彪哥带你学强化学习

¹¹.可视化REINFORCE算法

DEEPLY UNDERSTAND REINFORCEMENT LEARNING

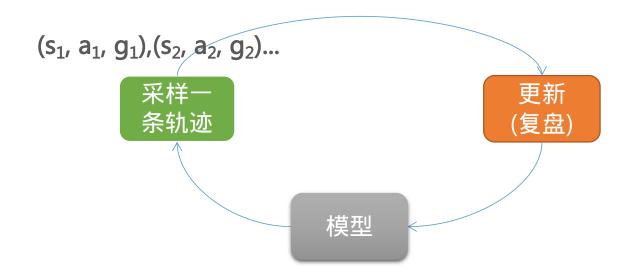
讲师: 韩路彪



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

$$J(heta) = \mathbb{E}_{ au \sim \pi_ heta}[G_{ au}]$$

$$J(heta) = \sum_{ au \sim \pi_ heta} P(au| heta) G_{ au}$$

$$abla_{ heta}J(heta) = \sum_{ au \sim \pi_{ heta}}
abla_{ heta}P(au| heta)G_{ au}$$

$$abla_{ heta}P(au| heta)=rac{
abla_{ heta}P(au| heta)}{P(au| heta)}P(au| heta)=P(au| heta)
abla_{ heta}lnP(au| heta)$$

$$egin{aligned}
abla_{ heta} J(heta) &= \sum_{ au \sim \pi_{ heta}}
abla_{ heta} P(au| heta) G_{ au} \ &= \sum_{ au \sim \pi_{ heta}} P(au| heta)
abla_{ heta} ln P(au| heta) G_{ au} \ &= \mathbb{E}_{ au \sim \pi_{ heta}} [
abla_{ heta} ln P(au| heta) G_{ au}] \end{aligned}$$

REINFORCE算法

$$abla_{ heta}J(heta) = \mathbb{E}_{ au\sim\pi_{ heta}}[
abla_{ heta}lnP(au| heta)G_{ au}]_{\infty}$$

$$egin{aligned} P(au| heta) &= p(s_0) \prod_{t=0}^\infty \pi(a_t|s_t, heta) p(s_{t+1}|s_t,a_t) \ lnP(au| heta) &= p(s_0) + \sum_{t=0}^\infty \pi(a_t|s_t, heta) + \sum_{t=0}^\infty p(s_{t+1}|s_t,a_t) \
abla_ heta lnP(au| heta) &= \sum_{t=0}^\infty
abla_ heta \pi(a_t|s_t, heta) \end{aligned}$$

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

$$abla_{ heta} J(heta) = \mathbb{E}_{ au \sim \pi_{ heta}}[(\sum_{t=0}^{\infty}
abla_{ heta} \pi(a_t | s_t, heta) g_t)] \qquad \qquad heta + lpha
abla_{ heta} J(heta)
ightarrow heta_{new}$$