COMP9414：人工智能第9b讲。强化学习

韦恩-沃布克

电邮：w.wobcke@unsw.edu.au

本讲座

- 强化学习与监督学习
- 优化模型
- 勘探与开发
- 时差学习
- Q-Learning

学习的类型

- 监督学习
  - 代理人被告知输入及其目标输出的例子，必须学习从输入到输出的函数，以达到与训练实例相一致，并对新的实例进行归纳总结
- 强化学习
  代理人没有为每个输入提供目标输出，但会定期获得奖励，并且必须学习如何最大化(长时间的回报
- 无监督学习
  代理人只得到一系列的输入，并且必须在这些输入中找到有用的模式。

监督学习

- 给定一个训练集和一个测试集，每个训练集由一组项目组成，训练集的每个项目都有一组特征和一个目标输出
- 学习者必须学习一个能够预测任何给定项目的目标输出的模型（由其特征集来描述）。
- 学习者被赋予训练集中每个项目的输入特征和目标输出。
  - 项目可以一次性提出（批处理）或按顺序提出（在线）。

- 项目可以随机或按时间顺序呈现（流）。
学习者在定义模型时完全不能使用测试集

- 模型的评估是通过预测测试集中每个项目的输出的性能来进行的。

学习行动

监督学习可以用来从情境-
行动对的训练集中学习行动（称为行为克隆）。

然而，在许多应用中，它是困难的、不适当的。
训练集，甚至不可能提供一个 "训练集"

- 最优控制
  - 移动机器人、撑杆平衡、驾驶直升飞机
- 资源分配
  - 工作车间调度，移动电话频道分配
- 分配和控制的混合
  - 电梯控制，五子棋

强化学习框架

- 代理人与环境的互动
- 有一个*状态*的集合S和一个*行动*的集合A
- 在每个时间步骤t，代理人处于某个状态s_t，必须选择一个行动a_t，然后进入状态$s_{t+1} = \delta(s_t, a_t)$，并获得奖励$r(s_t, a_t)$
- 一般来说，$r()$和$\delta()$可以是多值的，有一个随机元素
- 目的是找到一个最佳*政策*$\pi: S \rightarrow A$，使累积奖励最大化。

优化模型

一个快的五分钱值得一个慢的一毛钱吗？

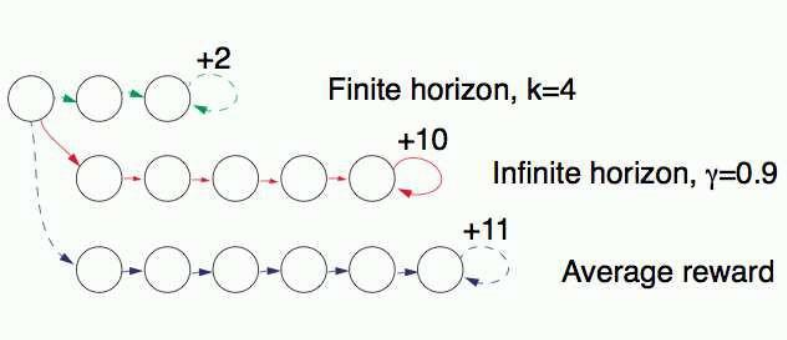
有限视界奖励 $\sum_{i=0}^{h} r_{t+i}$

平均奖励 $\lim_{h\rightarrow\infty} \frac{1}{h} \sum_{i=0}^{h} r_{t+i}$

无限折扣的奖励 $\sum_{i=0}^{\infty} \gamma^i r_{t+i}$。 $0 \leq \gamma < 1$

- 有限水平线奖励在计算上很简单
- 无限折扣的奖励更容易证明定理
- 平均奖励很难处理，因为无法在不久的小奖励和非常遥远的未来的大奖励之间做出明智的选择

优化模型的比较



环境类型

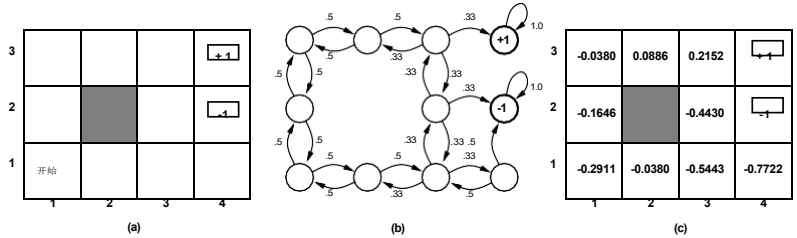
环境可以是

- 被动的和随机的
- 主动和决定性的（国际象棋）
- 主动的和随机的（西洋双陆棋）。

价值函数

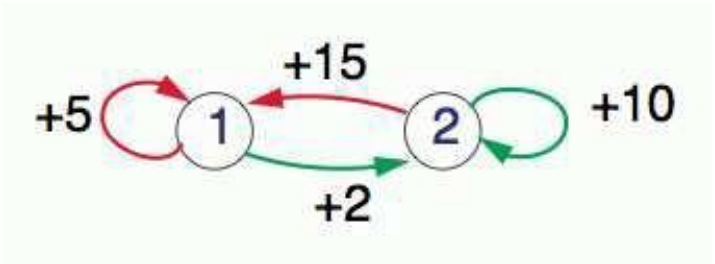
对于每个状态$s \in S$，状态s的价值，$V(s)$，由政策π和奖励函数r决定。

例子。政策是随机选择继承人



(b)中的政策决定了最终奖励下的价值函数(c)，直到最终状态{+1, -1}，奖励函数(a)

例子。无限折扣的奖励



问题。最佳政策是否取决于γ？

例子的计算方法

该定理。在一个确定的环境中，对于一个最佳政策，价值函数V^*满足贝尔曼方程。$V^*(s) = r(s, a) + \gamma V^*(\delta(s, a))$具有无限的贴现报酬，其中$a = \pi^*(s)$是$s$处的最优行动。

让$\delta^*(s)$为$\pi^*(s)$的过渡函数，假设$\gamma=0.9$

1. 假设 $\delta^*(s_1) = s_1$。那么 $V^*(s_1) = 5 + 0.9V^*(s_1)$ 所以 $V^*(s_1) = 50$ 假设 $\delta^*(s_2) = s_2$。那么 $V^*(s_2) = 10 + 0.9V^*(s_2)$ 所以 $V^*(s_2) = 100$
2. 假设 $\delta^*(s_1) = s_2$。那么 $V^*(s_1) = 2 + 0.9V^*(s_2)$ 所以 $V^*(s_1) = 92$ 假设 $\delta^*(s_2) = s_2$。那么 $V^*(s_2) = 10 + 0.9V^*(s_2)$ 所以 $V^*(s_2) = 100$

一个最佳政策π^*，决定了一个最佳价值函数V^*

3. 假设$\delta^*(s_1) = s_2$。那么$V^*(s_1) = 2 + 0.9V^*(s_2)$所以$V^*(s_1) = 81.6$
   假设$\delta^*(s_2) = s_1$。那么$V^*(s_2) = 15 + 0.9V^*(s_1)$所以$V^*(s_2) = 88.4$所以2是最优政策。

新南威尔士大学 ©W.Wobcke等人，2019-2022年

勘探/开发的权衡

大多数时候，代理人应该选择 "最佳 "行动

然而，为了确保能够学到最佳策略，代理人必须偶尔选择不同的行动，例如

- 在5%的时间内选择一个随机行动，或
- 使用玻尔兹曼分布来选择下一个行动

$$P(a) = \frac{e^{\hat{V}(a)/T}}{\sum_{b \in A} e^{\hat{V}(b)/T}}$$

K型武装匪徒问题

只有一种状态的主动随机环境的特殊情况被称为K-Armed Bandit问题，因为它就像在一个有几个（友好的）老虎机的房间里，在有限的时间内，试图收集尽可能多的钱。
每个行动（老虎机）提供不同的平均奖励

时差学习

TD(0) [也叫AHC，或Widrow-Hoff规则]

$$\hat{V}(s) \leftarrow \hat{V}(s) + \eta[\, r(s,a) + \gamma \hat{V}(\delta(s,a)) - \hat{V}(s)] \text{。}$$

(η = 学习率)

下一个状态的（折现的）价值，加上眼前的奖励，被用作当前状态的目标值。

一个更复杂的版本，称为TD(λ)，使用一个加权平均的未来国家

Q-Learning

对于每个$s \in S$，让$V^*(s)$是可从s获得的最大折现报酬，让$Q(s, a)$是先做行动a，然后采取最佳行动的折现报酬。

那么最优政策是

$$\pi^*(s) = \arg\max_a Q(s, a)$$

其中 $Q(s,a) = r(s,a) + \gamma V^*(\delta(s,a))$

那么 $V^*(s) = \max_a Q(s,a)$

所以 $Q(s,a) = r(s,a) + \gamma \max_b Q(\delta(s,a), b)$

这使得Q的迭代逼近可以通过以下方式进行

$$\hat{Q}(s,a) \leftarrow r(s,a) + \gamma \max_b \hat{Q}(\delta(s,a), b)$$

新南威尔士大学 ©W.Wobcke等人，2019-2022年

理论结果

该定理。
对于任何确定性的马尔科夫决策过程，假设有适当的随机化策略，Q-learning最终将收敛到最优策略。

(Watkins & Dayan 1992)

该定理。TD-learning也会收敛，概率为1。

(Sutton 1988, Dayan 1992, Dayan & Sejnowski 1994)

理论结果的局限性

- 延迟强化
  一个行动产生的奖励可能要在几个时间步骤之后才能收到，这也会减慢学习速度。
- 搜索空间必须是有限的
  - 如果搜索空间很大，收敛速度很慢
  - 依赖于无限频繁地访问每个州
- 对于 "真实世界 "的问题，不能依赖查找表。
  - 需要有某种归纳（如TD-Gammon）。

摘要

- 强化学习是一个活跃的研究领域
- 数学成果（比人工智能的其他领域更多）
- 需要有一个适当的代表
- 未来的算法，它选择自己的表现形式？
- 许多实际应用