彪哥带你学强化学习

8.深入理解 SARSA 算法及使用场景

DEEPLY UNDERSTAND REINFORCEMENT LEARNING

讲师：韩路彪

SARSA算法

SARSA：S_t A_t R S_{t+1} A_{t+1}

SARSA算法：学习的是Q^{π}，进而可以根据状态选出获得最大价值的动作

s_1

a_1	2
a_2	4
a_3	3

	s_1	s_2	s_3	s_4	。。。
a_1					
a_2					
a_3					

SARSA算法和Q算法的区别是什么？既然已经有Q算法了，为什么还需要SARSA算法？

SARSA算法

时序差分算法：temporal difference

理论依据：哈夫曼方程

训练好之后的结果
没训好时候不成立

$$Q^{\pi}(s_t, a_t) = \mathbb{E}^{\pi}_{S_{t+1}, A_{t+1}}[R(s_t, a_t, S_{t+1}) + \gamma Q^{\pi}(S_{t+1}, A_{t+1})]$$

$$\underbrace{Q^{\pi}(s_t, a_t)}_{A} \sim \underbrace{r(s_t, a_t, s_{t+1}) + \gamma Q^{\pi}(s_{t+1}, a_{t+1})}_{B}$$

蒙特卡洛采样

tderror = B - A

A_{new} = A + α*tderror　　其中α∈(0,1], 学习率

SARSA算法

$$Q^{\pi}(s_t, a_t) \sim r(s_t, a_t, s_{t+1}) + \gamma Q^{\pi}(s_{t+1}, a_{t+1})$$

$$Q_t^{\pi} \sim r + \gamma Q_{t+1}^{\pi} \qquad tderror = B - A$$

$$tderror = r + \gamma Q_{t+1}^{\pi} - Q_t^{\pi}$$

$$Q_{t_{new}}^{\pi} = Q_t^{\pi} + \alpha(r + \gamma Q_{t+1}^{\pi} - Q_t^{\pi})$$

SARSA算法

采样策略

探索性强，没有利用经验
容易多次踩坑
适用于开始训练阶段

随机策略：在可能的动作里边随机采样

经验利用，探索不足
容易错过更优动作
适用于训练后期或推理阶段

贪婪策略：哪个动作的价值高就用哪个

ε-贪婪策略：ε的概率随机，(1-ε)的概率贪婪　$\varepsilon \in [0,1]$

SARSA算法

SARSA学习步骤

1、创建Q表格，并全部置零

	s_1	s_2	s_3	s_4	。。。
a_1	0	0	0	0	0
a_2	0	0	0	0	0
a_3	0	0	0	0	0

2、采样S_t A_t

3、在S_t执行动作A_t，得的奖励 R 和状态S_{t+1}

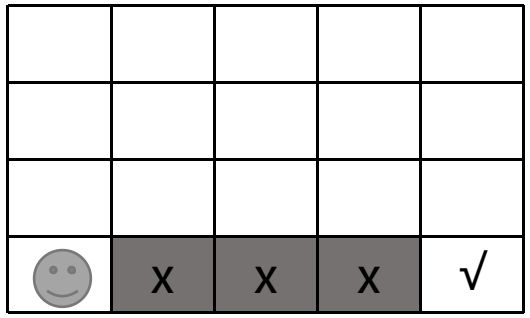
4、使用策略π在状态S_{t+1}选择动作A_{t+1}

5、根据公式更新Q(S_t, A_t)

$$Q^{\pi}_{t_{new}} = Q^{\pi}_{t} + \alpha(r + \gamma Q^{\pi}_{t+1} - Q^{\pi}_{t})$$

6、循环3~5步，直到所有Q值收敛

SARSA算法



悬崖漫步

- 从左下角出发，每走一步奖励-1
- 如果在边缘，下个位置不存在，位置不动，奖励-1
- 掉入悬崖奖励-100，结束
- 走到右下角奖励100，结束

	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{21}	S_{22}	S_{23}	S_{24}	S_{25}	S_{31}	S_{32}	S_{33}	S_{34}	S_{35}	S_{41}	S_{42}	S_{43}	S_{44}	S_{45}
上																				
下																				
左																				
右																				