lec04

没有03，03没什么内容
虽然04似乎也没什么内容，估计不考，随便看看

强化学习的关键特征

- 强化学习是什么?
  - 一种人工智能的方法
  - 从交互中学习
  - 目标导向的学习
  - 从与外部环境的交互中学习
  - 学习该做什么——如何将情景映射到动作——以最大化数值奖励信号

标准代理

- 时间相关
- 持续学习和规划
- 代理影响环境
- 环境是随机且不确定的

环境类型

- 确定性 vs. 随机性:
  - 下一个状态是否完全由当前状态决定（如果除了其他代理的动作外，环境是确定的，则环境是战略性的）
  - 例如，出租车驾驶是随机的（无法预测交通，轮胎可能爆胎），填字迷宫是确定的
- 完全可观察 vs. 部分可观察:
  - 代理的传感器可以访问环境的（不）完整状态
  - 例如，国际象棋是完全可观察的，出租车驾驶是部分可观察的
- 情景性 vs. 顺序性:
  - 代理的经验是否分为原子的“情景”。决策不依赖于先前的决策/行动。可以持续有限时间的任务，即具有终端状态的任务，称为情景任务
  - 例如，出租车驾驶是顺序的，迷宫运行是情景的
- 动态 vs. 静态:
  - 代理的传感器可以访问环境的（不）完整状态
  - 例如，国际象棋是完全可观察的，出租车驾驶是部分可观察的
- 离散 vs. 连续:
  - 有限数量的不同、明确定义的状态和动作
  - 例如，出租车驾驶是连续的，扑克是离散的
- 单智能体 vs. 多智能体:
  - 代理单独在环境中操作
  - 例如，出租车驾驶是多智能体的，填字游戏是单智能体的

关键特征

- 学习者没有被告知要采取哪些行动
- 试错搜索
- 延迟奖励的可能性
  - 牺牲短期利益以获得更大的长期收益
- 需要探索和利用
- 考虑到目标导向代理与不确定环境交互的整个问题

监督学习 vs. 强化学习

- 监督学习：
  - 系统输入输出
  - 训练需要期望的（目标）输出
- 强化学习：
  - 目标：获得尽可能多的奖励
  - 系统输入输出=动作
  - 训练需要奖励/反馈

强化学习的元素

- 策略：做什么
  - 策略定义了学习代理在给定时间的行为方式
- 奖励：什么是好的
  - 奖励信号指示在即时意义上什么是好的
- 价值：什么是好的，因为它预测了奖励
  - 价值函数指定了从长远来看什么是好的
- 模型：什么跟随什么
  - 环境的行为，允许推断环境将如何行为

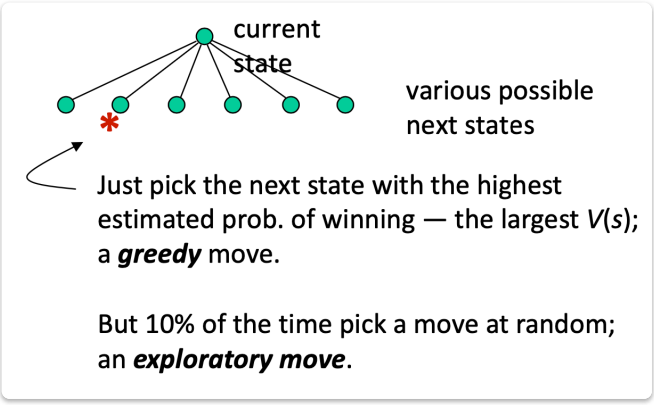
示例：井字棋

- 建立一个每个状态都有一个条目的表：
  - 估计的获胜概率 $V(s)$

State	V(s)	
	0.5	?
	0.5	?
…	…	…
	1	win
…	…	…
	0	loss
…	…	…
	0	draw

- 现在进行很多游戏：

- 选择我们的动作，向前看一步：



强化学习方法

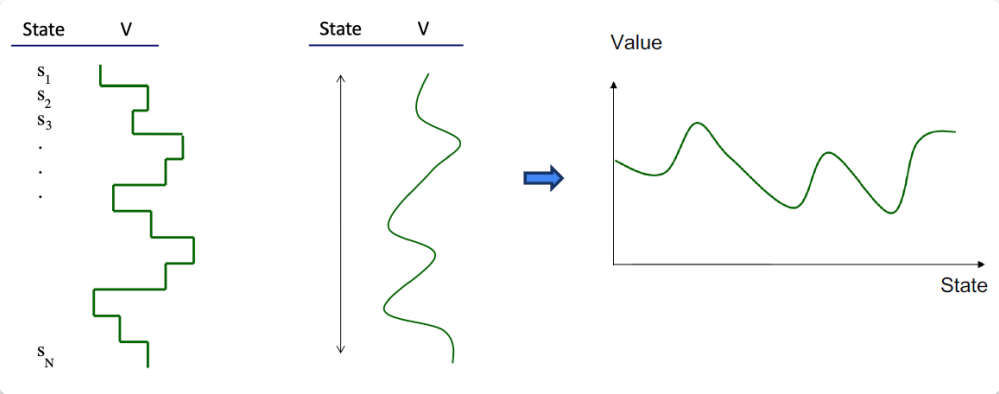
- 贪心探索
- 我们将状态s的值向状态s'的值逼近：
  - 备份
  - $V(s) \leftarrow V(s) + \alpha[Vs' - V(s)]$

如何改进井字棋玩家?

- 利用对称性
- 我们需要“随机”动作吗？为什么？
- 我们能从“随机”动作中学习吗？
- 我们能离线学习吗？
  - 从自我对弈中预训练？
  - 使用学到的对手模型？
  - .....

泛化：价值函数近似器

- 计算状态的值



井字棋为什么太简单?

- 有限的小状态数
- 总是可以进行一步预见
- 状态完全可观察

- .....