深度强化学习中的探索综述

田鸿龙

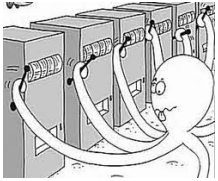
Software Institute, Nanjing University

November 25, 2020

Table of Contents

Table of Contents

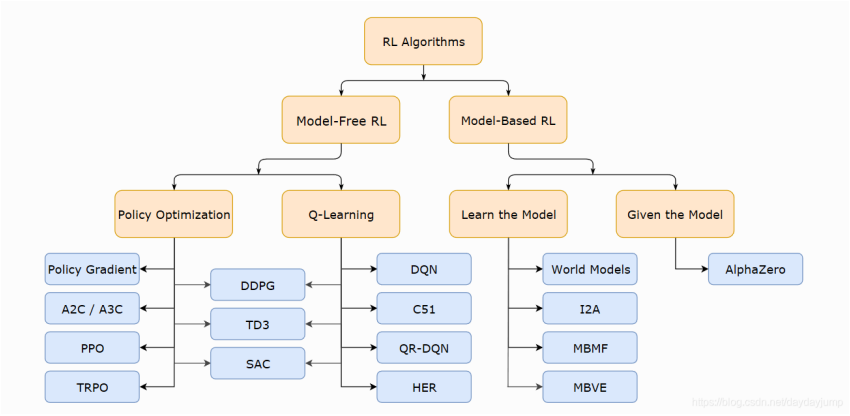
Bandit



MAB 问题：你进了一家赌场，假设面前有 K 台老虎机（arms）。我们知道，老虎机本质上就是个运气游戏，我们假设每台老虎机 i 都有一定概率 p_i 吐出一块钱，或者不吐钱（概率 $i - p_i$）。假设你手上只有 T 枚代币（tokens），而每摇一次老虎机都需要花费一枚代币，也就是说你一共只能摇 T 次，那么如何做才能使得期望回报（expected reward）最大呢？

Table of Contents

深度强化学习主流算法



两种思想

- Value-Based：和传统强化学习一样，试图学到一个值函数（Q Function 或者 V Function），通过这个值函数贪心（或在贪心的基础之上探索）形成策略，理论基础是广义策略迭代。
- Policy-Based：基于函数逼近的方法，因为深度学习强大的拟合能力而成为深度强化学习的主流方法，直接学习 $\pi : S \to A$，理论基础是策略梯度定理。
- 两种思想结合形成 Actor-Critic 方法。

References