彪哥带你学强化学习

10.可视化REINFORCE算法

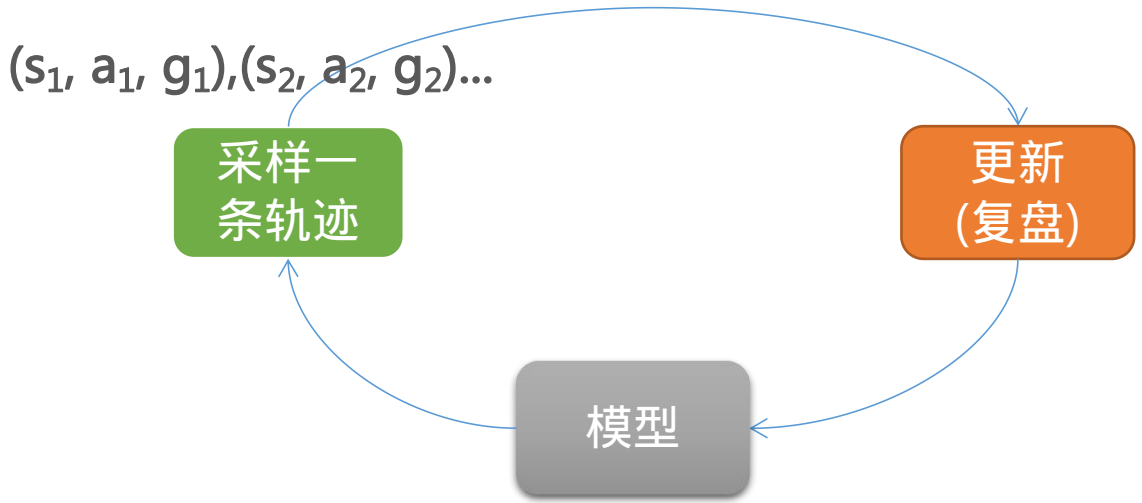
DEEPLY UNDERSTAND REINFORCEMENT LEARNING

讲师：韩路彪

REINFORCE算法

策略学习思路：根据探索得到的回报不断优化策略

REINFORCE：通过蒙特卡洛采样获取回报



REINFORCE算法

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[G_\tau] \qquad J(\theta) = \sum_{\tau \sim \pi_\theta} P(\tau|\theta) G_\tau$$

$$\nabla_\theta J(\theta) = \sum_{\tau \sim \pi_\theta} \nabla_\theta P(\tau|\theta) G_\tau \qquad \nabla_\theta P(\tau|\theta) = \frac{\nabla_\theta P(\tau|\theta)}{P(\tau|\theta)} P(\tau|\theta) = P(\tau|\theta) \nabla_\theta ln P(\tau|\theta)$$

$$\begin{aligned} \nabla_\theta J(\theta) &= \sum_{\tau \sim \pi_\theta} \nabla_\theta P(\tau|\theta) G_\tau \\ &= \sum_{\tau \sim \pi_\theta} P(\tau|\theta) \nabla_\theta ln P(\tau|\theta) G_\tau \\ &= \mathbb{E}_{\tau \sim \pi_\theta}[\nabla_\theta ln P(\tau|\theta) G_\tau] \end{aligned}$$

REINFORCE算法

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta lnP(\tau|\theta)G_\tau]$$

$$P(\tau|\theta) = p(s_0)\prod_{t=0}^{\infty}\pi(a_t|s_t,\theta)p(s_{t+1}|s_t,a_t)$$

$$lnP(\tau|\theta) = p(s_0) + \sum_{t=0}^{\infty}\pi(a_t|s_t,\theta) + \sum_{t=0}^{\infty}p(s_{t+1}|s_t,a_t)$$

$$\nabla_\theta lnP(\tau|\theta) = \sum_{t=0}^{\infty}\nabla_\theta\pi(a_t|s_t,\theta)$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[(\sum_{t=0}^{\infty}\nabla_\theta\pi(a_t|s_t,\theta))G_\tau]$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[(\sum_{t=0}^{\infty}\nabla_\theta\pi(a_t|s_t,\theta)g_t)]$$

$$\theta + \alpha\nabla_\theta J(\theta) \rightarrow \theta_{new}$$