RL

Dimtirov Vladimir

10 февраля 2023 г.

Содержание

Markov decision process

MDP is tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$, where:

- $\mathcal{S}$ - set of states of the world
- $\mathcal{A}$ - set of actions
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \mapsto \triangle(\mathcal{S})$ - state-transition function, giving us $p\left(s_{t+1} \mid s_t, a_t\right)$
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ - reward function, given us $\mathbb{E}_R\left[R\left(s_t, a_t\right) \mid s_t, a_t\right]$

Reward hypothesis (R.Sutton)

> That all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)

Cumulative rewards is called a return:

$G_t \triangleq R_t + R_{t+1} + R_{t+2} + \ldots + R_T$

There are **2 problems** in the design of the system:

1. Infinite sum -> solve: discounting coeeficient
2. Simple solve -> watch the pic 1

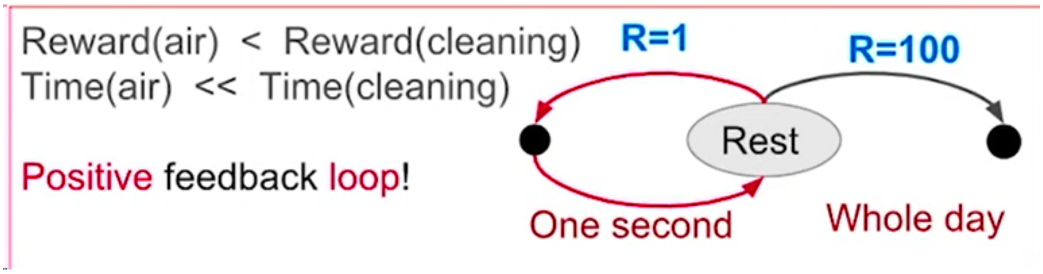

Рис. 1: Simple solution

Solving the infinite sum problem: get discounting coefficient $(0 \leq \gamma < 1)$ and cumulative rewards will take the form:

$$\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

This method has its pluses:

- Human likeness
- Mathematical convenience
- fast optimization

Notes: multiplying by γ changes the task and it's solution!

Take away 1: reward only for what, but never for how

Take away 2: do not subtract mean from rewards

Transformation politic for reward (ML advice):

- Rewards scaling - division by positive constant

- Reward shaping - we could add to all rewards in MDP values of potential-based shaping function

Expected objective

Optimal policy maximizes expected return:

$$
\begin{aligned}
\mathbb{E}\left[G_0\right] &= \mathbb{E}\left[R_0 + \gamma R_1 + \ldots + \gamma^T R_T\right] \\
&= \mathbb{E}_{E,\pi_\theta}\left[G_0\right] \\
&= \mathbb{E}_{\pi_\theta}\left[G_0\right] \\
&= \mathbb{E}\left[G_0 \mid \pi_\theta\right] \\
&= \mathbb{E}_{s_{0:T}}\left[G_0\right] \\
&= \mathbb{E}_{a_{0:T}}\left[\mathbb{E}_{a_0|s_0}\left[R_0 + \mathbb{E}_{s_1|s_0,a_0}\left[\mathbb{E}_{a_1|s_1}\left[\gamma R_1 + \ldots\right]\right]\right]\right] \\
&= \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim p_\theta}\left[\gamma^t R_t\right]
\end{aligned}
$$

State value function

" We want to know value function not only from 0 point in time "

Is the expected return conditional on state:

$$
\begin{aligned}
v_\pi(s) &\triangleq \mathbb{E}_\pi\left[G_t \mid S_t=s\right] \\
&=\mathbb{E}_\pi\left[R_t+\gamma G_{t+1} \mid S_t=s\right] \\
&=\sum_a \pi(a \mid s) \sum_{r, s^{\prime}} p\left(r, s^{\prime} \mid s, a\right)\left[r+\gamma \mathbb{E}_\pi\left[G_{t+1} \mid S_{t+1}=s^{\prime}\right]\right] \\
&=\sum_a \pi(a \mid s) \sum_{r, s^{\prime}} p\left(r, s^{\prime} \mid s, a\right)\left[r+\gamma v_\pi\left(s^{\prime}\right)\right]
\end{aligned}
$$

Intuition: mean value of following policy π from state s

In Russian: 'внешнее ожидание по политики и внутренее мат ожидание по среде'

Action value function

Is expected return conditional on state and action:

$$
\begin{aligned}
q_\pi\left(s_t a\right) &=\mathbb{E}_\pi\left[G_t \mid S_t=s, A_t=a\right] \\
&=\mathbb{E}_\pi\left[R_t+\gamma G_{t+1} \mid S_t=s, A_t=a\right] \\
&=\sum_{r, s^{\prime}} p\left(r, s^{\prime} \mid s, a\right)\left[r+\gamma \mathbb{E}_\pi\left[G_{t+1} \mid S_{t+1}=s^{\prime}\right]\right] \\
&=\sum_{r, s^{\prime}} p\left(r, s^{\prime} \mid s, a\right)\left[r+\gamma v_\pi\left(s^{\prime}\right)\right]
\end{aligned}
$$

Intuition: value of following policy π after committing action a in state s

In Russian: 'Какую ценность имеет политика π, если я в состоянии s

сделаю действие a '

Expression value function through state function:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{r,s'} p\left(r, s' \mid s, a\right)\left[r + \gamma v_\pi\left(s'\right)\right]$$
$$= \sum_a \pi(a \mid s) q_\pi(s, a)$$

Bellman expectation equation

$v(s)$ is calculated:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{r,s'} p\left(r, s' \mid s, a\right)\left[r + \gamma v_\pi\left(s'\right)\right]$$
$$= \mathbb{E}_\pi\left[R_t + \gamma v_\pi\left(S_{t+1}\right) \mid S_t = s\right]$$

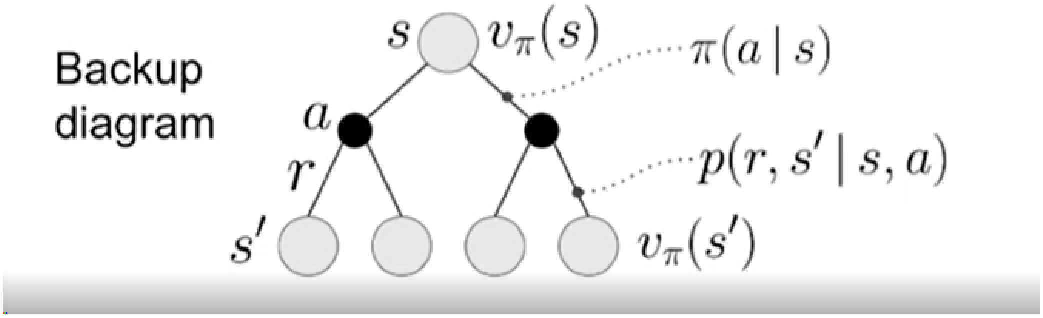

Рис. 2: Backup diagram for $v(s)$

$q(s)$ is calculated:

$$q_\pi(s, a) = \sum_{r,s'} p\left(r, s' \mid s, a\right)\left[r + \gamma v_\pi\left(s'\right)\right]$$
$$= \sum_{r,s'} p\left(r, s' \mid s, a\right)\left[r + \gamma \sum_{a'} \pi\left(a' \mid s'\right) q_\pi\left(s', a'\right)\right]$$

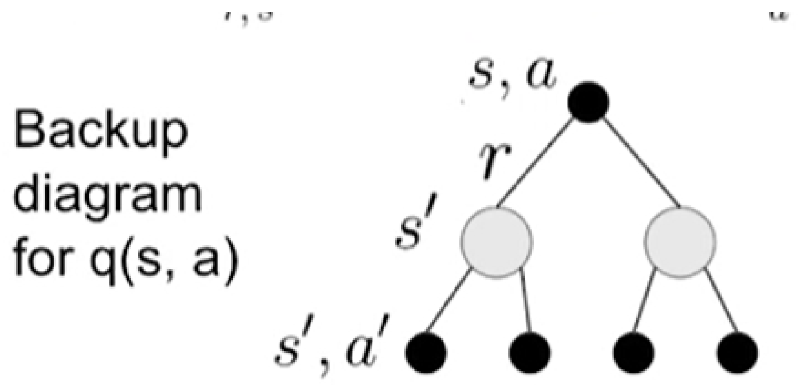

Рис. 3: Backup diagram for $q(s)$

Bellman optimality equation

We could compare policies on the basis of $v(s)$

$$\pi \geq \pi' \quad \Leftrightarrow \quad v_\pi(s) \geq v_{\pi'}(s) \quad \forall s$$

$$v_*(s) = \max_\pi v_\pi(s)$$

$$q_*(s,a) = \max_\pi q_\pi(s,a)$$

Notes: in any finite MDP there is always at least one deterministic optimal policy

$$v_*(s) = \max_a \sum_{r,s'} p\left(r, s' \mid s, a\right)\left[r + \gamma v_*\left(s'\right)\right]$$

$$= \max_a \mathbb{E}\left[R_t + \gamma v_*\left(S_{t+1}\right) \mid S_t = s, A_t = a\right]$$

$$q_*(s,a) = \mathbb{E}\left[R_t + \gamma \max_{a'} q_*\left(S_{t+1}, a'\right) \mid S_t = s, A_t = a\right]$$

$$= \sum_{r,s'} p\left(r, s' \mid s, a\right)\left[r + \gamma \max_{a'} q_*\left(s', a'\right)\right]$$

Generalized Policy Iteration

1. Policy Evaluation

2. Policy Improvement

Policy evaluation

Policy evaluation is also a called **prediction problem**.

Predict value function for a particular policy

Bellman expectation equation:

$$\begin{aligned} v_\pi(s) &= \sum_a \pi(a \mid s) \sum_{r,s'} p\left(r, s' \mid s, a\right)\left[r + \gamma v_\pi\left(s'\right)\right] \\ &= \mathbb{E}_\pi\left[R_t + \gamma v_\pi\left(S_{t+1}\right) \mid S_t = s\right] \end{aligned}$$

is basically a system of linear equation

Algorithm:

Input π, the policy to be evaluated
Initialize an array $V(s) = 0$, for all $s \in \mathcal{S}^+$
Repeat
  $\Delta \leftarrow 0$
  For each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $\boxed{V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\big[r + \gamma V(s')\big]}$ — Bellman expectation equation for v(s)
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$ (a small positive number)
Output $V \approx v_\pi$

Рис. 4:

Policy improvement

Idea: once we know what $v(s)$ for a particular policy, we could improve it by acting greedily w.r.t v(s)!

$$\pi'(s) \leftarrow \arg\max_a \overbrace{\sum_{r,s'} p\left(r, s' \mid s, a\right)\left[r + \gamma v_\pi\left(s'\right)\right]}^{q_\pi(s,a)}$$

This procedure is guaranteed to produce a better policy.