# 彪哥带你学强化学习

## 7.深入理解Q学习算法及优化思路

**DEEPLY UNDERSTAND REINFORCEMENT LEARNING**

讲师：韩路彪

# Q-learning

Q：最大动作状态价值函数$Q^*(S_t, A_t)$，指在状态$S_t$采取动作$A_t$的最大价值

Q-learning：学习的是$Q^*$，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 |

| | $s_1$ | $s_2$ | $s_3$ | $s_4$ | ∘ ∘ ∘ |
|---|---|---|---|---|---|
| $a_1$ | | | | | |
| $a_2$ | | | | | |
| $a_3$ | | | | | |

# Q-learning

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

$$Q^\star(s_t,a_t)=\mathbb{E}_{S_{t+1}\sim p(\cdot|s_t,a_t)}[R(s_t,a_t,S_{t+1})+\gamma\max_{A_{t+1}\in\mathcal{A}}Q^\star(S_{t+1},A_{t+1})]$$

$$Q^\star(s_t,a_t)\sim r(s_t,a_t,s_{t+1})+\gamma\max_{A_{t+1}\in\mathcal{A}}Q^\star(s_{t+1},A_{t+1})]$$

A　　B

蒙特卡洛采样

td = B - A

$A_{new}$ = A + α*td　　其中α∈(0,1], 学习率

## Q-learning

$$Q^\star(s_t, a_t) \sim r(s_t, a_t, s_{t+1}) + \gamma \max_{A_{t+1} \in \mathcal{A}} Q^\star(s_{t+1}, A_{t+1})]$$

$$Q_t^\star \sim r + \gamma \max Q_{t+1}^\star \qquad td = B - A$$

$$td = r + \gamma \max Q_{t+1}^\star - Q_t^\star$$

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

# Q-learning

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　$\varepsilon \in (0,1]$

# Q-learning

## Q-learning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$S_t$ $A_t$

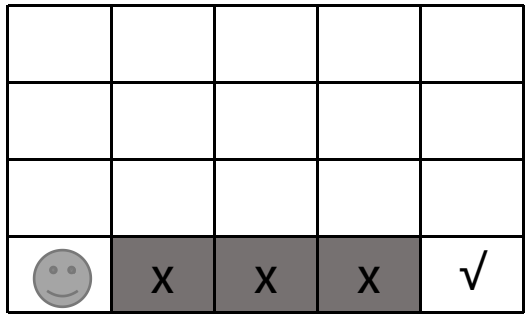
3、在$S_t$执行动作$A_t$，得的奖励r和状态$S_{t+1}$

4、根据公式更新$Q(S_t, A_t)$

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

5、循环2~4步，直到所有Q变化不大

# Q-learning




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 | | | | | | | | | | | | | | | | | | | | |