RLChina Reinforcement Learning Summer School

RLChina 2022

实践课三：经典强化学习算法实践

Prof. 田政

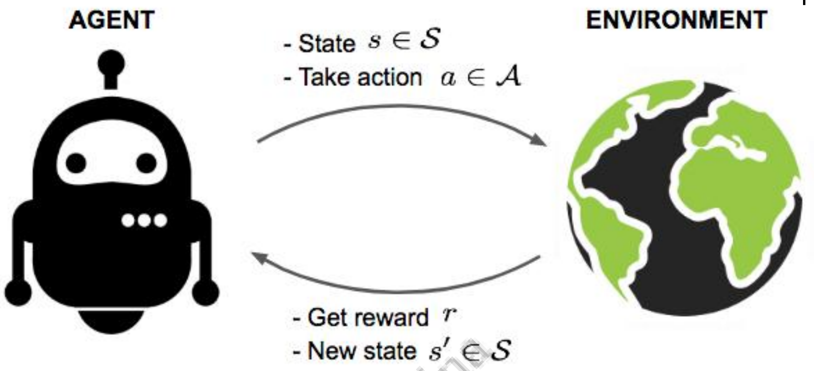
上海科技大学
ShanghaiTech University

August 17, 2022

目录

- MDP
- Dynamic Programming
  - Policy Iteration
  - Value Iteration
- Value-based:
  - Model free:
    - SARSA
    - n-step SARSA
    - Q-learning
  - Model-based:
    - Dyna-Q
  - Deep model free:
    - DQN
- Policy gradients:
  - REINFORCE

Markov Decision Process(MDP)


AGENT
ENVIRONMENT
- State $s \in \mathcal{S}$
- Take action $a \in \mathcal{A}$
- Get reward r
- New state $s' \in \mathcal{S}$

[1]

A MDP can be defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \gamma, \mathcal{R})$, which consists of:

- $\mathcal{S}$: a set of states called the state space;
- $\mathcal{A}$: a set of actions called the action space ;
- $P(s'|s,a)$: states transition function, describing the probability that action *a* and state *s* at time step *t* will lead to state *s'* at time step *t+1*;
- $\mathcal{R}(s,a)$: reward function, describing the immediate reward received after taking action *a* at state *s*.
- $\gamma \in [0,1]$: discount factor, which will discount the future rewards.

[1] *https://blog.paperspace.com/reinforcement-learning-for-machine-learning-folks/*

Goal: Maximizing Cumulated Return

- $$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k}$$

Tools:

- Policy: $\pi(a \mid s) = P\left(A_t = a \mid S_t = s\right)$, describing the probability of taking action *a* at state *s*. Policy can be stochastic or deterministic. Policy depending on current state is sufficient to be optimal in MDP.
- State-value function: $V^\pi(s) = \mathbb{E}_\pi\left[G_t \mid S_t = s\right]$
- Action-value function: $Q^\pi(s,a) = \mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right]$
- Bellman expectation function:

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_\pi[R_t + \gamma V^\pi(S_{t+1}) | S_t = s] \\ &= \sum_{a \in A} \pi(a|s) \left(r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V^\pi(s') \right) \end{aligned}$$

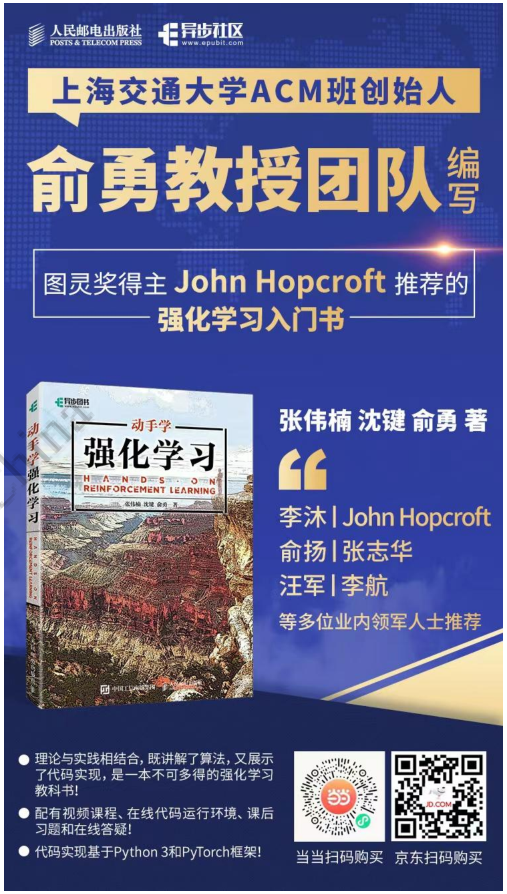
$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}_\pi[R_t + \gamma Q^\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a] \\ &= r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) \sum_{a' \in A} \pi(a'|s') Q^\pi(s',a') \end{aligned}$$

- Bellman optimal function:

$$V^*(s) = \max_{a \in \mathcal{A}} \{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V^*(s') \}$$

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \max_{a' \in \mathcal{A}} Q^*(s',a')$$

Main Resources



人民邮电出版社
POSTS & TELECOM PRESS
异步社区
www.epubit.com

上海交通大学ACM班创始人

俞勇教授团队编写

图灵奖得主 John Hopcroft 推荐的
强化学习入门书

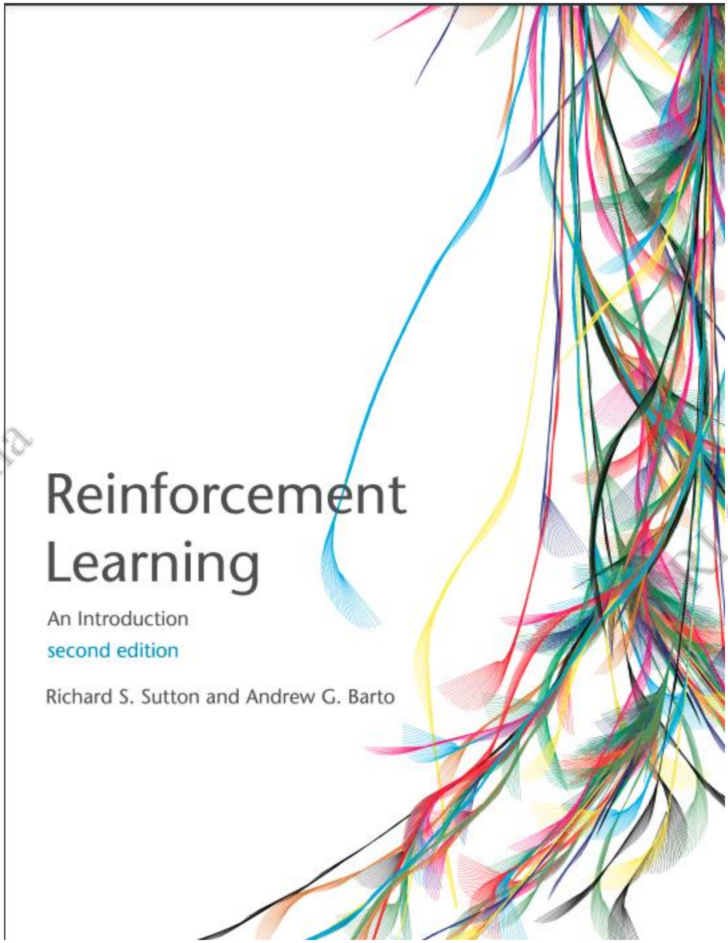
张伟楠 沈键 俞勇 著

“

李沐 | John Hopcroft
俞扬 | 张志华
汪军 | 李航
等多位业内领军人士推荐

- 理论与实践相结合，既讲解了算法，又展示了代码实现，是一本不可多得的强化学习教科书！
- 配有视频课程、在线代码运行环境、课后习题和在线答疑！
- 代码实现基于Python 3和PyTorch框架！

当当扫码购买　京东扫码购买



Reinforcement Learning

An Introduction

second edition

Richard S. Sutton and Andrew G. Barto

https://hrl.boyuai.com/

http://incompleteideas.net/book/the-book-2nd.html

上海数字大脑研究院
Digital Brain Laboratory

· 招聘 ·

上海数字大脑研究院面向中国和全球数字化业务需求，聚焦决策智能大模型、多智能体强化学习、机器学习驱动的运筹优化算法、人在环路算法、数字孪生等新一代人工智能关键技术研究与应用。

HR邮箱：jing.liu@digitalbrain.cn
联系电话：13024157621（微信）
联系人：刘小姐

社招职位

- 机器学习系统开发leader
- 机器学习平台开发工程师
- 强化学习研究员（博士/博士后）
- 后端开发工程师
- 后端开发工程师（GPU方向）
- 算法工程师（大模型预训练方向–文本）
- 算法工程师（大模型预训练方向–多模态）
- 算法工程师（工业智能能源方向）
- 算法工程师（3D数字人方向）

社招官网

实习生职位

- Unity3D开发
- 后端开发
- 项目管理
- 产品助理
- 算法
- 大模型平台
- 游戏AI
- 工业智能
- 智能交互
- 商务助理

实习生招聘官网