

# 第 1 章

## 强化学习基础

强化学习(Reinforcement Learning, RL)是机器学习的一个分支,它关注的是如何让智能体(Agent)通过与环境的交互来学习并改进其行为,以达到最大化某种累积奖励信号的目标。本章将详细讲解强化学习的基础知识,介绍强化学习的应用领域和基本概念。

## 1.1 强化学习概述

在强化学习中，智能体通过不断采取行动来探索环境，观察环境对其行动的反馈，并根据这些反馈来调整其行为策略，以使其在特定任务中表现得更好。这种学习方式与监督学习不同，因为在强化学习中，智能体必须在没有明确标签或指导的情况下在试错中学习。



扫码看视频

### 1.1.1 强化学习的背景

强化学习的背景可以追溯到人工智能和机器学习领域的早期历史，下面列出了强化学习的一些关键背景和历史要点。

(1) 早期计算机科学和人工智能。强化学习的概念在早期计算机科学和人工智能领域有根深蒂固的基础。早期的研究包括理论上的探索，但当时的计算能力和数据可用性有限，因此实际应用受到了限制。

(2) 伯克斯的细胞自动机和后来其他人的学习自动机。在 20 世纪 50 年代，冯·诺伊曼的助手伯克斯(Arthur W. Burks)提出了细胞自动机的概念。M. L. Tsetlin 等人在 1961 年提出了学习自动机的数学模型，该模型称为固定结构学习自动机(FSSA)。这是一种能够通过试错来学习行为的机械装置。这个概念启发了强化学习中的探索与利用思想。

(3) 动态规划。20 世纪 50 年代和 60 年代，数学家理查德·贝尔曼(Richard Bellman)的动态规划理论为强化学习奠定了坚实的数学基础。贝尔曼方程(Bellman Equation)成为强化学习中的核心概念，用于解决马尔可夫决策过程(Markov Decision Process, MDP)等问题。

(4) TD(Temporal Difference, 时序差分)学习。在 20 世纪 80 年代，计算机科学家理查德·萨顿(Richard Sutton)和计算神经科学家戴维·明格尔(David Mingolla, 也译作大卫·明格拉)提出了 TD 学习方法，这一方法是强化学习中的重要突破，用于训练智能体从经验中学习。

(5) Q-learning(Q 学习)。Q-learning 是一种重要的强化学习算法，由克里斯托弗·沃特金斯(Christopher Watkins)于 1989 年首次提出。Q-learning 是一个基于值迭代的算法，用于找到最佳策略以最大化奖励。

(6) 深度强化学习(Deep Reinforcement Learning)。近年来，深度学习技术的兴起促进了强化学习领域的快速发展。深度强化学习结合了深度神经网络和强化学习方法，成功地应用于各种复杂任务，如图像处理、自动驾驶和游戏玩法。

(7) 应用领域扩展。强化学习已经被广泛应用于各种领域，包括机器人控制、游戏智能、自动驾驶、金融、医疗保健、自然语言处理和能源管理等。其应用还在不断扩展和深化。

总的来说，强化学习的背景涵盖了早期的数学和计算理论、心理学研究、动态规划的发展以及近年来深度学习的崛起。这些因素共同推动了强化学习领域的发展和创新，使其成为解决复杂决策问题的重要工具。

### 1.1.2 强化学习的核心特点

强化学习的核心特点有以下几个方面。

(1) 交互性(Interactivity)。在强化学习中,智能体与环境之间存在交互。智能体采取行动并与环境互动,然后观察环境的反馈,这个反馈影响了智能体未来的行动选择。这种交互性模拟了现实世界中的决策过程,智能体必须在与环境互动的过程中学习和改进。

(2) 试错学习(Trial and Error Learning)。强化学习是一种试错学习方法。智能体通过尝试不同的行动,观察其结果,然后根据奖励信号来调整其策略。这种学习方式与监督学习不同,因为没有明确的标签或指导,智能体必须自己发现最佳策略。

(3) 延迟奖励(Delayed Rewards)。在强化学习中,智能体的行动可能会导致延迟的奖励。这意味着某个行动的结果可能在未来的执行步骤中才能被感知到,智能体必须能够将当前行动与未来的奖励联系起来,以做出明智的决策。

(4) 累积奖励(Cumulative Rewards)。智能体的目标是最大化累积奖励,而不仅仅是单个行动的奖励。这意味着智能体必须考虑其行动对长期目标的影响,而不只是眼前的奖励。

(5) 策略和价值函数。在强化学习中,智能体的学习可以通过两种主要的方法来表示:一是策略(Policy),它定义了智能体如何根据状态选择行动;二是价值函数(Value Function),它评估了在特定状态或状态-行动对下的预期奖励。智能体使用这些方法来指导其决策过程。

(6) 探索与利用(Exploration and Exploitation)。智能体需要在探索新策略以发现更好行动的需求与利用已知策略以最大化当前奖励之间找到平衡。这是一个重要的挑战,因为纯粹的探索或纯粹的利用都可能导致不理想的结果。

(7) 马尔可夫性(Markovian Property)。许多强化学习问题被建模为马尔可夫决策过程(Markov Decision Process, MDP),其中当前状态包含足够的信息,以便智能体可以根据它做出最佳决策。这个假设在强化学习中是常见的,但并不总是成立。

### 1.1.3 强化学习与其他机器学习方法的区别

强化学习与其他机器学习方法有一些区别,主要集中在以下几个方面。

#### 1. 学习方式

(1) 监督学习(Supervised Learning)。在监督学习中,模型通过从有标签的数据中学习来预测给定输入的标签。模型接收输入和相应的目标标签,并通过最小化预测与目标之间的差距来进行训练。监督学习中的目标是在已知数据上进行准确的预测。

(2) 无监督学习(Unsupervised Learning)。无监督学习旨在从未标记的数据中发现模式和结构。它通常用于聚类、降维和密度估计等任务,而不需要明确的标签或目标。

(3) 强化学习。强化学习是一种试错学习方法,智能体通过与环境的交互来学习,根据奖励信号来调整策略,以最大化长期奖励。与监督学习和无监督学习不同,强化学习没有明确的标签或指导,智能体必须自己发现最佳策略。

## 2. 反馈信号

(1) 监督学习。监督学习使用有标签的数据，其中每个输入都有一个对应的目标标签，用于指导模型的训练。反馈信号是明确且确定的，通常是用于评估预测的损失函数。

(2) 无监督学习。无监督学习通常没有明确的目标或反馈信号，模型试图从数据中自动学习隐藏的结构或特征。

(3) 强化学习。强化学习使用环境提供的奖励信号来评估智能体的行为。奖励信号可能是稀疏、延迟的，甚至是随机的，智能体必须通过与环境的交互来探索并学习最佳策略。

## 3. 任务类型

(1) 监督学习和无监督学习通常用于解决特定的任务，如分类、回归、聚类等。

(2) 强化学习更适合于决策制定任务，其中智能体必须在与环境的互动中学习策略，以达到最大化累积奖励的目标。

## 4. 环境交互

(1) 监督学习和无监督学习通常在静态数据集上进行训练，不需要与外部环境进行交互。

(2) 强化学习涉及与动态环境的交互，智能体的决策会影响环境的状态和未来奖励，因此需要在线学习策略。

## 5. 应用领域

(1) 监督学习广泛用于图像分类、语音识别、自然语言处理等领域。

(2) 无监督学习常用于数据分析、降维、聚类等任务。

(3) 强化学习在自动化决策制定领域广泛应用，包括机器人控制、自动驾驶、游戏玩法优化、金融交易和医疗决策等。

总之，强化学习与监督学习和无监督学习在学习方式、反馈信号、任务类型、环境交互和应用领域等方面都存在显著差异。强化学习的独特之处在于其能够处理动态环境下的决策制定问题，并通过与环境的互动来学习最优策略。

# 1.2 强化学习的应用领域

强化学习在各个领域都有广泛的应用，包括机器人控制、自动驾驶、游戏玩法优化、金融交易和医疗决策等。通过不断地试错和学习，强化学习使机器能够在复杂和不确定的环境中自主决策，并取得优秀的性能。



扫码看视频

## 1.2.1 机器人控制与路径规划

强化学习在机器人控制和路径规划领域有广泛的应用，它提供了一种有效的方法来训练和优化机器人的决策策略，使机器人能够在复杂和不确定的环境中执行任务。下面是强化学习在这两个领域中的常见应用。

(1) 机器人导航和路径规划。强化学习可用于开发自主导航系统，使机器人能够规划和执行路径以到达目标位置。机器人可以通过与环境的交互来学习最佳路径规划策略，考

虑避开静态及动态障碍物和不同类型的地形等因素。

(2) 无人飞行器控制。强化学习可用于训练无人飞行器(如无人机)的控制策略,以执行各种任务,包括飞行路径规划、目标跟踪、搜索与救援等。无人飞行器(如无人机)可以根据不同的环境条件和任务要求来调整其行为。

(3) 机器人操作。在工业自动化中,强化学习可以用于训练机器人执行复杂的操作,如装配、拾取和放置物体。通过强化学习,机器人可以学习如何准确地控制其运动,以最小化误差并提高生产效率。

(4) 协作机器人。多个机器人之间的协作是一个复杂的问题,强化学习可以用于协调多个机器人的行动,以完成协同任务,如团队搜索、物流操作和协作装配。

(5) 环境监测与勘探。在危险或难以进入的环境中,如深海勘探、空中监测或火灾救援,机器人可以使用强化学习来规划其路径和行动,以最大化信息获取直到任务完成。

(6) 自动驾驶汽车。自动驾驶汽车需要在复杂的城市道路和高速公路环境中做出决策,以实现安全驾驶。强化学习可用于训练自动驾驶汽车的控制策略,以遵守交通规则、避免碰撞和适应交通流量。

(7) 仿真环境训练。在许多机器人应用中,可以首先在仿真环境中通过强化学习进行训练,然后将学到的策略迁移到实际机器人上。这可以减少在实际环境中的试错成本。

总之,强化学习在机器人控制和路径规划中提供了一种灵活而强大的方法,可以让机器人根据不断变化的情境和任务要求来学习和优化其决策策略,从而在各种应用领域中实现更高的自主性和性能。

## 1.2.2 游戏与虚拟环境

强化学习在游戏与虚拟环境领域有着广泛的应用,这些应用包括游戏智能、虚拟角色控制、游戏设计优化以及虚拟世界模拟等。强化学习在游戏与虚拟环境领域中的主要应用如下。

(1) 游戏智能。强化学习被广泛应用于游戏智能领域,包括电子游戏、棋类游戏和策略游戏。强化学习代理可以学习玩游戏并逐渐提高其性能,甚至在某些情况下击败人类玩家。例如,AlphaGo是一个基于强化学习的系统,成功击败了世界围棋冠军。

(2) 虚拟角色控制。在虚拟世界和游戏中,强化学习可用于训练虚拟角色(如游戏中的角色或机器人)的控制策略。这些角色可以学习如何行走、奔跑、跳跃、射击、躲避障碍物等,以更自然和智能的方式参与游戏或模拟。

(3) 游戏设计和优化。游戏开发者可以使用强化学习来改进游戏的设计和平衡,以确保游戏具有足够的挑战性和娱乐性。通过模拟玩家或角色的行为,游戏设计者可以调整游戏中的关卡难度、奖励结构和游戏规则。

(4) 虚拟世界模拟。在虚拟现实(Virtual Reality, VR)和增强现实(Augmented Reality, AR)应用中,强化学习可以用于创建逼真的虚拟环境,模拟多种场景和任务。这可以用于培训、模拟和测试,如飞行模拟器、外科手术模拟和城市交通仿真。

(5) 自动游戏测试。游戏开发者可以使用强化学习代理来测试其开发的游戏。代理可以自动玩游戏并检测潜在的问题或漏洞,有助于提高游戏质量和稳定性。

(6) 游戏玩法优化。强化学习可以用于个性化游戏玩法体验。游戏公司可以使用强化学习来自动调整游戏中的难度、提示和建议，以满足不同玩家的需求和技能水平。

总之，强化学习在游戏与虚拟环境领域的应用使游戏变得更具挑战性、娱乐性和智能化，并且为虚拟世界中的模拟和互动提供了新的可能性。这些应用不仅影响游戏产业，还扩展到了虚拟培训、虚拟旅游和虚拟现实等应用领域。

### 1.2.3 金融与投资决策

强化学习在金融和投资决策领域有广泛的应用，它可以用于制定复杂的金融策略、优化投资组合、风险管理和市场预测。强化学习在金融与投资决策领域的主要应用如下。

(1) 股票交易。强化学习可用于开发自动化交易系统，帮助投资者制定股票交易策略。智能体可以学习如何在不同的市场条件下进行买入和卖出决策，以最大化投资回报或最小化风险。

(2) 投资组合管理。强化学习可以用于优化投资组合的配置，以在不同的资产类别之间分配资金。它可以考虑多种因素，如风险、回报、流动性和投资目标，以生成最佳的资产配置策略。

(3) 风险管理。在金融中，风险管理至关重要。强化学习可以用于建立风险模型，识别和量化潜在的风险因素，并采取相应的风险控制措施，以减轻投资组合的风险。

(4) 市场预测。强化学习可以用于分析市场数据，预测未来的价格趋势和市场波动。智能体可以通过学习历史数据来识别模式和趋势，从而为市场预测提供依据。

(5) 高频交易。在高频交易环境中，强化学习可以用于制定快速的交易策略，以在毫秒级别的时间内做出决策。这可以帮助交易公司获得微小但频繁的交易机会。

(6) 市场策略制定和优化。金融机构可以使用强化学习来改进市场策略制定，以在不同的市场策略条件下实现最佳运行。这有助于减少交易成本和滑点。

(7) 信用风险评估。强化学习可以用于评估借款人的信用风险，根据不同的因素和历史数据来预测违约概率，从而指导信贷决策。

(8) 量化金融。量化金融是利用数学和计算方法来制定金融策略的领域。强化学习可以是量化金融策略的一部分，用于改进决策制定和交易执行。

需要注意的是，金融市场是复杂和不确定的，因此强化学习在这个领域的应用需要谨慎处理风险，并且需要考虑监管和合规性要求。尽管如此，强化学习提供了一种强大的工具，可以帮助金融从业者更好地理解市场、制定更智能的决策，并优化其投资和交易策略。

### 1.2.4 自动驾驶与智能交通

强化学习在自动驾驶和智能交通领域有着广泛的应用，它可以帮助实现更安全、高效和自主的交通系统。强化学习在自动驾驶与智能交通领域中的主要应用如下。

(1) 自动驾驶汽车。强化学习用于训练自动驾驶汽车的控制策略，使其能够在复杂的道路条件下行驶，遵守交通规则，预测其他车辆的行为，并应对各种不确定性，这包括路径规划、速度控制、车道保持和交叉路口处理等任务。

(2) 智能交通管理。强化学习可用于城市交通管理系统，帮助优化交通信号灯的定

时、协调交通流量、减少交通拥堵和改善路网效率。它可以根据实时交通情况来调整信号灯的控制策略。

(3) 交通事故预防。自动驾驶汽车和交通监控系统可以使用强化学习来预测潜在交通事故的发生,并采取预防措施,如紧急制动或发出车道偏离警告,以减少交通事故发生的可能性。

(4) 智能交通导航。强化学习可用于开发智能交通导航系统,这些系统可以根据交通状况、路况和用户偏好来制定最佳路线,并提供实时导航和交通信息。

(5) 自动停车。强化学习可用于训练车辆自动停车系统,包括并行停车、垂直停车和自动泊车。这可以提高停车的效率,减少停车空间的浪费。

(6) 交通流建模。强化学习可用于建立交通流模型,以理解城市交通系统的行为和性能。这有助于城市规划者更好地管理交通和基础设施。

(7) 公共交通优化。在公共交通系统中,强化学习可以用于改进公交车、地铁和电车的调度和运营,以提高运输效率和服务质量。

(8) 自动驾驶车队。强化学习可用于自动驾驶车队的管理,包括任务分配、路线规划和资源分配,以确保车队的高效运营。

这些应用使强化学习成为实现智能、自主和安全交通系统的关键技术之一。随着技术的不断进步和研究的深入,预计强化学习将在自动驾驶和智能交通领域发挥更大的作用,有望改善道路安全、减少交通拥堵,并提供更便捷的出行方式。

### 1.2.5 自然语言处理

强化学习在自然语言处理(Natural Language Processing, NLP)领域有多个重要应用,它可以帮助机器理解和生成自然语言文本,提高自然语言处理系统的性能。下面是强化学习在自然语言处理中的一些主要应用。

(1) 对话系统和聊天机器人。强化学习可用于训练对话系统和聊天机器人,使它们更具智能、自然和适应性。智能体可以通过与用户的对话互动来学习如何生成自然语言响应,提高对话质量和连贯性。

(2) 自动摘要生成。在文本摘要任务中,强化学习可用于训练模型,使其能够自动提取文本中的关键信息并生成简明扼要的摘要。这在新闻摘要、文档摘要和文章总结等领域有一定应用。

(3) 机器翻译。强化学习可以用于训练机器翻译系统,以提高翻译质量和语句流畅度。通过引入奖励信号,可以使机器翻译系统更好地捕捉语法结构和语义信息。

(4) 信息检索和搜索引擎。在信息检索任务中,强化学习可以帮助搜索引擎优化搜索结果的排名,以提供更相关和有用的搜索结果。这有助于改善用户搜索体验。

(5) 自然语言生成。自然语言生成任务包括文本生成、故事生成和广告生成等。强化学习可以用于训练生成模型,以生成高质量的自然语言文本。

(6) 情感分析。在情感分析中,强化学习可以用于训练情感分类模型,以识别文本中的情感倾向(如正面、负面或中性)。这在社交媒体分析和情感监测中有广泛应用。

(7) 文本分类。强化学习可用于文本分类任务,如新闻分类、垃圾邮件检测和情感分

类。智能体可以通过反馈信号来调整分类模型的权重，以提高分类准确性。

(8) 语音识别。虽然语音识别通常被视为音频信号处理的领域，但在将语音转换为文本的过程中，自然语言处理仍然发挥着关键作用。强化学习可用于改进语音识别系统的错误纠正和后处理步骤。

(9) 问答系统。在问答系统中，强化学习可以帮助系统理解问题并生成准确的答案。这对于智能搜索和虚拟助手非常有用。

以上这些应用使强化学习成为自然语言处理领域的一个重要工具，有助于提高自然语言处理系统的性能、适应性和用户友好性。随着深度学习和强化学习技术的不断进步，可以期待看到更多创新和改进，以改善自然语言处理在各种应用中的效果。

## 1.3 强化学习中的常用概念

强化学习中有许多常用的重要概念，这些概念是理解和应用强化学习的关键，熟悉这些概念是学习和应用强化学习的重要一步。



扫码看视频

### 1.3.1 智能体、环境与交互

在强化学习中，有3个基本概念非常重要，即智能体(Agent)、环境(Environment)和交互(Interaction)，这3个概念构成了强化学习问题的核心框架。

#### 1. 智能体

智能体是强化学习问题中的决策制定者或学习者，它可以是一个机器人、程序、自动驾驶汽车、虚拟角色或任何需要从环境中学习并采取行动的实体。智能体的任务是根据其目标从环境中获取最大的奖励或累积最大的回报。

#### 2. 环境

环境是智能体操作的外部系统或情境，包括智能体所处的位置、状态、行动以及与环境互动所获得的奖励。环境既可以是静态的，也可以是动态的，智能体的行动可能会改变环境的状态。

#### 3. 交互

交互是指智能体与环境之间的信息传递和相互作用。智能体负责采取行动，然后环境响应智能体的行动，提供下一个状态和相应的奖励。这个过程是不断迭代的，智能体根据先前的经验和当前的观察来做出新的决策。

智能体通过与环境的交互来学习，试图找到一种策略，即一种从状态到行动的映射，以最大化长期奖励或回报。在交互中，智能体根据其策略采取行动，观察环境的反馈，根据奖励信号来评估其行动的好坏，并不断调整策略以改进性能。这个交互过程是强化学习的核心，它使智能体能够在不断学习和优化的过程中适应不同的任务和环境。

### 1.3.2 奖励与回报的概念

在强化学习中，奖励(Reward)和回报(Return)是两个关键的概念，它们用于评估智能体采取行动的好坏，以指导学习过程。

#### 1. 奖励

(1) 奖励是一个标量值，用于量化智能体在特定状态下采取特定行动的即时效果的好坏程度。

(2) 奖励可以是正数(表示积极奖励，鼓励智能体采取这个行动)、负数(表示负面奖励，反对智能体采取这个行动)或零(表示中性奖励)。

(3) 奖励通常由环境提供，它反映了环境对智能体行动的反馈，但通常只涉及与当前行动和状态相关的瞬时奖励。

#### 2. 回报

(1) 回报是一个累积值，表示智能体从开始执行任务直到任务结束的整个任务期间所获得的总奖励。

(2) 回报是一种考虑了未来的奖励，因为它包括在未来执行步骤中获得的奖励。

(3) 回报是评估智能体策略效果的一种方法，智能体的目标通常是最大化长期回报。

理解奖励和回报的概念对于强化学习至关重要，因为它们构成了智能体学习和决策的基础。智能体的任务是在环境中采取行动，以最大化其预期回报。为了达到这个目标，智能体需要学会通过试错来调整其策略，以在不同状态下选择最佳的行动，从而最大化累积回报。

#### • 注意 •

强化学习算法的设计和性能评估通常都涉及奖励和回报的计算和优化，以使智能体能够有效解决各种决策问题。

### 1.3.3 马尔可夫性质与马尔可夫决策过程

在强化学习中，马尔可夫性质(Markov Property)和马尔可夫决策过程(Markov Decision Process, MDP)是两个重要的概念，它们在建模和解决强化学习问题中起着关键作用。

#### 1. 马尔可夫性质

马尔可夫性质是一个状态(或状态序列)的特性，它表示当前状态包含了足够的信息，以便完整地描述系统的状态，而不需要考虑先前的历史状态。简而言之，如果一个系统满足了马尔可夫性质，那么在给定当前状态的情况下，未来的状态只依赖于当前状态，而不依赖于之前的状态或行动。这一性质使强化学习问题可以被建模为马尔可夫决策过程，简化了问题的表示和求解。

#### 2. 马尔可夫决策过程

马尔可夫决策过程是对强化学习问题进行形式化建模的数学框架，它满足马尔可夫性

质。马尔可夫决策过程由五元组 $(S, A, P, R, \gamma)$ 组成，具体说明如下。

- (1)  $S$ (状态空间): 表示系统可能的状态集合。
- (2)  $A$ (动作空间): 表示智能体可以采取的行动集合。
- (3)  $P$ (转移概率): 表示在特定状态下采取特定行动后，系统转移到下一个状态的概率分布。
- (4)  $R$ (奖励函数): 表示在特定状态和采取特定行动后获得的即时奖励的函数。
- (5)  $\gamma$ (折扣因子): 表示未来奖励的折现因子，用于权衡当前和未来奖励的重要性。

在马尔可夫决策过程中，智能体的目标是找到一个策略，即从状态到行动的映射，以最大化预期回报或累积奖励。

马尔可夫决策过程提供了一种严格的数学方式来表示和求解强化学习问题，它允许智能体在不确定性和复杂性的环境中进行决策制定和学习。通过马尔可夫决策过程框架，可以运用各种强化学习算法来寻找最佳策略，从而使智能体能够在给定环境下获得最佳回报。

### 1.3.4 策略与价值函数

在强化学习中，策略(Policy)和价值函数(Value Function)是两个核心概念，用于描述和优化智能体的行为策略和性能评估。

#### 1. 策略

- (1) 策略是一种从状态到行动的映射，它定义了智能体在特定状态下应该采取哪种行动。
- (2) 策略通常用符号 $\pi$ 表示， $\pi(a|s)$ 表示在状态 $s$ 下采取行动 $a$ 的概率或概率分布。
- (3) 策略可以是确定性的，即对于每个状态只选择一个特定的行动，也可以是随机性的，即对于每个状态可能选择不同的行动的概率分布。
- (4) 优化策略是强化学习的主要目标之一，目的是找到一个策略，以最大化累积奖励或回报。

#### 2. 价值函数

- (1) 价值函数是一种度量状态或状态-行动对的预期累积奖励的函数，它用于评估智能体在不同状态下的价值。
- (2) 常见的价值函数包括状态值函数( $V$ 函数)和状态-行动值函数( $Q$ 函数)。
- (3) 状态值函数 $V(s)$ 表示在状态 $s$ 下按照某个策略 $\pi$ 执行后获得的预期累积奖励。
- (4) 状态-行动值函数 $Q(s, a)$ 表示在状态 $s$ 下采取行动 $a$ ，并按照某个策略 $\pi$ 执行后获得的预期累积奖励。
- (5) 价值函数是用于指导智能体行动选择的重要工具，通常通过贝尔曼(Bellman)方程等进行计算。

在强化学习中，策略和价值函数是非常重要的概念，它们帮助智能体评估不同行动的优劣以及状态的价值，从而指导智能体如何在环境中采取行动。强化学习的目标之一是找到最佳策略或价值函数，以最大化预期回报或累积奖励。不同的强化学习算法使用不同的方法来估计和优化策略或价值函数。