《强化学习》第一讲 简介

叶强

深度学习 / 强化学习 / 机器学习 / 算法 / 眼科学

本讲是对于强化学习整体的一个简单介绍，描述了强化学习是什么，解决什么问题，大概用什么样的方式来解决问题。介绍了强化学习中常用的概念。这些概念非常重要，贯穿于整个强化学习始终，但是在这一讲，读者仅需对这些概念有个初步的印象。

引子

推荐教材(下载地址)

1. An Introduction to Reinforcement Learning, Sutton and Barto, 1998

2. Algorithms for Reinforcement Learning, Szepesvari, 2009

强化学习在不同领域有不同的表现形式：神经科学、心理学、计算机科学、工程领域、数学、经济学等有不同的称呼。

强化学习是机器学习的一个分支：监督学习、无监督学习、强化学习

强化学习的特点：

1. 没有监督数据、只有奖励信号
2. 奖励信号不一定是实时的，而很可能是延后的，有时甚至延后很多。
3. 时间（序列）是一个重要因素
4. 当前的行为影响后续接收到的数据

强化学习有广泛的应用：像直升机特技飞行、经典游戏、投资管理、发电站控制、让机器人模仿人类行走等

强化学习问题的提出

- 奖励 Reward

R_t 是信号的反馈，是一个标量，它反映个体在t时刻做得怎么样。个体的工作就是最大化累计奖励。

强化学习主要基于这样的”奖励假设”：所有问题解决的目标都可以被描述成最大化累积奖励。

Definition (Reward Hypothesis)

All goals can be described by the maximisation of expected cumulative reward

- **序列决策 Sequential Decision Making**

目标：选择一定的行为系列以最大化未来的总体奖励

这些行为可能是一个长期的序列

奖励可能而且通常是延迟的

有时候宁愿牺牲即时（短期）的奖励以获取更多的长期奖励

- **个体和环境 Agent & Environment**

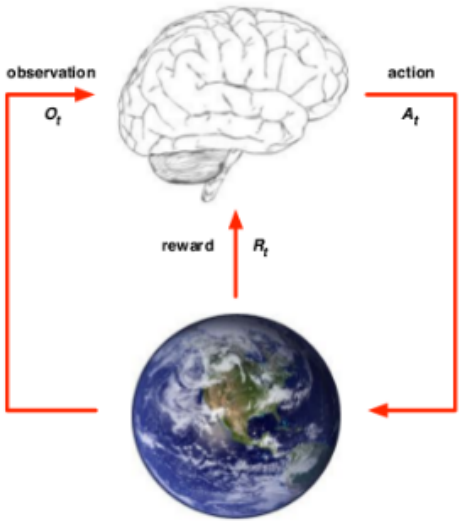
可以从个体和环境两方面来描述强化学习问题。

在 t 时刻，个体可以：

1. 有一个对于环境的观察评估 O_t ，
2. 做出一个行为 A_t ，
3. 从环境得到一个奖励信号 R_{t+1} 。

环境可以：

1. 接收个体的动作 A_t ，
2. 更新环境信息，同时使得个体可以得到下一个观测 O_{t+1}，
3. 给个体一个奖励信号 R_{t+1} 。



- **历史和状态 History & State**

历史

历史是观测、行为、奖励的序列： $H_t = O_1, R_1, A_1, \ldots, O_{t-1}, R_{t-1}, A_{t-1}, O_t, R_t, A_t$

状态

状态是所有决定将来的已有的信息，是关于历史的一个函数： $S_t = f(H_t)$

环境状态

是环境的私有呈现，包括环境用来决定下一个观测/奖励的所有数据，通常对个体并不完全可见，也就是个体有时候并不知道环境状态的所有细节。即使有时候环境状态对个体可以是完全可见的，这些信息也可能包含着一些无关信息。

个体状态

是个体的内部呈现，包括个体可以使用的、决定未来动作的所有信息。个体状态是强化学习算法可以利用的信息，它可以是历史的一个函数：$S_t^a = f(H_t)$

信息状态

包括历史上所有有用的信息，又称Markov状态。

- **马儿可夫属性 Markov Property**

一个状态St是马尔可夫的，当且仅当：$P[S_{t+1}|S_t] = P[S_{t+1}|S_1, S_2, \ldots, S_t]$

Definition

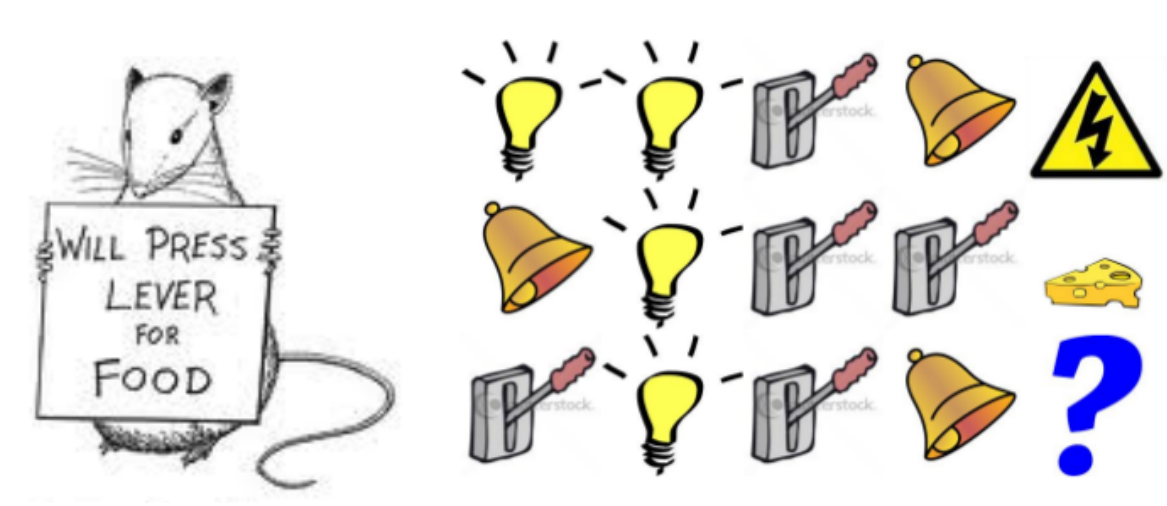
A state S_t is Markov if and only if

$$\mathbb{P}[S_{t+1} \mid S_t] = \mathbb{P}[S_{t+1} \mid S_1, \ldots, S_t]$$

也就是说，如果信息状态是可知的，那么所有历史信息都可以丢掉，仅需要 t 时刻的信息状态就可以了。例如：环境状态是Markov的，因为环境状态是环境包含了环境决定下一个观测/奖励的所有信息；同样，（完整的）历史 H_t 也是马尔可夫的。

示例——马儿可夫性

有如下三个针对老鼠的事件序列，其中前两个最后的事件分别是老鼠遭电击和获得一块奶酪，现在请分析比较这三个事件序列的特点，分析第第三个事件序列中，老鼠是获得电击还是奶酪？



假如个体状态 = 序列中的后三个事件（不包括电击、获得奶酪，下同），事件序列3的结果会是什么？（答案是：电击）

假如个体状态 = 亮灯、响铃和拉电闸各自事件发生的次数，那么事件序列3的结果又是什么？（奶酪）

假如个体状态 = 完整的事件序列，那结果又是什么？（未知）

- **完全可观测的环境 Fully Observable Environments**

个体能够直接观测到环境状态。在这种条件下：

个体对环境的观测 = 个体状态 = 环境状态

正式地说，这种问题是一个马儿可夫决定过程（Markov Decision Process， MDP）

- **部分可观测的环境 Partially Observable Environments**

个体间接观测环境。举了几个例子：

1. 一个可拍照的机器人个体对于其周围环境的观测并不能说明其绝度位置，它必须自己去估计自己的绝对位置，而绝对位置则是非常重要的环境状态特征之一；
2. 一个交易员只能看到当前的交易价格；
3. 一个扑克牌玩家只能看到自己的牌和其他已经出过的牌，而不知道整个环境（包括对手的牌）状态。

在这种条件下：

个体状态 ≠ 环境状态

正式地说，这种问题是一个部分可观测马儿可夫决策过程。个体必须构建它自己的状态呈现形式，比如：记住完整的历史：
$S_t^a = H_t$

这种方法比较原始、幼稚。还有其他办法，例如 ：

1. Beliefs of environment state：此时虽然个体不知道环境状态到底是什么样，但个体可以利用已有经验（数据），用各种个体已知状态的概率分布作为当前时刻的个体状态的呈现：

$$S_t^a = (\mathbb{P}[S_t^e = s^1], ..., \mathbb{P}[S_t^e = s^n])$$

2. Recurrent neural network：不需要知道概率，只根据当前的个体状态以及当前时刻个体的观测，送入循环神经网络(RNN)中得到一个当前个体状态的呈现：

$$S_t^a = \sigma(S_{t-1}^a W_s + O_t W_o)$$

强化学习个体的主要组成部分

强化学习中的个体可以由以下三个组成部分中的一个或多个组成：

- **策略 Policy**

策略是决定个体行为的机制。是从状态到行为的一个映射，可以是确定性的，也可以是不确定性的。

- **价值函数 Value Function**

是一个未来奖励的预测，用来评价当前状态的好坏程度。当面对两个不同的状态时，个体可以用一个Value值来评估这两个状态可能获得的最终奖励区别，继而指导选择不同的行为，即制定不同的策略。同时，一个价值函数是基于某一个特定策略的，不同的策略下同一状态的价值并不相同。某一策略下的价值函数用下式表示：

$$v_\pi(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... \mid S_t = s \right]$$

这里暂不对此公式进行详细解释。

- **模型 Model**

个体对环境的一个建模，它体现了个体是如何思考环境运行机制的（how the agent think what the environment was.），个体希望模型能模拟环境与个体的交互机制。

模型至少要解决两个问题：一是状态转化概率，即预测下一个可能状态发生的概率：

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

另一项工作是预测可能获得的即时奖励：

$$\mathcal{R}_s^a = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$$

模型并不是构建一个个体所必需的，很多强化学习算法中个体并不试图（依赖）构建一个模型。

注：模型仅针对个体而言，环境实际运行机制不称为模型，而称为环境动力学(dynamics of environment)，它能够明确确定个体下一个状态和所得的即时奖励。

强化学习个体的分类

解决强化学习问题，个体可以有多种工具组合，比如通过建立对状态的价值的估计来解决问题，或者通过直接建立对策略的估计来解决问题。这些都是个体可以使用的工具箱里的工具。因此，根据个体内包含的“工具”进行分类，可以把个体分为如下三类：

1. 仅基于价值函数的 Value Based：在这样的个体中，有对状态的价值估计函数，但是没有直接的策略函数，策略函数由价值函数间接得到。
2. 仅直接基于策略的 Policy Based：这样的个体中行为直接由策略函数产生，个体并不维护一个对各状态价值的估计函数。
3. 演员-评判家形式 Actor-Critic：个体既有价值函数、也有策略函数。两者相互结合解决问题。

此外，根据个体在解决强化学习问题时是否建立一个对环境动力学的模型，将其分为两大类：

1. 不基于模型的个体: 这类个体并不视图了解环境如何工作，而仅聚焦于价值和/或策略函数。
2. 基于模型的个体：个体尝试建立一个描述环境运作过程的模型，以此来指导价值或策略函数的更新。

学习和规划 Learning & Planning

- 学习：环境初始时是未知的，个体不知道环境如何工作，个体通过与环境进行交互，逐渐改善其行为策略。
- 规划: 环境如何工作对于个体是已知或近似已知的，个体并不与环境发生实际的交互，而是利用其构建的模型进行计算，在此基础上改善其行为策略。

一个常用的强化学习问题解决思路是，先学习环境如何工作，也就是了解环境工作的方式，即学习得到一个模型，然后利用这个模型进行规划。

探索和利用 Exploration & Exploitation

强化学习类似于一个试错的学习，个体需要从其与环境的交互中发现一个好的策略，同时又不至于在试错的过程中丢失太多的奖励。探索和利用是个体进行决策时需要平衡的两个方面。

一个形象的比方是，当你去一个餐馆吃饭，“探索”意味着你对尝试新餐厅感兴趣，很可能会去一家以前没有去过的新餐厅体验，“利用”则意味着你就在以往吃过的餐厅中挑一家比较喜欢的，而不去尝试以前没去过的餐厅。这两种做法通常是一对矛盾，但对解决强化学习问题又都非常重要。

其它一些例子，在线广告推广时，显示最受欢迎的广告和显示一个新的广告；油气开采时选择一个已知的最好的地点同在未知地点进行开采；玩游戏时选择一个你认为最好的方法同实验性的采取一个新的方法。

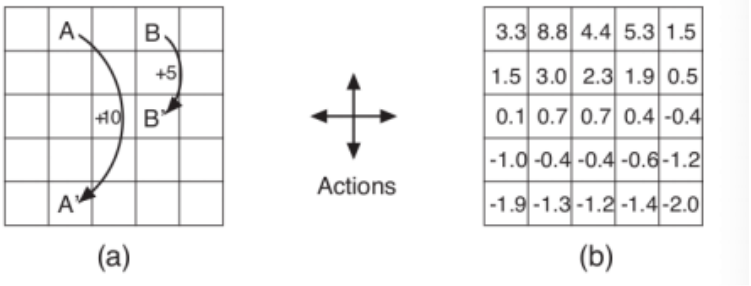
预测和控制 Prediction & Control

在强化学习里，我们经常需要先解决关于预测（prediction）的问题，而后在此基础上解决关于控制（Control）的问题。

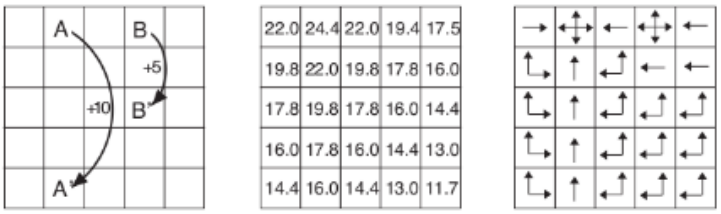
- 预测：给定一个策略，评价未来。可以看成是求解在给定策略下的价值函数（value function）的过程。How well will I(an agent) do if I(the agent) follow a specific policy?
- 控制：找到一个好的策略来最大化未来的奖励。

举了一个例子来说明预测和控制的区别。

预测：现在给出了从A到A’的奖励以及从B到B’的奖励，在“随机选择4个方向进行移动”的策略下，如何得知每一个位置的价值。



控制：同样的条件，在所有可能的策略下最优的价值函数是什么？最优策略是什么？



课程提纲

整个视频公开课分为十讲，分为两个部分。其中前5讲是第一部分，偏重于基础理论；后5讲是第二部分，偏重于解决大规模问题的应用理论。每讲具体主题如下：

第一部分：强化学习基础理论

1. 强化学习简介： 本讲

2. 马儿可夫决策过程： 理论基础，对于描述强化学习问题很重要

3. 动态规划 小规模强化学习问题的一种解决方案

4. 不基于模型的预测 理论核心

5. 不基于模型的控制 全课重点及核心

第二部分：实践中的强化学习

6. 价值函数的近似表示 基于价值函数解决大规模问题的常用技巧

7. 策略梯度方法 基于策略本身解决大规模问题时的常用技巧

8. 整合学习与规划 联合模型解决大规模问题

9. 探索和利用 理论介绍如何平衡探索和利用

10. 案例学习（选） 强化学习在游戏（博弈）中的应用

编辑于 2017-08-09
「真诚赞赏，手留余香」
赞赏