第一讲：强化学习概述

主讲人 陈达贵

清华大学自动化系

在读硕士

深蓝学院
shenlanxueyuan.com

强化学习理论与实践

2018-12-07

目录

先修课程

- Python: (基本用法、数值计算库 numpy 等、面向对象编程基础)
- 高等数学: (微积分)
- 线性代数: (矩阵运算)
- 概率论

教材推荐

- Reinforcement Learning: An Introduction[1]
  1. 作者：Richard S. Sutton 和 Andrew G. Barto
  2. 免费下载
- Algorithms for Reinforcement Learning[2]
  - 作者：Szepesvari, Csaba
  - 免费下载

[1] Sutton and Barto 1998.

[2] Szepesvari 2010.

课程定位

- 内容：以强化学习为主，深度强化学习作为拓展
- 对象：主要面对强化学习初学者
- 形式：以理论基础为主，结合一定的编程实践，系统地讲述强化学习
- 目的：建立主流强化学习的理论体系，学会用强化学习解决实际问题

评分与答疑

评分

作业 (60%) 和 Project(40%) 两部分构成

答疑

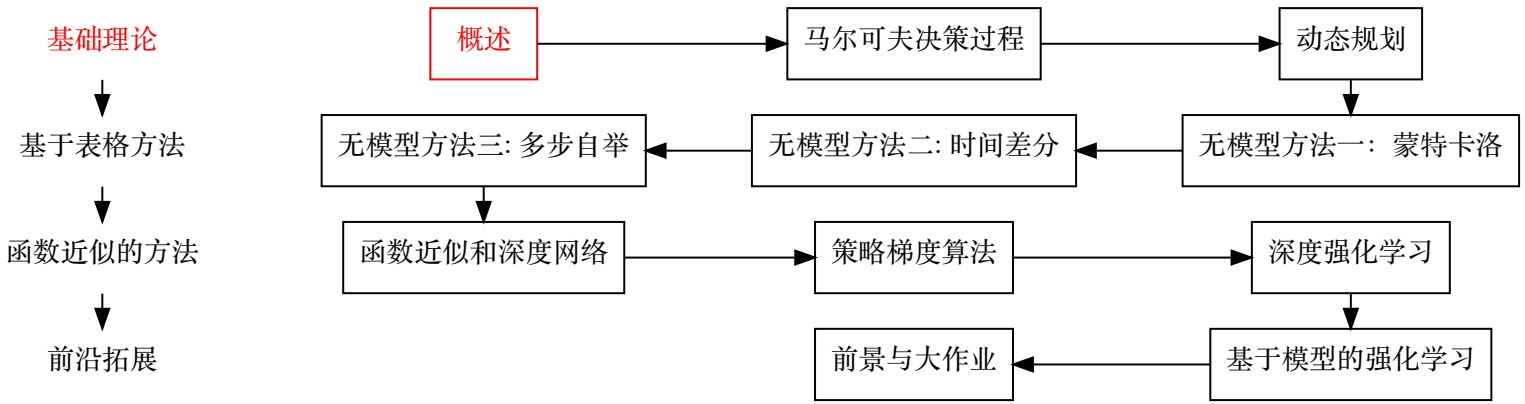
- 微信群：主要负责通知
- 深蓝学院讨论区
  1. 可以插入公式代码和图片
  2. 可以相互讨论，可以当成阅读材料
  3. 方便其他同学检索；提高提问质量；可以将问题整理.

版权说明

- 本课程中使用了一些来自 Reinforcement Learning：An Introduction 的例子。
- PPT 中较多地借鉴了 David Silver 的课程 PPT
- 同时也对网络平台上的强化学习资料有所借鉴

以上借鉴主要是因为上述材料的某些例子足够优秀，更方便组织逻辑，有利于知识的传播。如果中间发生侵权的行为，希望能直接联系本人，本人会将其删除。

课程目录说明


基础理论
基于表格方法
函数近似的方法
前沿拓展
概述
马尔可夫决策过程
动态规划
无模型方法一：蒙特卡洛
无模型方法二: 时间差分
无模型方法三: 多步自举
函数近似和深度网络
策略梯度算法
深度强化学习
基于模型的强化学习
前景与大作业

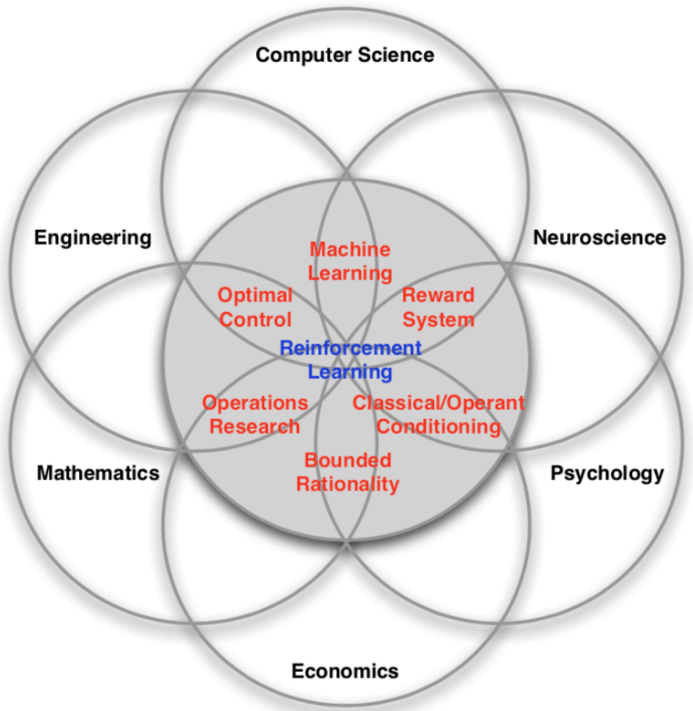
本节目录说明

目录

什么是强化学习?

1. 面向智能体的学习——通过与一个环境进行交互来实现目标。
   - 所学习的样本来自于和环境的交互
2. 通过试错和优化进行学习——用试错后的奖励 (或惩罚) 进行学习。
   - 智能体并不知道怎么做最好

强化学习和其它学科的联系[3]



[3] David Silver 2015

强化学习和其他机器学习的关系

机器学习

- 监督学习 (supervised learning): 有即时标签的学习 (分类、回归等)
- 非监督学习 (unsupervised learning): 无标签学习 (聚类问题)
- 强化学习：有延迟奖励的学习问题，介于监督和非监督之间

- 监督学习本质上也可以认为是强化学习的一种特殊形式 (无延迟场景)
- 强化学习更贴近人类的学习过程
- 强化学习可能是通往通用人工智能的道路

强化学习与其他机器学习的关系

这里拿一个例子说明监督学习，非监督学习和强化学习的区别
比如给了一些人脸图片

- 监督学习：同时，给定标签（人名），然后通过数据学习这些这些人脸是谁的脸。
  - 监督学习要求带标签的数据，这些数据是比较昂贵的
  - 标注数据也是一门学问
- 非监督学习：没有标签，但是我们可以直接观察这些人脸，来判断哪些头像是同一个人。
  - 无标签数据的数量非常庞大且容易获得
- 强化学习：没有标签信号，只有奖励信号。
  - 即时的奖励：如果分错了，只告诉你错了
  - 延迟的奖励：等全部分类完毕之后，告诉总分数如何

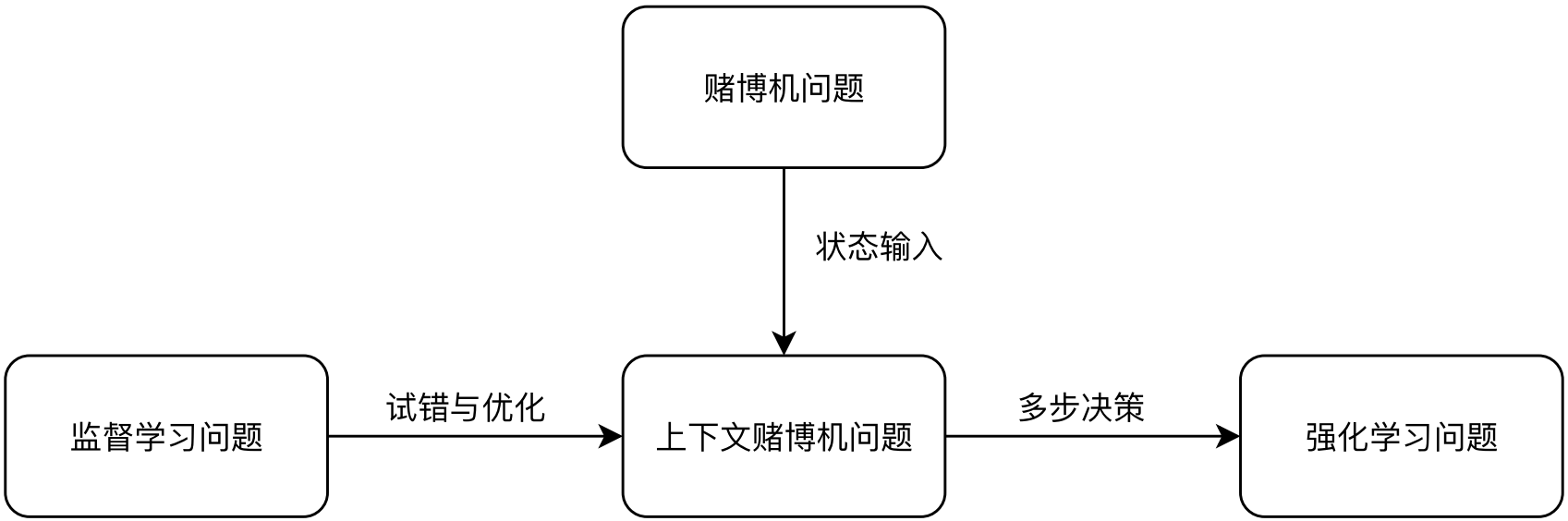
注：课后要求对这些概念有所理解，尤其是理解监督学习

强化学习的直观特性

- 没有监督信号，只有奖励信号 (具体 vs 抽象)
- 奖励信号大都是延迟的，不是瞬时的
- 问题是优化问题 (怎么做最好)
- 数据具有时间相关性，不满足独立同分布 (iid) 假设
- 智能体的动作是可以影响它之后的数据

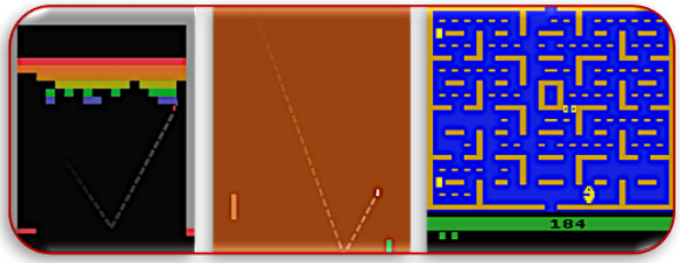
和监督学习和赌博机问题的关系

- 赌博机 (bandit) 问题：单步优化问题，且没有输入状态。每一个时刻都有奖励，每次决策跟后续状态没有关系。
- 上下文赌博机 (contextual bandit) 问题：在赌博机问题上增加了状态输入，所作出的决策是跟状态相关的。
- 上下文赌博机问题和监督学习的区别是给标签还是给奖励。
- 当上下文赌博机问题拓展到多步时会变成强化学习问题



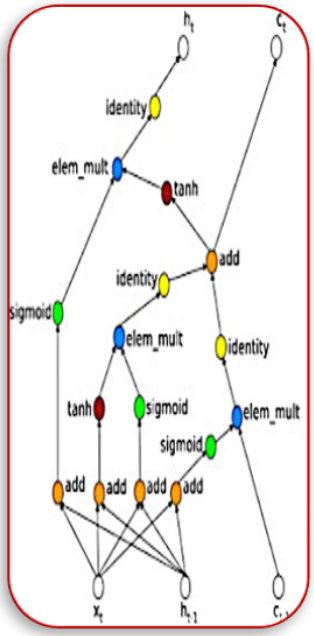
强化学习例子

Atari视频游戏

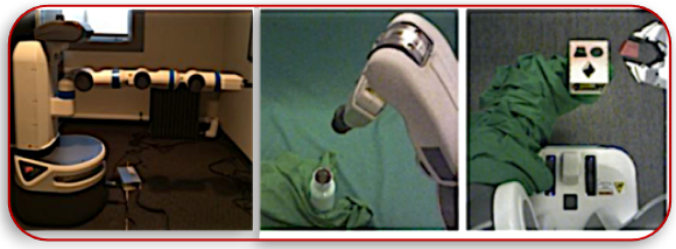


AlphaGo

学会学习



机器人控制



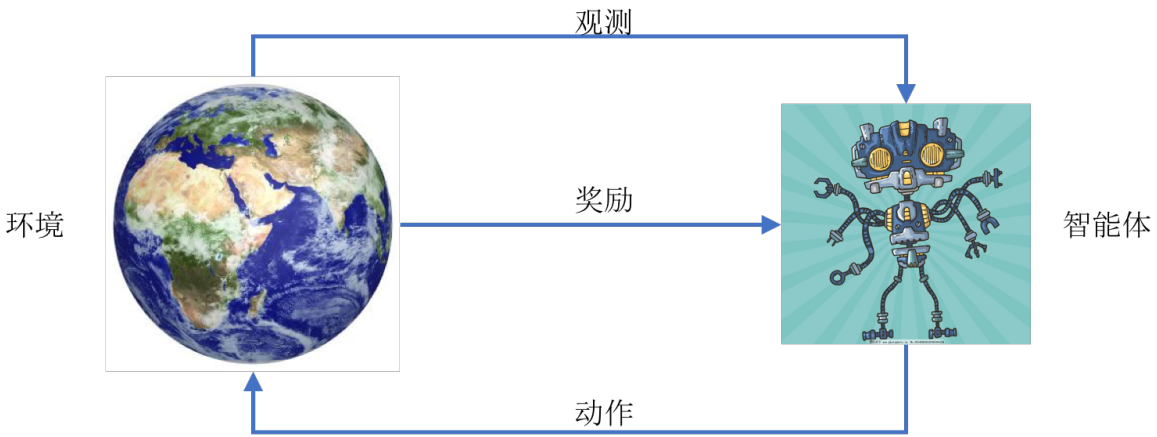
无人驾驶

强化学习例子

- 无人机特技飞行
- 管理投资策略系统
- 控制电力系统
- 广告投放策略控制
- ...

目录

整体结构



每个时刻 t,

- 智能体（agent）
  - 智能体执行动作 A_t，并在环境中得到观测 O_t和奖励 R_t
- 环境（environment）
  - 环境会对智能体的动作 A_t的做出反应，然后发送新的观测 O_{t+1}和奖励R_{t+1}

智能体与环境

- 智能体是指强化学习需要优化的部分，是我们能够精确控制的部分
- 环境是我们不能直接控制的部分
- 区分智能体和环境是强化学习的第一步
- 环境并不是指自然环境
- 不同的问题，智能体和环境的划分也有所区别
  - 机器人探索房间 vs. 机器人行走控制
  - 仿真环境中的控制 vs. 实际环境中的控制

目录

奖励与奖励假设

- 奖励 (reward) 是强化学习的核心
  - 可以没有观测，但是不能没有奖励。
  - 奖励是强化学习区别其他机器学习的标志特征。
- 奖励的特点
  - 奖励 R_t 是一个标量反馈
  - 它衡量了智能体在时间 t 上做得有多好
  - 智能体的目标就是最大化累计奖励

奖励假设

强化学习的目标就是最大化期望累计奖励

注: 如果一个问题不满足奖励假设，那么就不能用强化学习去解决!

奖励的示例

- 无人机和无人车控制
  - + 奖励：如果按预定轨迹运行
  - -奖励：碰撞或翻车
- 下围棋
  - +/-奖励：赢了/输了
- Atari 游戏
  - +/-奖励：得分增加/减少
- 机械臂控制
  - + 奖励：成功抓到东西

注：奖励并不要求一定要有正有负，只有正的奖励和负的奖励都可以！

长期奖励

- 每一个动作都可能有一个长期的结果
- 奖励可能是延迟的
- 有时我们需要牺牲一些短期奖励来获得更多的长期奖励

例子

- 下围棋：只有在最后才能获得奖励
- 打砖块游戏：当球运行一段到砖块时才有奖励
- （牺牲的例子）：抄袭虽然可以获得短暂的奖励，但是为了长期奖励（取得进步），我们要牺牲短期奖励。

奖励值与回报值

当智能体在时间 t 做出动作 A_t 时，会在未来收到奖励序列 $R_t, R_{t+1}, R_{t+2}, \cdots$。我们的目的是使累计奖励最大。一种通用的累计奖励的定义方式是将这些奖励值进行加权求和:

$$G_t = w_t R_t + w_{t+1} R_{t+1} + w_{t+2} R_{t+2} + \cdots$$

其中 w_t 表示不同时间的加权系数

- 我们把上面的 G_t 称为回报值（Return）
- 回报值衡量了动作 A_t 对未来结果的影响
- 强化学习的目的即变成了在每个时刻，使未来的期望回报值最大

回报值

- $w_{t+n} = 1, \forall n$
  - 我们将所有时刻的奖励看成一样重要的
  - 无衰减回报值 (undiscounted return)
- $w_{t+n} = \gamma^n, \gamma \in [0, 1]$
  - $G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots$
  - $\gamma = 0$ 时，退化成赌博机问题
  - $\gamma = 1$ 时，变成无衰减回报值
  - $\gamma \in (0, 1)$ 时，γ 值衡量了我们对未来奖励的关注度，γ 越大，表示我们越关注长期奖励，γ 越小表示我们越关注短期奖励
  - 衰减回报值 (discounted return)
  - γ 被称为衰减系数

衰减系数的理解

- 未来的奖励还会受到其他动作的影响
- 在估计未来奖励时，我们的把握也越来越小

目录

历史和状态

- 历史 (history)是一个观测、动作和奖励的序列

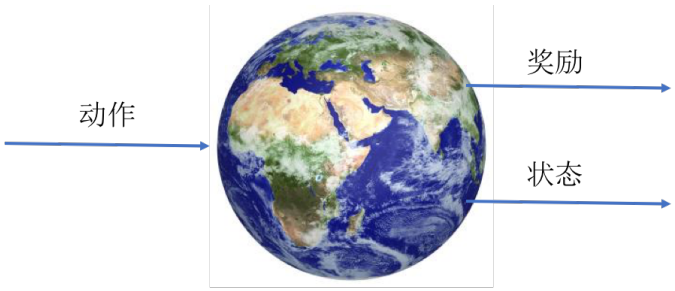
$$H_t = O_1, R_1, A_1, O_2, R_2, A_2, \cdots, A_{t_1}, O_t, R_t$$

- 就是智能体在时间 t 以前的所有的交互变量
- 根据历史：
  - 智能体选择动作 A_t
  - 环境产生新的观测 O_{t+1} 和奖励 R_{t+1}
- 状态是历史的一种表达，根据状态我们可以判断接下来发生什么
- 本质上，状态是历史的一个函数 $S_t = f(H_t)$

状态的例子

- 下棋时，棋盘现在的布局可以被认为是状态 $S_t = O_t$
- 玩打砖块游戏时，前几帧的观测被认为是状态 $S_t = O_{t-3}, O_{t-2}, O_{t-1}, O_t$
- 玩 CS 时，可能整个历史都被认为是状态 $S_t = H_t$

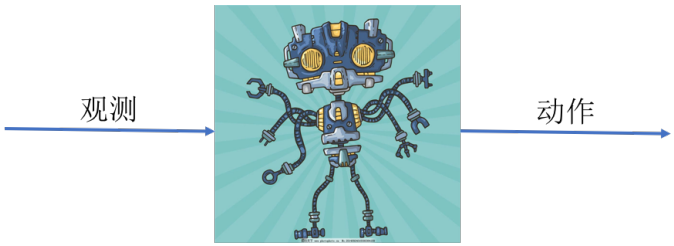
环境状态



环境状态

- 环境状态 S_t^e 是环境的内部表达
- 所有能够影响环境产生观测/奖励的数据都被认为是环境状态的一部分
- 环境状态一般是智能体观察不到的
- 即使环境状态 S_t^e 可见的，一般也包含了不相关的信息

智能体状态



- 智能体状态 S_t^a 是智能体的内部表达
- 所有能够影响智能体做出下一个动作的数据都被认为是智能体状态的一部分
- 强化学习中使用的状态
- 智能体状态可能是历史的任何函数

$$S_t^a = f(H_t)$$

环境状态和智能体状态对比

- 机器人控制
  - 环境状态：所有机械零件的参数，状态等
  - 智能体状态：传感器所获得的数据
- 3D 游戏
  - 环境状态：所有的游戏参数以及对手信息
  - 智能体状态：玩家所能看到的观测

要点

- 对于智能体来说，环境状态是未知的，智能体状态是已知的
- 智能体通过智能体的状态来做出相应的动作
- 没有特殊说明的情况下，我们所说的状态均指智能体状态 $S_t = S_t^a$

全观测和部分观测环境

- 全观测
  - 智能体能够直接观测到环境状态 $O_t = S_t^a = S_t^e$
  - 或者说智能体状态等价于环境状态
  - 这是强化学习的主要研究问题（之后详述）——马尔可夫决策过程
- 部分观测
  - 智能体不直接观测到环境状态
  - 智能体状态 $\neq$ 环境状态
  - 部分观测下的马尔可夫决策问题

目录

动作

- 动作是智能体主动和环境交互的媒介
- 动作必须对环境起到一定的控制作用 (尤其是对奖励)
- i.e. 动作序列 $A_1, A_2, A_3, \cdots$ 能够影响智能体的回报值

要点

- 动作要能改变未来所能获得的奖励
  - 打砖块时上/下动作不能改变环境
- 奖励的设置要能被动作改变
  - 比如时间的流逝是不受改变的

注: 如果上述两点不满足，那么就无法解这个强化学习问题

目录

智能体的主要组成部分

智能体 (agent)是强化学习需要优化的部分。

- 一个强化学习的智能体主要以下一个或几个部分
  - 策略 (Policy)
  - 值函数 (value function)
  - 模型 (model)

注: 这里只是简述，后续的课程会对这几个部分详述

目录

策略

- 策略是智能体的核心，我们最终的目的就是找到一个策略
- 通过策略可以描述智能体的行为
- 它是一个从状态到动作的映射
- 直观上描述就是：当智能体在什么状态时应该做什么事
- 策略分两大类：
  - 确定性策略：$a = \pi(s)$
  - 随机策略：$\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$

目录

值函数

- 值函数是对回报值的预测（期望）
- 值函数主要用来评价不同状态的好坏
- 可以用来指导动作的选择

$$v_\pi(s) = \mathbb{E}_\pi\left[G_t|S_t = s\right] = \mathbb{E}\left[R_{t+1} + \gamma R_{t+2} + \cdots|S_t = s\right]$$

目录

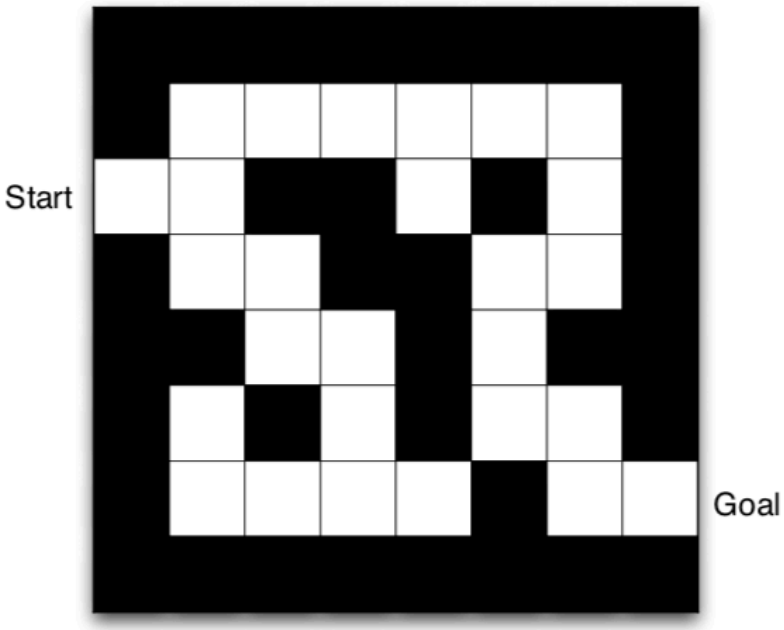
模型

- 这里的模型指智能体所拥有的对环境的预测模型，主要包含两部分：
  - $\mathcal{P}$：预测下一个状态是什么
  - $\mathcal{R}$：预测下一奖励是多少

$$\mathcal{P}_{ss'}^{a} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$$
$$\mathcal{R}_{s}^{a} = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$$

- 这里我们将环境看成一个黑盒子，只关心其输入输出
- $\mathcal{P}_{ss'}^{a}$ 是一个概率
  - 确定性环境
  - 随机性环境
- $\mathcal{P}$ 和 $\mathcal{R}$ 可能是不精确的

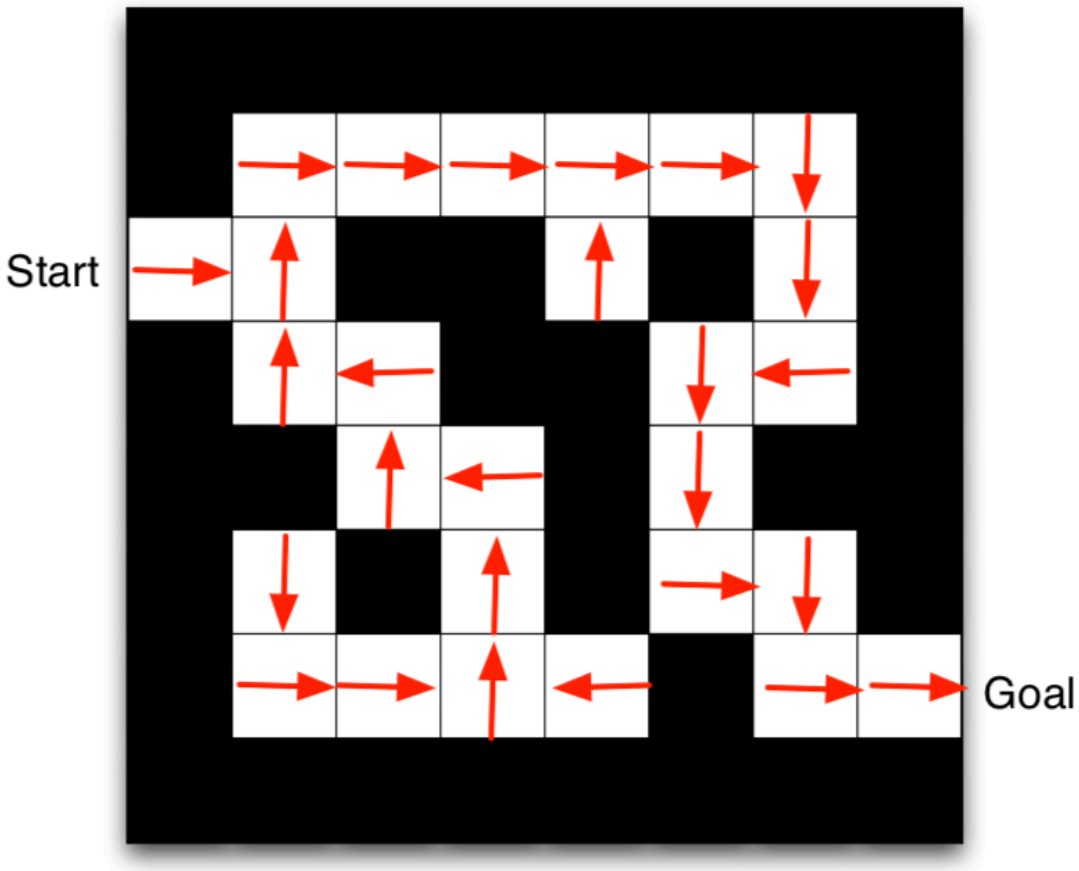
迷宫例子[4]



- 奖励：每走一步获得-1 的奖励
- 动作：上下左右
- 状态：智能体的位置

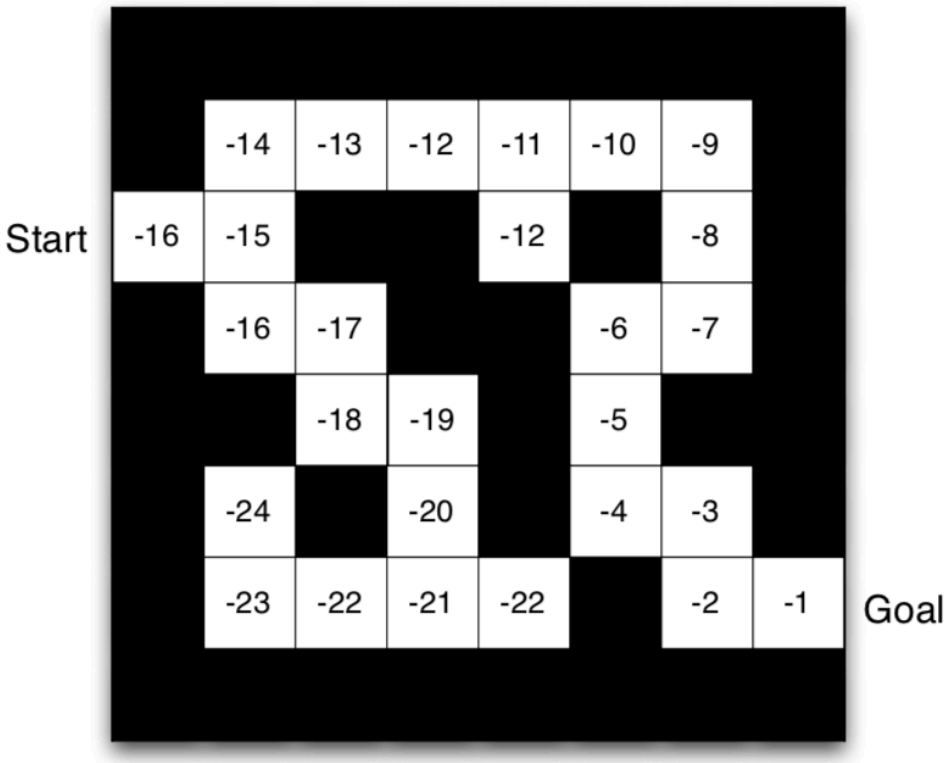
[4] David Silver 2015

迷宫策略



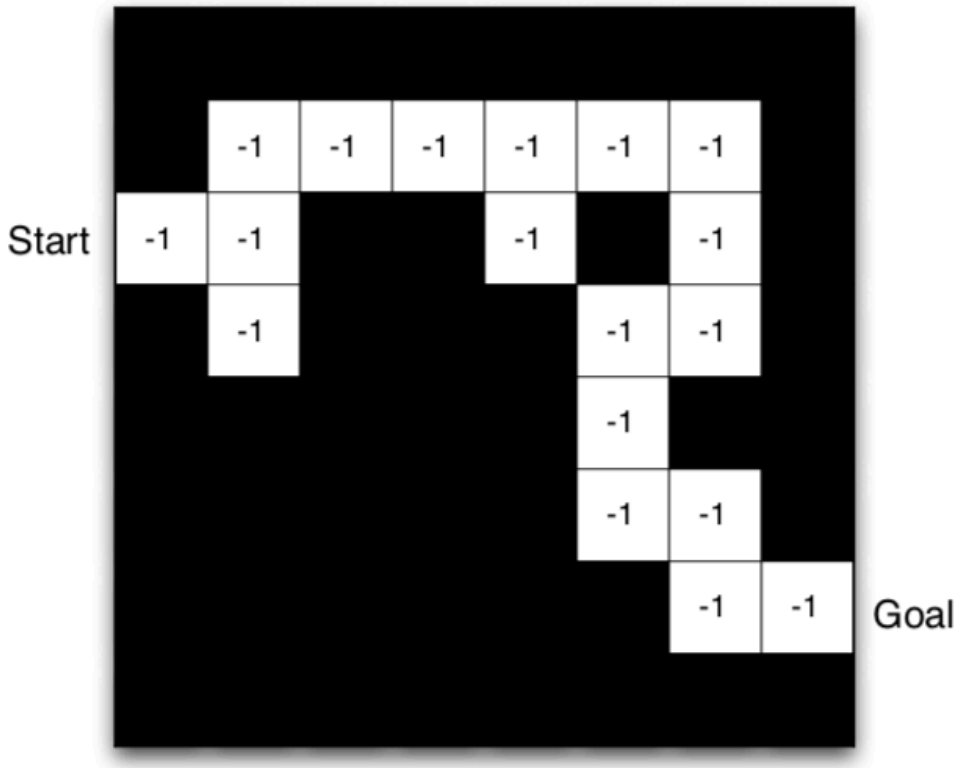
- 图中的箭头描述了 $\pi(s)$——确定性策略

迷宫值函数



- 图中的数字代表了 $v_\pi(s)$——按照 π 进行的值函数

迷宫模型



- 网格的布局即代表了 $P^a_{ss'}$——确定性环境
- 数值代表了 R^a_s
- 注：这里省略了左下角的图例

目录

按环境分类

- 全观测环境下的强化学习
- 部分可观测环境下的强化学习

按智能体的成分分类

- 基于值函数: 学习值函数
- 基于策略: 学习策略
- Actor Critic: 同时学习值函数和策略

按有无模型分类

- 无模型强化学习
- 基于模型的强化学习

按使用的手段分类

- 传统强化学习
- 深度强化学习

目录

学习与规划

对于序列决策问题有两个基本问题

- 强化学习
  - 环境未知
  - 智能体与环境进行交互
  - 智能体会不断地改善自己的策略
- 规划 (Planning)
  - 环境已知
  - 智能体可以根据模型直接计算，不用交互
  - 智能体会不断地改善自己的策略

注: 既利用模型进行规划，又与环境交互进行强化学习，这就构成了基于模型的强化学习

注: 当有很精确的环境模型时，可以直接用规划的方式解

探索与利用

- 探索 (Exploration) 和利用 (Exploitation) 是强化学习的根本问题
- 强化学习会根据过去的经验得到一个好的策略
- 但是还有没有更好的策略呢?
- 探索是为了能够发现环境的更多信息
- 利用是为了利用当前已知的信息来最大化回报值
- 两者同等重要

评价和优化

在很多地方叫预测（Prediction）和控制（Control）

- 评价：给一个策略，评价该策略的好坏，即求对应的值函数
- 优化：找到最优的策略

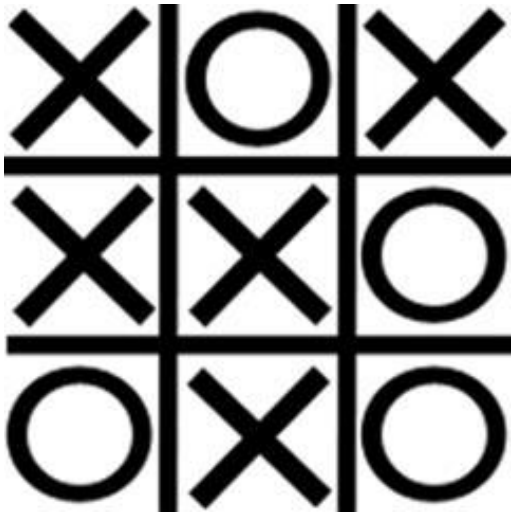
这两个过程构成了强化学习的基本思路

目录

文字作业

- 选择你认为可以建模成强化学习问题的两个场景回答下面的问题
  1. 为什么需要用 RL 来建模？为什么不能使用 SL？
  2. 环境和智能体分别指什么？
  3. 状态，动作，奖励分别代表什么？
  4. 是全观测的还是部分观测的，为什么？
  5. 环境模型已知还是未知？
  6. 你所认为的最佳策略应该是确定性的还是随机性的？

TieTacToe



- 实现一个这样的环境
- 实现一个智能体，其策略是随机的
- 体会环境与智能体之间的关系，理解状态、奖励和动作是怎么在两者之间进行交互的

引用

Sutton, Richard S and Andrew G Barto (1998).
Reinforcement learning: An introduction. Vol. 1. 1. MIT press Cambridge.
Szepesvari, Csaba (2010). "Algorithms for Reinforcement Learning". In:
International Conference on Computing, p. 103.