Reinforcement Learning

Saturday, May 5, 2018 14:26

Markov Decision Process

(S, A, R, P, γ)

S: state, A: action, R: reward, P: action → reward, γ: discount factor.

Policy π: $\pi^* = \sum_t \gamma^t r$ $\pi^* := \max \sum_t \gamma^t r$

$$\pi^* = \arg\max_{\pi} E\left[\sum_{t\geq 0} \gamma^t r_t | \pi\right]$$

Bellman Equation.

$$Q^*(s,a) = E_{s' \sim \varepsilon}\left[r + \gamma \max_{a'} Q^*(s',a') | s, a\right].$$

Value iteration:

$$Q_{i+1}(s,a) = E\left[r + \gamma \max_{a'} Q_i(s',a') | s, a\right].$$

$$\underbrace{Q(s,a;\theta)}_{\text{approximator}} \approx Q^*(s,a).$$

forward

$$L_i(\theta_i) = E_{s,a \sim p(\cdot)}\left[(y_i - Q(s,a;\theta_i))^2\right].$$

where: $y_i = E_{s' \sim \varepsilon}\left[r + \gamma \max_{a'} Q(s',a';\theta_{i-1}) | s, a\right].$

backward:

$$\nabla_{\theta_i} L_i(\theta_i) = E_{s,a \sim p(\cdot); s' \sim \varepsilon}\left[r + \gamma \max_{a'} Q(s',a';\theta_{i-1}) - Q(s,a;\theta_i)\right.$$
$$\left.\nabla_{\theta_i} Q(s,a;\theta_i)\right]$$

parametrized policies $\Pi = \{\pi_\theta, \theta \in \mathbb{R}^m\}$

for each policy: $J(\theta) = E_{\tau \sim p(\tau;\theta)}[r(\tau)]$

$$= \int_\tau r(\tau) p(\tau;\theta) d\tau$$

$r(\tau)$ is the reward of a trajectory $\tau = (s_0, a_0, r_0, s_1)$

trick $\nabla_\theta p(\tau;\theta) = p_\theta(\tau;\theta) \frac{\nabla_\theta p(\tau;\theta)}{p(\tau;\theta)} = p(\tau;\theta) \cdot \nabla_\theta \log p(\tau;\theta)$

$$\nabla_\theta J(\theta) = \int_\tau (r(\tau) \nabla_\theta \log p(\tau;\theta)) p(\tau;\theta) d\tau.$$
$$= E_{\tau \sim p(\tau;\theta)}[r(\tau) \nabla_\theta \log p(\tau;\theta)].$$ Monte Carlo.

$$p(\tau;\theta) = \prod_{t\geq 0} p(s_{t+1}|s_t, a_t) \pi_\theta(a_t|s_t)$$

Thus: $\log p(\tau;\theta) = \sum_{t\geq 0} \log p(s_{t+1}|s_t, a_t) + \log \pi_\theta(a_t|s_t)$

$$\nabla_\theta \log p(\tau;\theta) = \sum_{t\geq 0} \nabla_\theta \log \pi_\theta(a_t|s_t)$$

$$\nabla_\theta J(\theta) \approx \sum_{t\geq 0} r(\tau) \nabla_\theta \log \pi_\theta(a_t|s_t)$$

baseline:

$$\nabla_\theta J(\theta) \approx \sum_{t\geq 0} \left(\sum_{t'\geq t} \gamma^{t'-t} r_{t'} - b(s_t)\right) \nabla_\theta \log \pi_\theta(a_t|s_t)$$

Q function as baseline.

$$\nabla_\theta J(\theta) \approx \sum_{t\geq 0} (Q^{\pi_\theta}(s_t,a_t) - V^{\pi_\theta}(s_t)) \nabla_\theta \log \pi_\theta(a_t|s_t)$$

Universal approximation theorem.