policy can be found by maximising over $q_*(s,a)$,

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a \in \mathcal{A}} q_*(s,a) \\ 0 & \text{otherwise} \end{cases} \tag{17}$$