机器学习系列

Packt

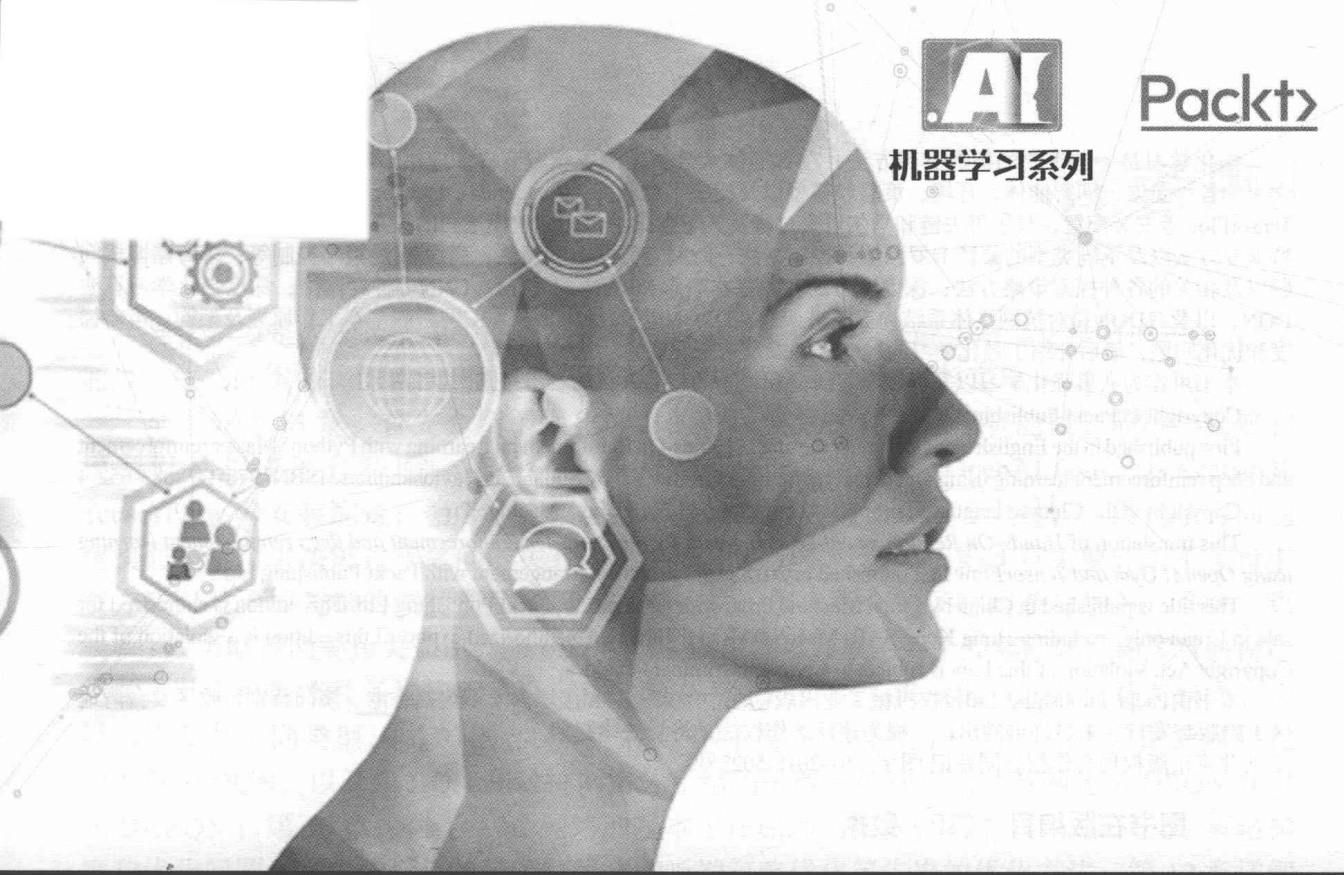
Hands-On Reinforcement Learning with Python :
Master reinforcement and deep reinforcement learning using OpenAI Gym and TensorFlow

Python 强化学习实战

应用OpenAI Gym和TensorFlow精通强化学习和深度强化学习

[印度]苏达桑·拉维尚迪兰（Sudharsan Ravichandiran） 著
连晓峰 等译

机械工业出版社
CHINA MACHINE PRESS



机器学习系列

Packt

Hands-On Reinforcement Learning with Python :
Master reinforcement and deep reinforcement learning using OpenAI Gym and TensorFlow

Python 强化学习实战

应用OpenAI Gym和TensorFlow精通强化学习和深度强化学习

[印度]苏达桑·拉维尚迪兰（Sudharsan Ravichandiran） 著
连晓峰 等译

机械工业出版社
CHINA MACHINE PRESS

强化学习是一种重要的机器学习方法，在智能体及分析预测等领域有许多应用。本书共13章，主要包括强化学习的各种要素，即智能体、环境、策略和模型以及相应平台和库；Anaconda、Docker、OpenAI Gym、Universe和TensorFlow等安装配置；马尔可夫链和马尔可夫过程及其与强化学习问题建模之间的关系，动态规划的基本概念；蒙特卡罗方法以及不同类型的蒙特卡罗预测和控制方法；时间差分学习、预测、离线/在线策略控制等；多臂赌博机问题以及相关的各种探索策略方法；深度学习的各种基本概念和RNN、LSTM、CNN等神经网络；深度强化学习算法DQN，以及双DQN和对抗网络体系结构等改进架构；DRQN以及DARQN；A3C网络的基本工作原理及架构；策略梯度和优化问题；最后介绍了强化学习的最新进展以及未来发展。

本书可作为从事强化学习以及深度学习研究和教学的相关人员的参考用书。

图书在版编目（CIP）数据

Python 强化学习实战：应用 OpenAI Gym 和 TensorFlow 精通强化学习和深度强化学习 /（印）苏达桑·拉维尚迪兰著；连晓峰等译 . —北京：机械工业出版社，2018.11

（机器学习系列）

书名原文：Hands-On Reinforcement Learning with Python : Master reinforcement and deep reinforcement learning using OpenAI Gym and TensorFlow

ISBN 978-7-111-61288-9

Ⅰ . ① P… Ⅱ . ①苏… ②连… Ⅲ . ①软件工具 – 程序设计 Ⅳ . ① TP311.561

中国版本图书馆 CIP 数据核字（2018）第 249823 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）

策划编辑：顾 谦 责任编辑：顾 谦

责任校对：李 杉 责任印制：李 昂

北京宝昌彩色印刷有限公司印刷

2019 年 1 月第 1 版第 1 次印刷

184mm × 240mm · 13.5 印张 · 300 千字

0001—4000 册

标准书号：ISBN 978-7-111-61288-9

定价：69.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务 网络服务

服务咨询热线：010-88361066 机 工 官 网 ：www.cmpbook.com

读者购书热线：010-68326294 机 工 官 博 ：weibo.com/cmp1952

010-88379203 金 书 网 ：www.golden-book.com

封面无防伪标均为盗版 教育服务网 ：www.cmpedu.com

译者序

强化学习是人工智能相关的前沿领域。本书主要针对零基础学习强化学习的相关人员。通过一些实际案例，深入浅出地介绍了强化学习相关的理论知识与实现。

全书分为 13 章：首先介绍了强化学习的各种要素，其中包括智能体、环境、策略和模型以及相应平台和库；第 2 章学习了 Anaconda、Docker、OpenAI Gym、Universe 和 TensorFlow 等安装配置；第 3 章介绍了什么是马尔可夫链和马尔可夫过程、强化学习问题建模与马尔可夫决策过程之间的关系、动态规划的基本概念等；第 4 章主要介绍了蒙特卡罗方法以及不同类型的蒙特卡罗预测和控制方法；第 5 章介绍了时间差分学习、预测、离线 / 在线策略控制等相关知识；第 6 章介绍了强化学习领域的一个经典问题，即多臂赌博机问题，其中主要涉及各种探索策略方法；第 7 章主要介绍了深度学习的各种基本概念、神经网络及其不同类型，如 RNN、LSTM 和 CNN；第 8 章介绍了一种应用广泛的深度强化学习算法 DQN，以及双 DQN 和对抗网络体系结构等改进架构；第 9 章阐述了 DRQN 并引入 DARQN；第 10 章探讨了 A3C 网络的基本工作原理及架构；第 11 章主要是关于策略梯度和优化问题，其中包括信赖域策略优化和近端策略优化等先进的优化方法；第 12 章详细介绍了利用对抗性 DQN 构建赛车智能体的具体实现过程；第 13 章介绍了强化学习的最新进展以及未来发展。

全书内容丰富，结构合理，并通过实际项目的实现来阐述相关的理论知识，易于读者理解和掌握。作者在强化学习方面具有丰富的实践经验，同时通过各种博客科技文章的撰写积累了理论与实践结合的表述能力。

本书主要由连晓峰翻译，另外，刘鹏华、郭柯、闫峰、李林、王宇龙、赵晓平、刘萌、潘峰、叶璐、孙冬、王炎、赵旭、潘兵、贾涵、王磊等人参与了部分翻译工作。鉴于译者水平有限，难免存在不当与错误之处，恳请广大读者批评指正！

译　者

原书前言

强化学习是一种自我进化型的机器学习，可以更接近于实现真正的人工智能。本书通过 Python 语言编写的丰富示例进行了详细解释。

本书读者

本书主要针对对人工智能感兴趣，并想要零基础学习强化学习的从事机器学习开发人员和热衷于深度学习的爱好者。通过阅读本书，并通过在工作或项目中实现实际案例，将会使读者成为强化学习方面的专家。对线性代数、微积分和 Python 编程语言有一定基础将有助于理解本书的内容。

本书主要内容

第 1 章 强化学习简介，有助于理解什么是强化学习及其是如何工作的。在此将学习强化学习的各种要素，如智能体、环境、策略和模型，同时将会分析用于强化学习的不同类型的环境、平台和库。在本章结束处，还将介绍强化学习的一些应用。

第 2 章 从 OpenAI 和 TensorFlow 入门，主要是针对各种强化学习任务配置所用计算机。在此将学习如何通过安装 Anaconda、Docker、OpenAI Gym、Universe 和 TensorFlow 来配置计算机。然后，将学习如何在 OpenAI Gym 中模拟智能体，还将学习如何构建一个视频游戏机器人。另外，还将学习 TensorFlow 的基本原理，并了解如何使用 TensorBoard 进行可视化。

第 3 章 马尔可夫决策过程和动态规划，首先解释了什么是马尔可夫链和马尔可夫过程，然后将学习如何将强化学习问题建模为马尔可夫决策过程（MDP）。另外，还将了解几个基本概念，如值函数、Q 函数和 Bellman 方程。接下来，分析什么是动态规划，以及如何使用值函数和策略迭代来解决冰冻湖问题。

第 4 章 基于蒙特卡罗方法的博弈游戏，介绍了蒙特卡罗方法以及不同类型的蒙特卡罗预测方法，如首次访问蒙特卡罗和每次访问蒙特卡罗。另外，还将学习如何利用蒙特卡罗方法玩二十一点游戏。然后，探讨了不同的在线策略和离线策略的蒙特卡罗控制方法。

第 5 章 时间差分学习，其中包括时间差分（TD）学习、时间差分预测、时间差分离线策略和在线策略控制方法，如 Q 学习和 SARSA。另外，还将学习如何利用 Q 学习和 SARSA 解决出租车问题。

第 6 章 MAB 问题，其中涉及强化学习的一个经典问题：多臂赌博机（MAB）或 k 臂赌博机问题。在此将学习如何利用各种探索策略来解决该问题，如贪婪、Softmax 探索、UCB 和 Thompson 采样。在本章结束处，还将分析如何通过 MAB 向用户显示正确的广告标识。

第 7 章 深度学习基础，其中涵盖了深度学习的各种基本概念。首先，将学习什么是神

经网络，然后是不同类型的神经网络，如 RNN、LSTM 和 CNN。在此将通过构建一些诸如生成歌词和分类时尚产品等任务的应用程序来进行学习。

第 8 章 基于 DQN 的 Atari 游戏，其中介绍了一种应用最广泛的深度强化学习算法，称为深度 Q 网络（DQN）。将通过分析其各个组成部分来学习了解 DQN，然后讨论如何构建一个智能体利用 DQN 来玩 Atari 游戏。接下来，将研究一些改进的 DQN 架构，如双 DQN 和对抗网络体系结构。

第 9 章 基于 DRQN 玩 Doom 游戏，解释了深度递归 Q 网络（DRQN）及其与 DQN 的不同。在此将讨论如何构建一个智能体利用 DRQN 来玩 Doom 游戏。在本章结束处，还将学习深度注意力递归 Q 网络（DARQN），这是将注意力机制添加到 DRQN 架构中。

第 10 章 A3C 网络，解释了异步优势行为者评论家（A3C）网络的工作原理。详细探讨了 A3C 架构，并将学习如何利用 A3C 构建一个爬山的智能体。

第 11 章 策略梯度和优化，涵盖了策略梯度是如何有助于在无需 Q 函数情况下找到正确策略，还将探讨深度确定性策略梯度（DPPG）法。在本章结束处，还将学习最先进的策略优化方法，如信赖域策略优化（TRPO）和近端策略优化（PPO）。

第 12 章 Capstone 项目——基于 DQN 的赛车游戏，提供了构建一个利用对抗性 DQN 赢得赛车比赛的智能体的详细方法。

第 13 章 最新进展和未来发展，提供了有关强化学习的各种相关进展信息，如想象力增强智能体（I2A）、基于人类偏好的学习、基于演示的深度 Q 学习（DQfd）和事后经验回放（HER），然后将讨论不同类型的强化学习方法，如分层强化学习（HRL）和逆向强化学习。

更好地利用本书

阅读本书需要以下软件：

- Anaconda。
- Python。
- 任何 Web 浏览器。
- Docker。

下载示例代码文件

可以在 www.packtpub.com 上根据账户下载本书的示例代码文件。如果是从其他途径购买本书，可以访问 www.packtpub.com/support 并注册，从而可直接将文件邮件发送。

可以通过以下步骤下载代码文件：

1）在 `www.packtpub.com` 登录或注册。

2）选择 SUPPORT 标签页。

3）点击 Code Downloads & Errata。

4）在 Search box 中输入书名，并按照提示进行操作。

下载完成后，请确保采用以下最新版本进行文件夹解压缩：

- Windows 系统下的 WinRAR/7-Zip。
- Mac 系统下的 Zipeg/iZip/UnRarX。
- Linux 系统下的 7-Zip/PeaZip。

本书的代码包还在 Github 上托管，具体地址为 https://github.com//PacktPublishing/Hands- On- Reinforcement- Learning- with- Python。如果代码有更新，同时也会在现有的 Github 代码库中更新。

另外，还可以从 https://github.com/PacktPublishing/ 丰富的图书和视频目录中获得其他代码包。敬请关注！

彩页下载

本书还提供了包含书中截屏 / 图表的所有彩色图像的 PDF 文件。可以从 http://www.packtpub.com/sites/default/files/downloads/HandsOnReinforcementLearningwithPython_ColorImages.pdf 下载。

约定惯例

本书中使用了一些文本约定。

CodeInText：表示文本中的代码关键字、数据库表名、文件夹名、文件名、文件扩展名、路径名、虚拟 URL、用户输入和 Twitter 句柄。下面是一个示例："Mount the downloaded WebStorm-10*.dmg diskimage file as another disk in your system"。

代码块的设置如下：

```
policy_iteration():
    Initialize random policy
    for i in no_of_iterations:
        Q_value = value_function(random_policy)
        new_policy = Maximum state action pair from Q value
```

任何命令行输入 / 输出表示如下：

```
bash Anaconda3-5.0.1-Linux-x86_64.sh
```

粗体：表示一个新词、一个重要的词或在屏幕上看到的单词。例如，在菜单或对话框中以文本形式显示的单词。

警告或重要提示在此显示。

提示和技巧在此显示。

联系我们

欢迎读者反馈。

一般反馈：发送电子邮件到 feedback@packtpub.com，并在邮件主题注明书名。如果对本书有任何疑问，请发送邮件到 questions@packtpub.com。

勘误：尽管已尽一切努力确保内容的准确性，但难免还是有错误存在。如果在本书中发现错误，请告知我们，将不胜感激。请访问 www.packtpub.com/submit-errata，选择具体图书，点击 Errata Submission Form 链接，并输入详细信息。

盗版：如果在网上看到任何非法复制图书作品，请提供具体地址和网站名称，我们将不胜感激。请将上述材料发送到 copyright@packtpub.com。

有兴趣成为作者：如果有您擅长的主题，并有兴趣撰写图书，请访问 authors.packtpub.com。

评论

请不吝赐教。既然阅读并使用了本书，为何不在购买本书的网站上留下评论呢？潜在的读者可以根据您的客观评价决定是否购买本书，Packt 出版社完全理解您对本书的评价，同时作者也非常乐意收到您对本书的反馈意见。非常感谢！

有关 Packt 出版社的更多信息，请访问 www.packtpub.com。

作者简介

Sudharsan Ravichandiran 是一位数据科学家、研究员、人工智能爱好者以及 YouTuber（搜索 Sudharsan reinforcement learning），获得了 Anna 大学信息技术学士学位。他的研究领域包括深度学习和强化学习的实现，其中包括自然语言处理和计算机视觉。他曾是一名自由职业的网页开发人员和设计师，所设计开发的网站屡获殊荣，同时也热衷于开源，擅长解答堆栈溢出问题。

原书审稿人简介

Sujit Pal 是 Elsevier 实验室的技术研究总监，Elsevier 实验室是 Reed-Elservier 集团公司下的一个先进技术团队，研究领域包括语义检索、自然语言处理、机器学习和深度学习。他在 Elsevier 实验室主要从事搜索质量检测与改进、图像分类和重复率检测、医学和科学语料库的标注与本体开发。他曾与 Antonio Gulli 合作撰写了一本关于深度学习的著作，并在博客 Slamon Run 上撰写了一些科技文章。

Suriyadeepan Ramamoorthy 是一名来自印度 Puducherry 的 AI 研究人员和工程师，主要研究领域是自然语言理解和推理，同时积极撰写有关深度学习的博客文章。在 SAAMA 技术中，他将先进的深度学习技术应用于生物医学文本分析，同时也是一名积极推动 FSFTN 领域发展的免费软件宣传者，另外对社交网络、数据可视化和创造性编程也非常感兴趣。

目 录

第 1 章
强化学习简介

强化学习（RL）是机器学习的一个分支，其中学习是通过与环境交互而进行的。这是一种目标导向的学习，学习者并未被告知应采取何种行为，相反学习者是从其行为后果中进行学习的。随着各种算法的提出，该方法发展迅速，现已是**人工智能**（AI）方面最活跃的研究领域之一。

本章的主要内容包括：

- 强化学习的基本概念；
- 强化学习算法；
- 智能体环境接口；
- 强化学习的环境类型；
- 强化学习平台；
- 强化学习的应用。

1.1 什么是强化学习

想象一下训练小狗抓球的场景，但不能很具体地去训练狗抓球，而是仅扔出去一个球，每次狗抓住球，就会奖赏一块饼干。如果没有抓住球，就没有奖赏。这样狗就会明白哪些行为可以使之得到饼干，从而不断重复执行这些动作。

同理，在强化学习环境中，也不会训练智能体做什么或怎么做，而是根据智能体的每次行为给予奖励。这种奖励可以是正面的也可以是负面的。然后，智能体将开始执行能够使之得到正面奖励的行为。因此，这是一种反复试验的过程。在上述比喻中，狗代表智能体。一旦狗抓住球就会得到正面奖励的饼干，而如果没有抓住球，不给饼干就是负面奖励。

奖励可能会有延迟，即可能不会在每一个步骤都得到奖励。奖励只能是在完成任务后才能得到。在某些情况下，每个步骤得到奖励会表明是否犯错。

想象要训练一个机器人行走且不会因遇到山坡而不知所措，但不能明确告诉机器人不要朝山的方向运动（见图 1.1）。

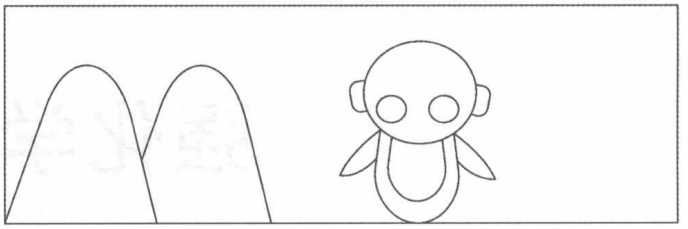


图 1.1

相反，如果机器人撞上山坡而陷入困境，那么就会扣除 10 分，这样机器人就会明白撞上山坡会产生负面奖励，从而不会再朝这个方向运动（见图 1.2）。

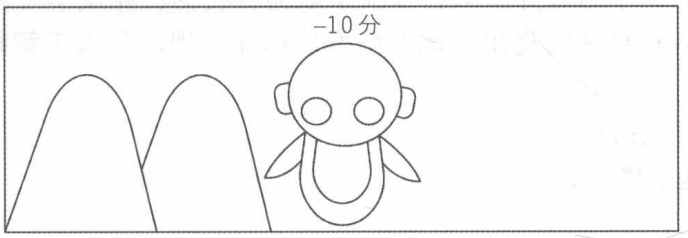


图 1.2

当机器人沿正确方向行走而不会陷入困境时，将给予 20 分。因此，机器人就会了解哪条路径是正确的，并朝着正确方向运动来尽量获得最大奖励（见图 1.3）。

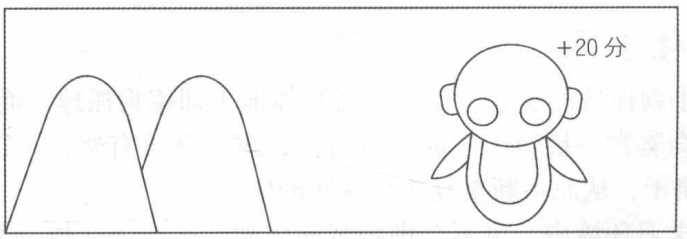


图 1.3

强化学习智能体可以**探索**可能会得到良好奖励的各种行为，或者可以**开发**（执行）能够获得良好奖励的先前行为。如果强化学习智能体探索了不同行为，那么可能会得到一个很差的奖励，这是因为所有行为都不是最佳的。如果强化学习智能体只采用了已知的最佳行为，那么也可能会错过可以获得更好奖励的最佳行为。探索和开发之间总是存在一种权衡关系。不能同时进行探索和开发。在后面的内容中将会详细讨论探索—开发问题。

1.2 强化学习算法

在典型的强化学习算法中，所涉及的步骤如下：

1）首先，智能体通过执行行为与环境进行交互。

2）智能体执行一个行为后，从一个状态转移到另一个状态。

3）然后，智能体将会根据所执行的行为获得相应奖励。

4）根据所获得的奖励，智能体会知晓该行为是好还是坏。

5）如果行为是好的，即如果智能体得到正面奖励，那么将会倾向于执行该行为，否则智能体会尝试执行其他行为来获得正面奖励。因此，该方法本质上是一个试错学习过程。

1.3 强化学习与其他机器学习范式的不同

在监督学习中，机器（智能体）从具有标记的输入 / 输出训练数据集中学习。目的是使得模型对其学习进行外推和推广，从而能够很好地适用于未知数据。在此，需要有一个对环境具备完备知识基础的外部监督者，来监督智能体完成该任务。

考虑上述讨论的狗抓球的情况：在监督学习中，为了训练狗抓球，将会通过指定向左、向右、向前移动五步、抓球等动作来明确地训练狗。但是在强化学习中，只是扔出去一个球，每次狗抓住球，就奖赏一块饼干（奖励）。所以狗将学习如何抓住球，这意味着会得到饼干。

在无监督学习中，将提供一个仅有一组输入训练数据的模型，通过模型来学习确定输入数据中的隐藏模式。通常普遍会误认为强化学习是一种无监督学习，其实并非如此。在无监督学习中，通过模型来学习隐藏的结构，而在强化学习中，通过最大化奖励来学习模型。假设想要向用户推荐新电影。无监督学习会通过分析用户观看过的类似电影来推荐，而强化学习则不断接收来自用户的反馈信息，了解其偏好电影，并在此基础上构建一个知识库，来推荐一部新电影。

另外，还有一种称为半监督的学习，其本质上是监督学习和无监督学习的结合。该方法涉及标记数据和未标记数据的函数估计，而强化学习本质上是智能体与其环境之间的交互。因此，强化学习与所有其他机器学习范式完全不同。

1.4 强化学习的要素

下面介绍强化学习的要素。

1.4.1 智能体

智能体是指进行智能决策的软件程序，在强化学习中通常是学习者。智能体通过与环境交互来执行行为，并根据其行为来获得奖励，如超级马里奥在视频游戏中的动作行为。

1.4.2 策略函数

策略定义了智能体在环境中的行为。智能体决定执行何种行为取决于策略。假设从家到办公室存在不同的路线，其中有些路线很短，而有些路线相对很长。这些路线就可以称为策略，因为代表了为达到目标而选择执行的行为。策略通常用符号 π 表示。策略也可以

是查找表形式或复杂的搜索过程。

1.4.3 值函数

值函数是表示智能体在某一特定状态下的程度。这与策略相关，通常用 $v(s)$ 表示。值函数等价于智能体从初始状态开始所获得的总的期望奖励。值函数有多种形式。最优值函数是与其他值函数相比，所有状态下具有最大值的一种值函数。同理，最优策略是指具有最优值函数的策略。

1.4.4 模型

模型是指智能体对环境的表示。学习可以分为基于模型的学习和无模型学习两种类型。在基于模型的学习中，智能体利用先前学习到的信息来完成任务，而在无模型学习中，智能体仅是通过试错经验来执行正确行为。假设想要更快地从家到办公室。在基于模型的学习中，只需利用先前学习的经验（地图）来快速到达办公室，而在无模型学习中，不会使用之前的经验，而是尝试所有不同的路线，并从中选择较快的一种。

1.5 智能体环境接口

智能体是指在时刻 t 执行行为 A_t 从一个状态 S_t 转移到另一个状态 S_{t+1} 的软件智能体。根据行为，智能体从环境获得数值型奖励 R。最终，强化学习就是寻找使得数值奖励增大的最优行为（见图 1.4）。

在此通过一个迷宫游戏来理解强化学习的概念（见图 1.5）。

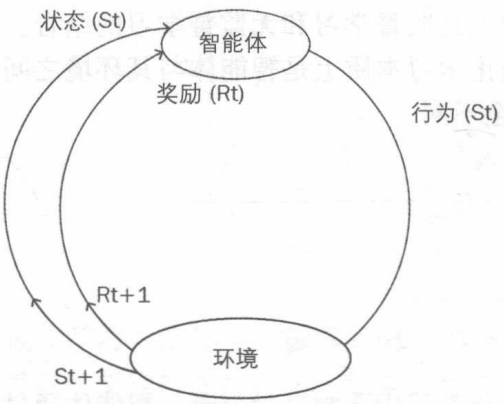


图 1.4

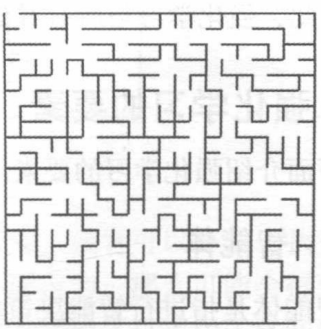


图 1.5

迷宫的目标是到达目的地而不会因障碍物而陷入困境。以下是具体工作流程：

- 智能体是指穿过迷宫的物体，即软件程序 / 强化学习算法。
- 环境是迷宫。
- 状态是智能体当前在迷宫中所处的位置。
- 智能体通过从一个状态转移到另一个状态来执行行为。
- 当智能体的行为没有因任何障碍物而困住时会得到正面奖励，而如果因障碍物而陷

入困境，导致无法到达目的地时会得到负面奖励。

- 目标是穿过迷宫，到达目的地。

1.6 强化学习的环境类型

与智能体交互的所有内容都称为环境。环境是指外部世界，包括除智能体之外的一切事物。环境有不同类型，这将在下面讨论。

1.6.1 确定性环境

如果根据当前状态就可以知道相应的结果，那么就称为确定性环境。例如，在国际象棋中，会知道移动每个棋子后的确切结果。

1.6.2 随机性环境

如果不能根据当前状态来确定相应的结果，那么这种环境就称为随机性环境。在这种环境中，会存在较大程度的不确定性。例如，在掷骰子时永远不知道会出现什么数字。

1.6.3 完全可观测环境

如果智能体在任何时候都能确定系统的状态，那么就称为完全可观测环境。例如，在国际象棋中，系统的状态，即棋盘上所有棋子的位置，都是可以获得的，因此棋手可以做出最优决策。

1.6.4 部分可观测环境

如果智能体无法在任何时候都能确定系统的状态，那么就称为部分可观测环境。例如，在玩扑克时，不知道对手的牌。

1.6.5 离散环境

如果从一个状态转移到另一个状态后只能有一个有限的行为状态，那么就称为离散环境。例如，在国际象棋中，只能有移动棋子后的有限集。

1.6.6 连续环境

如果从一个状态转移到另一个状态后可以有无限的行为状态，那么就称为连续环境。例如，在从出发地到目的地的旅行中可以有多条路线。

1.6.7 情景和非情景环境

情景环境也称为**非序贯**环境。在情景环境中，智能体的当前行为不会影响将来的行为，而在非情景环境中，智能体的当前行为会影响今后的行为，也称为**序贯**环境。也就是说，智能体在情景环境中执行独立的任务，而在非情景环境中，所有智能体的行为都是相关的。