彪哥带你学强化学习

11.actor-critic算法

DEEPLY UNDERSTAND REINFORCEMENT LEARNING

讲师：韩路彪

REINFORCE算法回顾

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

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

上: 0.01 左: 0.01 右: 0.97 下: 0.02	上: 0.00 左: 0.01 右: 0.01 下: 0.98	上: 0.31 左: 0.30 右: 0.07 下: 0.32	上: 0.31 左: 0.30 右: 0.07 下: 0.32	上: 0.10 左: 0.35 右: 0.34 下: 0.21
上: 0.94 左: 0.01 右: 0.03 下: 0.01	上: 0.01 左: 0.00 右: 0.99 下: 0.00	上: 0.00 左: 0.00 右: 0.99 下: 0.00	上: 0.00 左: 0.00 右: 1.00 下: 0.00	上: 0.00 左: 0.00 右: 0.00 下: 1.00
上: 0.92 左: 0.00 右: 0.05 下: 0.03	上: 0.07 左: 0.28 右: 0.32 下: 0.33	上: 0.11 左: 0.21 右: 0.34 下: 0.34	上: 0.20 左: 0.29 右: 0.32 下: 0.19	上: 0.00 左: 0.00 右: 0.00 下: 1.00
上: 0.00 左: 1.00 右: 0.00 下: 0.00				

公式:

probs + alpha * grad * (gamma * G + reward) = logits -> probsNew

0.01 + 0.1 * (-0.01) * (0.9 * (42.61) + (-1)) = -0.02 -> 0.02

0.00 + 0.1 * (1.00) * (0.9 * (42.61) + (-1)) = 3.73 -> 0.95

0.99 + 0.1 * (-0.99) * (0.9 * (42.61) + (-1)) = -2.71 -> 0.00

0.00 + 0.1 * (-0.00) * (0.9 * (42.61) + (-1)) = -0.01 -> 0.02

轨迹:

([4,1], 上, -1), ([3,1], 上, -1), ([2,1], 左, -1), ([2,1], 上, -1), ([1,1], 右, -1), ([1,2], 下, -1), ([2,2], 左, -1),

方差大，不稳定

actor-critic算法

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

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

actor：学习π_θ

critic：学习$q_\pi(s,a)$

可以用时序差分学习