HZ BOOKS
华章IT

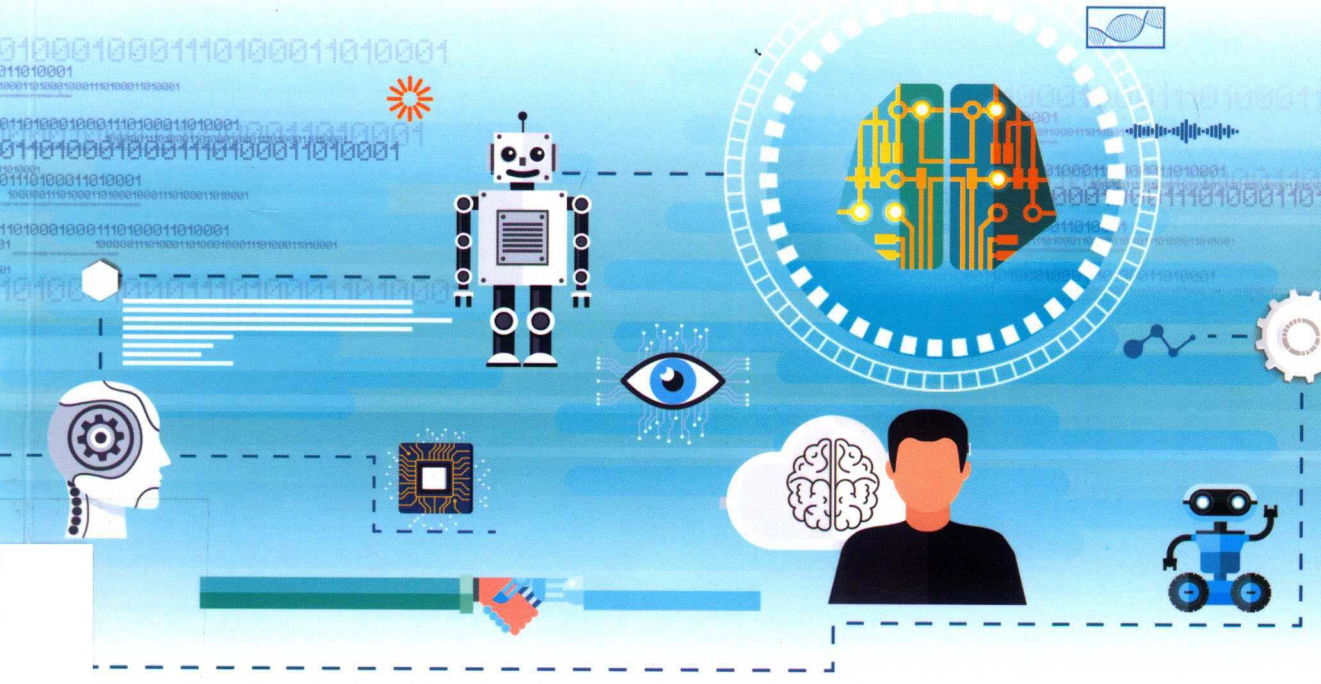
智能系统与技术丛书

理论完备，涵盖主流经典强化学习算法和深度强化学习算法
实战性强，基于Python、Gym、TensorFlow 2、AlphaZero等构建，配套代码与综合案例

Reinforcement Learning
Theory and Python Implementation

强化学习
原理与Python实现

肖智清 著



机械工业出版社
China Machine Press

智能系统与技术丛书

Reinforcement Learning
Theory and Python Implementation

强化学习
原理与Python实现

肖智清 著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

强化学习：原理与 Python 实现 / 肖智清著 . —北京：机械工业出版社，2019.8
（智能系统与技术丛书）

ISBN 978-7-111-63177-4

I. 强… II. 肖… III. 软件工具 - 程序设计 IV. TP311.561

中国版本图书馆 CIP 数据核字（2019）第 138489 号

强化学习：原理与 Python 实现

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：高婧雅 责任校对：殷 虹
印 刷：北京文昌阁彩色印刷有限责任公司 版 次：2019 年 8 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：15.5
书 号：ISBN 978-7-111-63177-4 定 价：89.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 读者信箱：hzit@hzbook.com

PREFACE

前　言

强化学习正在改变人类社会的方方面面：基于强化学习的游戏 AI 已经在围棋、星际争霸等游戏上战胜人类顶尖选手，基于强化学习的控制算法已经运用于机器人、无人机等设备，基于强化学习的交易算法已经部署在金融平台上并取得超额收益。由于同一套强化学习代码在使用同一套参数的情况下能解决多个看起来毫无关联的问题，所以强化学习常被认为是迈向通用人工智能的重要途径。在此诚邀相关专业人士研究强化学习，以立于人工智能的时代之巅。

内容梗概

本书介绍强化学习理论及其 Python 实现，全书分为三个部分。

- 第 1 章：介绍强化学习的基础知识与环境库 Gym 的使用，并给出一个完整的编程实例。
- 第 2 ～ 9 章：介绍强化学习的理论和算法。采用严谨的数学语言，推导强化学习的基本理论，进而在理论的基础上讲解算法，并为算法提供配套的 Python 实现。算法的讲解和 Python 实现逐一对应，覆盖了所有主流的强化学习算法。
- 第 10 ～ 12 章：介绍多个综合案例，包括电动游戏、棋盘游戏和自动驾驶。环境部分涵盖 Gym 库的完整安装和自定义扩展，也包括 Gym 库以外的环境。算法部分涵盖了《自然》《科学》等权威期刊发表的多个深度强化学习明星算法。

本书特色

本书完整地介绍了主流的强化学习理论。

- 全书采用完整的数学体系，各章内容循序渐进，严谨地讲授强化学习的理论基础，主要定理均给出证明过程。基于理论讲解强化学习算法，覆盖了所有主流强化学习算法，包括资格迹等经典算法和深度确定性梯度策略等深度强化学习算法。
- 全书采用一致的数学符号，并且与权威强化学习教程（如 R. Sutton 等的《 Reinforcement Learning: An Introduction（第 2 版）》和 D. Silver 的视频课程）完美兼容。

本书各章均提供 Python 代码，实战性强。

- 全书代码统一规范，基于最新的 Python 3.7（兼容 Python 3.6）、Gym 0.12 和 TensorFlow 2（兼容 TensorFlow 1）实现强化学习算法。所有代码在 Windows、macOS 和 Linux 三大操作系统上均可运行，书中给出了环境的安装和配置方法。
- 涉及环境全面。第 1 ～ 9 章提供算法的配套实现，强化学习环境只依赖于 Gym 的最小安装，使理论学习免受环境安装困扰；第 10 ～ 12 章的综合案例既涵盖 Gym 库的完整安装和自定义扩展，还包括 Gym 库以外的环境，让读者体验更加复杂的强化学习任务。
- 全书实现对硬件配置要求低。第 1 ～ 9 章代码在没有 GPU 的计算机上也可运行；第 10 ～ 12 章代码在配置普通 GPU 的计算机上即可运行。

代码下载和技术支持

本书代码下载地址为：http://github.com/zhiqingxiao/rl-book。笔者会不定期更新代码，以适应软件版本的升级。

在此推荐你加入本书学习交流 QQ 群：935702193。如果有任何意见、建议或经过网络搜索仍不能解决的问题，可以在 QQ 群里提问。笔者的邮箱是：xzq.xiaozhiqing@gmail.com。

致谢

在此感谢为本书出版做出贡献的所有工作人员。其中，机械工业出版社的高婧雅女士是本书的责任编辑，她对本书的写作提出了很多建设性意见。同时，还要感谢机械工业出版社的其他编辑为提升本书质量所做的大量工作，与他们合作是一个愉快的过程。我要特别感谢我的父亲肖林进和母亲许丽平，他们也参与了本书的编写。同时，还要感谢我的上级、同事和其他亲友，他们在本书写作期间给予我极大的支持。

感谢你选择本书。祝你学习快乐！

CONTENTS

目　录

CHAPTER 1

第 1 章

初识强化学习

强化学习（Reinforcement Learning，简称 RL，又译为“增强学习”）这一名词来源于行为心理学，表示生物为了趋利避害而更频繁实施对自己有利的策略。例如，我每天工作中会根据策略决定做出各种动作。如果我的某种决定使我升职加薪，或者使我免遭处罚，那么我在以后的工作中会更多采用这样的策略。据此，心理学家 Ivan Pavlov 在 1927 年发表的专著中用“强化”（reinforcement）这一名词来描述特定刺激使生物更趋向于采用某些策略的现象。强化行为的刺激可以称为“强化物”（reinforcer）。因为强化物导致策略的改变称为“强化学习”。

心理学家 Jack Michael 于 1975 年发表文章《Positive and negative reinforcement, a distinction that is no longer necessary》，说明了强化包括正强化（positive reinforcement）和负强化（negative reinforcement），其中正强化使得生物趋向于获得更多利益，负强化使得生物趋向于避免损害。在前面例子中，升职加薪就是正强化，免遭处罚就是负强化。正强化和负强化都能够起到强化的效果。

人工智能（Artificial Intelligence，AI）领域中有许多类似的趋利避害的问题。例如，著名的围棋 AI 程序 AlphaGo 可以根据不同的围棋局势下不同的棋。如果它下得好，它就会赢；如果下得不好，它就会输。它根据下棋的经验不断改进自己的棋艺，这就和行为心理学中的情况如出一辙。所以，人工智能借用了行为心理学的这一概念，把与环境交互中趋利避害的学习过程称为强化学习。

本章介绍人工智能领域中强化学习的基础知识，阐述强化学习的学习方法，并给出强化学习中智能体和环境交互的编程实例。

1.1 强化学习及其关键元素

在人工智能领域中，强化学习是一类特定的机器学习问题。在一个强化学习系统中，决策者可以观察环境，并根据观测做出行动。在行动之后，能够获得奖励。强化学习通过与环境的交互来学习如何最大化奖励。例如，一个走迷宫的机器人在迷宫里游荡（见

图 1-1）。机器人观察周围的环境，并且根据观测来决定如何移动。错误的移动会让机器人浪费宝贵的时间和能量，正确的移动会让机器人成功走出迷宫。在这个例子中，机器人的移动就是它根据观测而采取的行动，浪费的时间能量和走出迷宫的成功就是给机器人的奖励（时间能量的浪费可以看作负奖励）。

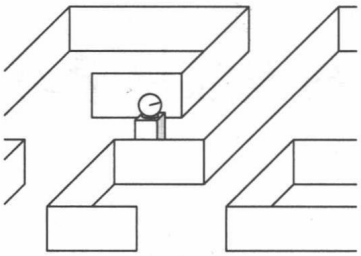


图 1-1　机器人走迷宫

强化学习的最大特点是在学习过程中没有正确答案，而是通过奖励信号来学习。在机器人走迷宫的例子中，机器人不会知道每次移动是否正确，只能通过花费的时间能量以及是否走出迷宫来判断移动的合理性。

一个强化学习系统中有两个关键元素：奖励和策略。

- **奖励**（reward）：奖励是强化学习系统的学习目标。学习者在行动后会接收到环境发来的奖励，而强化学习的目标就是要最大化在长时间里的总奖励。在机器人走迷宫的例子中，机器人花费的时间和能量就是负奖励，机器人走出迷宫就可以得到正奖励。
- **策略**（policy）：决策者会根据不同的观测决定采用不同的动作，这种从观测到动作的关系称为策略。强化学习的学习对象就是策略。强化学习通过改进策略以期最大化总奖励。策略可以是确定性的，也可以不是确定性的。在机器人走迷宫的例子中，机器人根据当前的策略来决定如何移动。

强化学习试图修改策略以最大化奖励。例如，机器人在学习过程中不断改进策略，使得以后能更快更省事地走出迷宫。

强化学习与监督学习和非监督学习有着本质的区别。

- 强化学习与监督学习的区别在于：对于监督学习，学习者知道每个动作的正确答案是什么，可以通过逐步比对来学习；对于强化学习，学习者不知道每个动作的正确答案，只能通过奖励信号来学习。强化学习要最大化一段时间内的奖励，需要关注更加长远的性能。与此同时，监督学习希望能将学习的结果运用到未知的数据，要求结果可推广、可泛化；强化学习的结果却可以用在训练的环境中。所以，监督学习一般运用于判断、预测等任务，如判断图片的内容、预测股票价格等；而强化学习不适用于这样的任务。

- 强化学习与非监督学习的区别在于：非监督学习旨在发现数据之间隐含的结构；而强化学习有着明确的数值目标，即奖励。它们的研究目的不同。所以，非监督学习一般用于聚类等任务，而强化学习不适用于这样的任务。

1.2 强化学习的应用

基于强化学习的人工智能已经有了许多成功的应用。本节将介绍强化学习的一些成功案例，让你更直观地理解强化学习，感受强化学习的强大。

- 电动游戏：电动游戏，主要指玩家需要根据屏幕画面的内容进行操作的游戏，包括主机游戏吃豆人（PacMan，见图 1-2）、PC 游戏星际争霸（StarCraft）、手机游戏 Flappy Bird 等。很多游戏需要得到尽可能高的分数，或是要在多方对抗中获得胜利。同时，对于这些游戏，很难获得在每一步应该如何操作的标准答案。从这个角度看，这些游戏的游戏 AI 需要使用强化学习。基于强化学习，研发人员已经开发出了许多强大的游戏 AI，能够超越人类能够得到的最佳结果。例如，在主机 Atari 2600 的数十个经典游戏中，基于强化学习的游戏 AI 已经在将近一半的游戏中超过人类的历史最佳结果。

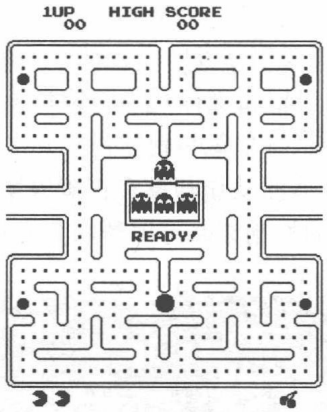


图 1-2 街机游戏吃豆人（本图片改编自 https://en.wikipedia.org/wiki/Pac-Man#Gameplay）

- 棋盘游戏：棋盘游戏是围棋（见图 1-3）、黑白翻转棋、五子棋等桌上游戏的统称。通过强化学习可以实现各种棋盘运动的 AI。棋盘 AI 有着明确的目标——提高胜率，但是每一步往往没有绝对正确的答案，这正是强化学习所针对的场景。Deepmind 公司使用强化学习研发出围棋 AI AlphaGo，于 2016 年 3 月战胜围棋顶尖选手李世石，于 2017 年 5 月战胜排名世界第一的围棋选手柯洁，引起了全社会的关注。截至目前，最强的棋盘游戏 AI 是 DeepMind 在 2018 年 12 月发表的 AlphaZero，它可以在围棋、日本将棋、国际象棋等多个棋盘游戏上达到最高水平，并远远超出人类的最高水平。

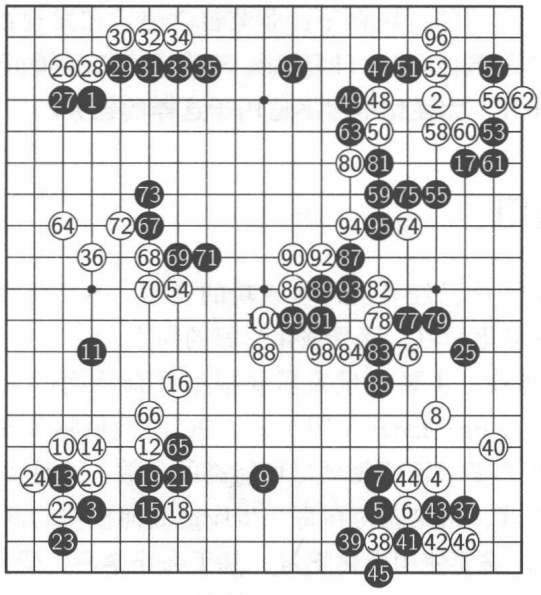


图 1-3　一局围棋棋谱（图中实心圆表示黑棋的棋子，空心圆表示白棋的棋子，圆里的数字记录棋子是在第几步被放在棋盘上，本图片改编自论文 D. Silver, et al. Mastering the game of Go without human knowledge, Nature, 2017）

- 自动驾驶：自动驾驶问题通过控制方向盘、油门、刹车等设备完成各种运输目标（见图 1-4）。自动驾驶问题既可以在虚拟环境中仿真（比如在电脑里仿真），也可能在现实世界中出现。有些任务往往有着明确的目标（比如从一个指定地点到达另外一个指定地点），但是每一个具体的动作却没有正确答案作为参考。这正是强化学习所针对的任务。基于强化学习的控制策略可以帮助开发自动驾驶的算法。

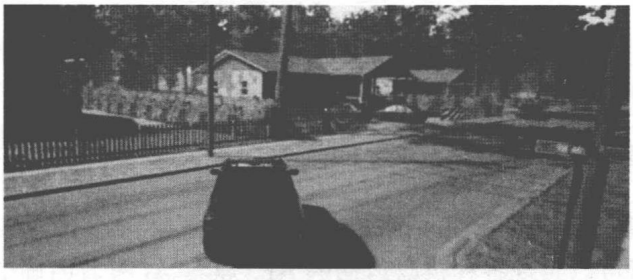


图 1-4　自动驾驶（本图截取自仿真平台 AirSimNH）

1.3　智能体 / 环境接口

强化学习问题常用**智能体 / 环境接口**（Agent-Environment Interface）来研究（见图 1-5）。

智能体 / 环境接口将系统划分为智能体和环境两个部分。

- **智能体**（agent）是强化学习系统中的决策者和学习者，它可以做出决策和接受奖励信号。一个强化学习系统里可以有一个或多个智能体。我们并不需要对智能体本身进行建模，只需要了解它在不同环境下可以做出的动作，并接受奖励信号。
- **环境**（environment）是强化系统中除智能体以外的所有事物，它是智能体交互的对象。环境本身可以是确定性的，也可以是不确定性的。环境可能是已知的，也可能是未知的。我们可以对环境建模，也可以不对环境建模。

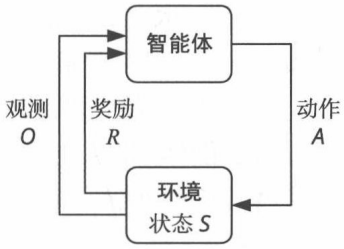


图 1-5　智能体 / 环境接口

智能体 / 环境接口的核心思想在于分隔主观可以控制的部分和客观不能改变的部分。例如，在工作的时候，我是决策者和学习者。我可以决定自己要做什么，并且能感知到获得的奖励。我的决策部分和学习部分就是智能体。同时，我的健康状况、困倦程度、饥饿状况则是我不能控制的部分，这部分则应当视作环境。我可以根据我的健康状况、困倦程度和饥饿状况来进行决策。

注意：强化学习问题不一定要借助智能体 / 环境接口来研究。

在智能体 / 环境接口中，智能体和环境的交互主要有以下三个环节：

- 智能体观测环境，可以获得环境的**观测**（observation），记为 O；
- 智能体根据观测做出决策，决定要对环境施加的**动作**（action），记为 A；
- 环境受智能体动作的影响，改变自己的**状态**（state），记为 S，并给出**奖励**（reward），记为 R。

在这三个环节中，观测 O、动作 A 和奖励 R 是智能体可以直接观测到的。

注意：状态、观测、动作不一定是数量（例如标量或矢量），也可以是“感觉到饿”、“吃饭”这样一般的量。在本书中用无衬线字体表示这样的量。奖励总是数量（而且往往是数量中的标量），本书中用衬线字体表示数量（包括标量或矢量）。

绝大多数的强化学习问题是按时间顺序或因果顺序发生的问题。这类问题的特点是具有先后顺序，并且先前的状态和动作会影响后续的状态等。例如，在玩电脑游戏时，游戏随着时间不断进行，之前玩家的每个动作都可能会影响后续的局势。对于这样的问题，我们可以引入时间指标 t，记 t 时刻的状态为 S_t，观测为 O_t，动作为 A_t，奖励为 R_t。

注意：用智能体 / 环境接口建模的问题并不一定要建模成和时间有关的问题。有些问题一共只需要和环境交互一次，就没有必要引入时间指标。例如，以不同的方式投掷一个给定的骰子并以点数作为奖励，就没有必要引入时间指标。

在很多任务中，智能体和环境是在离散的时间步骤上交互的，这样的问题可以将时间指标离散化，建模为离散时间智能体 / 环境接口。具体而言，假设交互的时间为 $t=0,1,2,3,\ldots$。在时刻 t，依次发生以下事情：

- 智能体观察环境得到观测 O_t；
- 智能体根据观测决定做出动作 A_t；
- 环境根据智能体的动作，给予智能体奖励 R_{t+1} 并进入下一步的状态 S_{t+1}。

注意：智能体 / 环境接口问题不一定能时间上离散化。有些问题在时间上是连续的，需要使用偏微分方程来建模环境。连续时间的问题也可以近似为离散时间的问题。

在智能体 / 环境接口的基础上，研究人员常常将强化学习进一步建模为 Markov 决策过程。本书第 2 章会介绍 Markov 决策过程。

1.4　强化学习的分类

强化学习的任务和算法多种多样，本节介绍一些常见的分类（见图 1-6）。

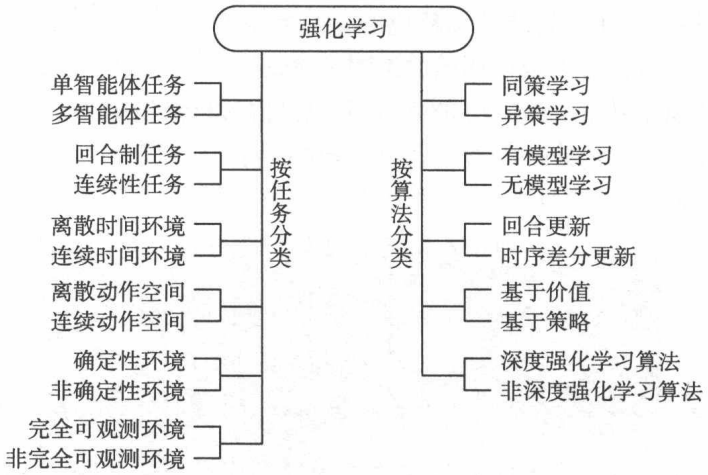


图 1-6　强化学习的分类

1.4.1　按任务分类

根据强化学习的任务和环境，可以将强化学习任务作以下分类。

- **单智能体任务**（single agent task）和**多智能体任务**（multi-agent task）：顾名思义，根据系统中的智能体数量，可以将任务划分为单智能体任务和多智能体任务。单智能体任务中只有一个决策者，它能得到所有可以观察到的观测，并能感知全局的奖励

值；多智能体任务中有多个决策者，它们只能知道自己的观测，感受到环境给它的奖励。当然，在有需要的情况下，多个智能体间可以交换信息。在多智能体任务中，不同智能体奖励函数的不同会导致它们有不同的学习目标（甚至是互相对抗的）。在本书没有特别说明的情况下，一般都是指单智能体任务。

- **回合制任务**（episodic task）和**连续性任务**（sequential task）：对于回合制任务，可以有明确的开始状态和结束状态。例如在下围棋的时候，刚开始棋盘空空如也，最后棋盘都摆满了，一局棋就可以看作是一个回合。下一个回合开始时，一切重新开始。也有一些问题没有明确的开始和结束，比如机房的资源调度。机房从启用起就要不间断地处理各种信息，没有明确的结束又重新开始的时间点。
- **离散时间环境**（discrete time environment）和**连续时间环境**（continuous time environment）：如果智能体和环境的交互是分步进行的，那么就是离散时间环境。如果智能体和环境的交互是在连续的时间中进行的，那么就是连续时间环境。
- **离散动作空间**（discrete action space）和**连续动作空间**（continuous action space）：这是根据决策者可以做出的动作数量来划分的。如果决策得到的动作数量是有限的，则为离散动作空间，否则为连续动作空间。例如，走迷宫机器人如果只有东南西北这 4 种移动方式，则其为离散动作空间；如果机器人向 360° 中的任意角度都可以移动，则为连续动作空间。
- **确定性环境任务**（deterministic environment）和**非确定性环境**（stochastic environment）：按照环境是否具有随机性，可以将强化学习的环境分为确定性环境和非确定性环境。例如，对于机器人走固定的某个迷宫的问题，只要机器人确定了移动方案，那么结果就总是一成不变的。这样的环境就是确定性的。但是，如果迷宫会时刻随机变化，那么机器人面对的环境就是非确定性的。
- **完全可观测环境**（fully observable environment）和**非完全可观测环境**（partially observable environment）：如果智能体可以观测到环境的全部知识，则环境是完全可观测的；如果智能体只能观测到环境的部分知识，则环境是非完全可观测的。例如，围棋问题就可以看作是一个完全可观测的环境，因为我们可以看到棋盘的所有内容，并且假设对手总是用最优方法执行；扑克则不是完全可观测的，因为我们不知道对手手里有哪些牌。

1.4.2 按算法分类

从算法角度，可以对强化学习算法作以下分类。

- **同策学习**（on policy）和**异策学习**（off policy）：同策学习是边决策边学习，学习者同时也是决策者。异策学习则是通过之前的历史（可以是自己的历史也可以是别人的历史）进行学习，学习者和决策者不需要相同。在异策学习的过程中，学习者并不一定要知道当时的决策。例如，围棋 AI 可以边对弈边学习，这就算同策学习；围

棋 AI 也可以通过阅读人类的对弈历史来学习，这就算异策学习。

- **有模型学习**（model-based）和**无模型学习**（model free）：在学习的过程中，如果用到了环境的数学模型，则是有模型学习；如果没有用到环境的数学模型，则是无模型学习。对于有模型学习，可能在学习前环境的模型就已经明确，也可能环境的模型也是通过学习来获得。例如，对于某个围棋 AI，它在下棋的时候可以在完全了解游戏规则的基础上虚拟出另外一个棋盘并在虚拟棋盘上试下，通过试下来学习。这就是有模型学习。与之相对，无模型学习不需要关于环境的信息，不需要搭建假的环境模型，所有经验都是通过与真实环境交互得到。
- **回合更新**（Monte Carlo update）和**时序差分更新**（temporal difference update）：回合制更新是在回合结束后利用整个回合的信息进行更新学习；而时序差分更新不需要等回合结束，可以综合利用现有的信息和现有的估计进行更新学习。
- **基于价值**（value based）和**基于策略**（policy based）：基于价值的强化学习定义了状态或动作的价值函数，来表示到达某种状态或执行某种动作后可以得到的回报。基于价值的强化学习倾向于选择价值最大的状态或动作；基于策略的强化学习算法不需要定义价值函数，它可以为动作分配概率分布，按照概率分布来执行动作。
- **深度强化学习**（Deep Reinforcement Learning，DRL）算法和非深度强化学习算法。如果强化学习算法用到了深度学习，则这种强化学习可以称为深度强化学习算法。

值得一提的是，强化学习和深度学习是两个独立的概念。一个学习算法是不是强化学习和它是不是深度学习算法是相互独立的（见图 1-7）。如果一个算法解决了强化学习的问题，这个算法就是强化学习的算法；如果一个算法用到了深度神经网络，这个算法就是深度学习算法。一个强化学习算法可以是深度学习算法，也可以不是深度学习算法；一个深度学习算法可以是强化学习算法，也可以不是强化学习算法。对于强化学习算法而言，在问题规模比较小时，能够获得精确解；当问题规模比较大时，常常使用近似的方法。深度学习则利用神经网络来近似复杂的输入 / 输出关系。对于规模比较大的强化学习问题，可以考虑利用深度学习来实现近似。如果一个算法既是强化学习算法，又是深度学习算法，则可以称它是深度强化学习算法。例如，很多电动游戏 AI 需要读取屏幕显示并据此做出决策。对屏幕数据的解读可以采用卷积神经网络这一深度学习算法。这时，这个 AI 就用到了深度强化学习算法。

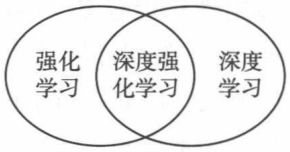


图 1-7　强化学习与深度学习之间的关系

1.5　如何学习强化学习

本节介绍强化学习需要的预备知识，以及如何学习强化学习，本节中还提供了一些参考资料。