第10讲 强化学习

主讲人：郝晓玲
信息管理与工程学院

目录

01

强化学习的基本概念

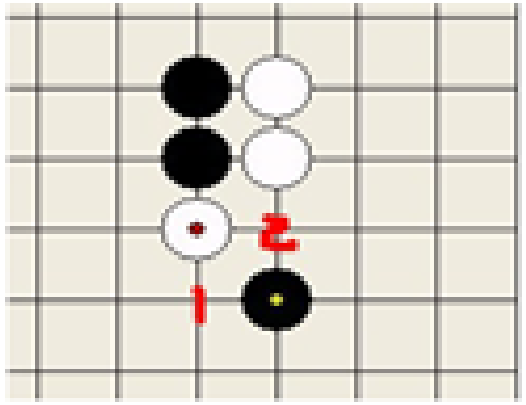
强化学习的概念

强化学习是一种机器学习领域中的方法，通过智能体与其环境之间的交互，学习如何采取行动以最大化某种累积奖励。

它是无监督学习的一种形式，智能体通过试错法来学习，无需明确的指导或反馈。

强化学习结合了动态规划、马尔可夫决策过程、蒙特卡罗方法等多个领域的技术。

强化学习的概念



（1）棋手1通过数学公式计算，发现位置1比位置2价值大。

（2）棋手2通过几次尝试，发现走位置1比走位置2赢棋的可能性大。

哪一种属于强化学习？

解决问题

强化学习能够训练智能体在复杂环境中找到最优解，提高智能体的决策和问题解决能力。

适应环境

通过试错和反馈机制，强化学习能够使智能体逐步适应环境，提高其对环境变化的应对能力。

自主学习能力

强化学习是一种无监督学习方法，智能体能够在没有人为干预的情况下自主学习，从而提高其智能素养。

强化学习的重要性

01

推动人工智能发展

强化学习是人工智能领域的重要分支，对于推动人工智能的发展和应用具有重要意义。

02

提高机器智能

通过强化学习，机器可以具备更高的智能水平，从而胜任更复杂的任务。

03

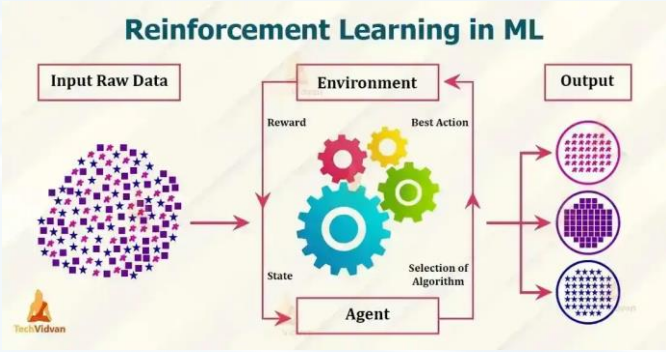
优化决策过程

强化学习能够找到最优策略，使得智能体在决策过程中获得最大利益。

强化学习与传统机器学习的区别

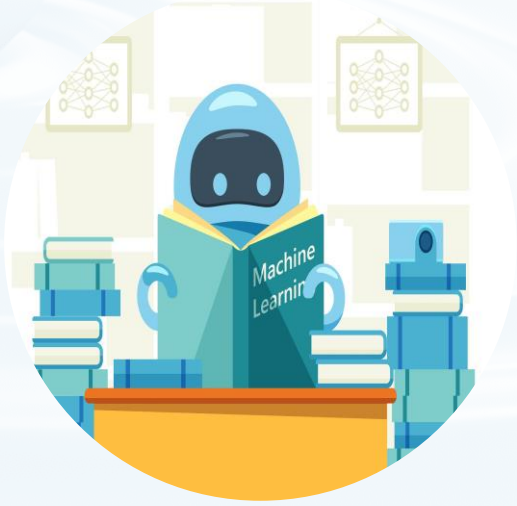
学习方式不同

传统机器学习主要依赖于监督学习或无监督学习，而强化学习通过**智能体与环境之间的交互**来学习。



目标设定不同

传统机器学习的目标是**最小化损失函数或误差**，而强化学习的目标是**最大化长期累积奖励**。



数据利用方式不同

传统机器学习主要利用**静态数据**集进行训练，而强化学习则依赖于智能体不断与环境交互产生的**动态数据**。

02

强化学习的基本要素

智能体

定义

智能体是指能够感知环境并通过动作与之交互的实体，是强化学习中的核心。

特性

智能体具有自主性、反应性、适应性和学习性等特性。

举例

智能体可以是一个机器人、一辆自动驾驶汽车或一个玩游戏的AI。

在围棋游戏中，AlphaGo就是智能体，它通过不断地与环境（即围棋棋盘和对手）交互来学习如何下棋。

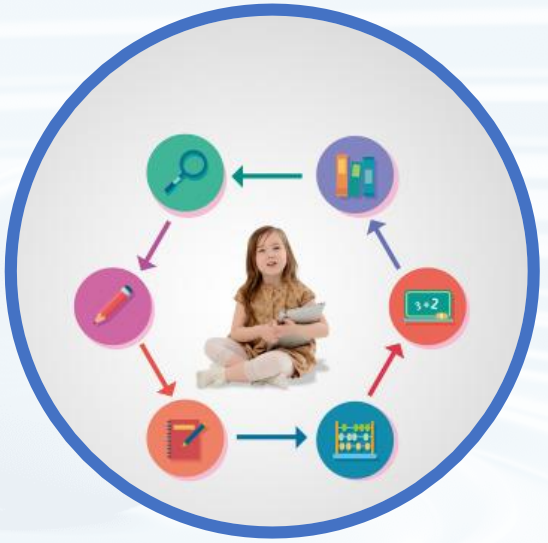
环境

定义

环境是智能体生存和交互的空间，包括状态、动作和奖励等元素。

特性

环境具有动态性、不确定性和可观测性等特性。



举例：

在围棋游戏中，环境就是围棋棋盘和对手；在自动驾驶场景中，环境就是道路、车辆、行人等交通元素构成的交通环境。

状态

定义

状态是描述环境或系统的某种特定情况或信息。
智能体做出决策取决于当前状态。

特性

状态具有可观测性、可描述性和时序性等特性。

表示方法

状态可用状态空间、状态向量或状态变量等表示。

状态可以是离散的（如围棋中的棋盘布局）或连续的（如自动驾驶汽车的速度和位置）。

举例：

在自动驾驶场景中，当前的道路状况（如车辆数量、速度、方向等）和汽车的位置、速度等信息构成了环境的状态。

行动

01

定义

动作是智能体根据当前状态做出的反应或决策，是智能体与环境进行交互的手段。

02

特性

动作具有离散性、有限性和影响环境等特性。

03

分类

动作可分为探索性动作和利用性动作等类型。

动作可以是离散的（如上下左右移动）或连续的（如控制汽车的速度和方向）。

举例：

在围棋游戏中，智能体的行动可以是落子的位置；

在自动驾驶场景中，智能体的行动可以是调整车辆的速度、方向或刹车等。

奖励

定义

奖励是智能体在环境中所获得的正面或负面的反馈，是指导智能体学习的信号。

特性

奖励具有即时性、稀疏性和延迟性等特性。

举例：

在围棋游戏中，赢得一局比赛可能获得正的奖励，而输掉比赛则获得负的奖励；

在自动驾驶场景中，安全行驶可能获得正的奖励，而发生事故则获得负的奖励。

策略

定义

策略是智能体在给定状态下选择动作的规则或方法，是强化学习的核心。

特性

策略具有最优性、平稳性、可学习性和风险等特性。

分类

策略可分为确定性策略和随机策略等类型。
确定性策略是指智能体在给定状态下选择唯一动作；随机策略则是指智能体在给定状态下以一定概率选择多个动作。

策略可以是确定性的（在给定状态下总是选择相同的动作）或随机的（在给定状态下根据概率分布选择动作）。

举例：

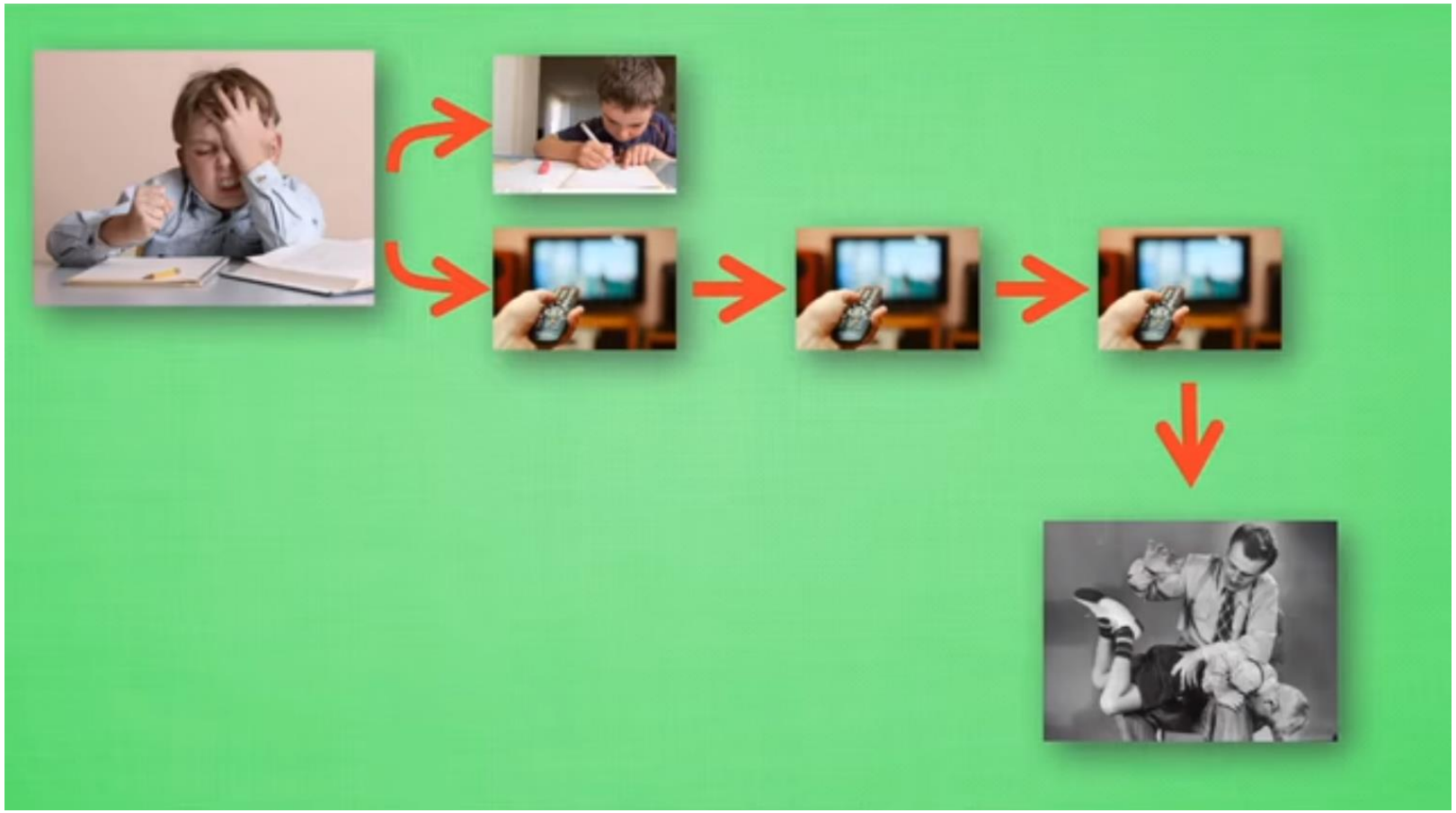
在围棋游戏中，一个策略可能是基于当前棋盘布局来选择下一步落子的位置；

在自动驾驶场景中，一个策略可能是根据当前道路状况和汽车状态来决定是否加速、减速或转弯。

03

强化学习的分类

https://www.bilibili.com/video/BV1kx411E7Yq/?spm_id_from=333.337.search-card.all.click



基于价值的强化学习

基本原理：

侧重于找出每个状态下每个行动的价值，这个价值表示在特定状态下采取特定行动的好处。

智能体的目标是选择能使该价值最大化的行动。

算法会学习一个值函数，用来预测每个行动的好坏，智能体通常会选择每个状态下数值最高的行动。

特点：

适用于具有离散行动空间的环境。

通过值函数间接学习如何行动，即先计算每个动作的价值，再选择价值最高的动作。

学习效率较高，因为可以直接通过比较价值来选择动作。

通过迭代更新Q值，求解状态-动作对的最优策略，如Sarsa算法。

基于策略的强化学习

基本原理：

直接学习策略，即学习一个从状态到动作的映射关系，这个映射关系被称为策略。

策略通常表示为行动的概率分布，智能体根据当前状态，通过策略选择动作。

学习过程中，智能体通过不断尝试和调整策略，以最大化累积奖励。

特点：

适用于高维或连续动作空间的环境。

能够学习到随机策略，这对于某些需要随机性的任务是有利的。

学习过程可能较为平滑，因为策略的调整通常是逐步的。

基于模型的强化学习

基本原理：

利用对环境的建模和预测来指导决策制定过程。

首先对环境进行建模，构建一个能够预测环境中状态转换和奖励信号的模型。

然后利用该模型来预测智能体在不同行动选择下的未来结果，并根据这些预测进行决策规划。

特点：

需要对环境进行建模，这通常需要大量的数据和计算资源。

一旦模型建立完成，可以利用模型进行高效的规划和决策。

适用于需要对未来结果进行预测和规划的任务。

在线强化学习与离线强化学习

学习对象

- 基于策略的方法直接学习策略或行动的概率分布。
- 基于价值的方法学习每个状态下每个行动的价值。
- 基于模型的方法则是对环境进行建模，预测状态转换和奖励信号。

适用环境

- 基于策略的方法更适合高维或连续动作空间的环境。
- 基于价值的方法通常用于具有离散行动空间的环境。
- 基于模型的方法则需要对环境进行建模，适用于需要对未来结果进行预测和规划的任务。

学习方式

- 基于策略的方法通过调整策略参数来最大化累积奖励。
- 基于价值的方法通过比较不同行动的价值来选择最优行动。
- 基于模型的方法则通过模拟仿真来预测未来结果，并根据预测结果进行决策规划。

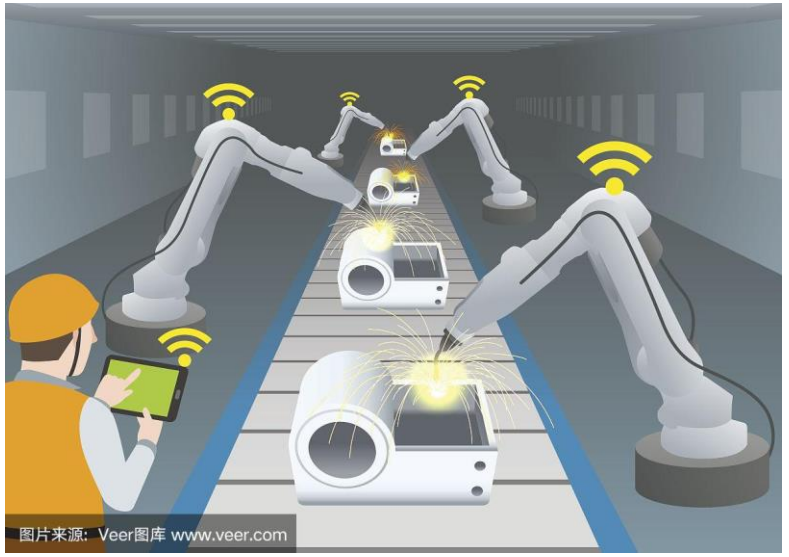
基于策略的强化学习应用场景

机器人控制：在机器人控制领域，基于策略的强化学习得到了广泛应用。机器人需要在不同时间步中不断调整动作以完成任务，这些任务往往涉及连续或高维动作空间。

基于策略的强化学习能够直接学习从状态到动作的映射关系，适用于这种复杂环境。

通过与环境的交互，机器人可以学会行走、抓取物体甚至进行复杂的任务，如在灾难场景中进行救援。

在这些场景中，每个决策会影响后续决策的效果，基于策略的强化学习能够通过反复试验找到最优的动作序列。



基于价值的强化学习应用场景

游戏AI：在游戏AI方面，基于价值的强化学习被广泛应用。以围棋为例，AlphaGo通过深度强化学习打败了世界顶尖围棋选手。

在这个场景中，智能体（即AI）在棋盘上进行试错学习，不断优化自己的策略，最终达到超越人类的水平。

基于价值的强化学习通过估计每个状态下执行某个动作的长期回报（即Q值），并选择具有最大Q值的动作来执行。

这种方法在游戏等具有离散动作空间的环境中非常有效。

基于模型的强化学习应用场景

自动驾驶：在自动驾驶领域，基于模型的强化学习具有潜在的应用价值。自动驾驶汽车需要在动态环境中做出实时决策，这要求智能体能够快速准确地预测未来状态并规划出最优路径。

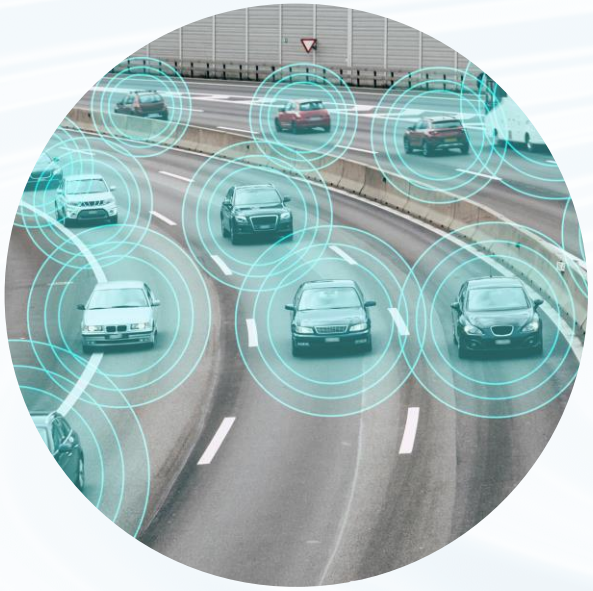
基于模型的强化学习首先对环境进行建模，构建一个能够预测环境中状态转换和奖励信号的模型。然后利用该模型来预测智能体在不同行动选择下的未来结果，并根据这些预测进行决策规划。这种方法能够显著提高自动驾驶汽车的安全性和可靠性。

需要注意的是，基于模型的强化学习需要大量的数据和计算资源来构建和训练模型，因此在实际应用中可能面临一些挑战。

04

强化学习的应用案例

自动驾驶汽车中的强化学习



路径规划

利用强化学习技术，自动驾驶汽车可以学习最优路径规划策略，从而在复杂交通环境中实现高效、安全的行驶。

决策与控制

强化学习可以帮助自动驾驶汽车在处理突发情况时做出快速、准确的决策，并通过实时调整车辆控制参数来确保行驶稳定性。

感知与预测

结合深度学习技术，强化学习可以提升自动驾驶汽车的感知能力，使其更准确地识别周围环境中的物体和事件，并预测它们的动态变化。

强化学习在行业自动化中的应用

智能制造

在制造业中，强化学习可以用于优化生产流程、提高设备利用率和降低能耗。例如，通过训练智能机器人学习高效的操作策略，可以实现生产线的自动化升级。

智能物流

强化学习可以帮助物流企业优化仓储管理、提高货物配送效率。例如，利用强化学习技术规划货物的存储位置和运输路线，可以减少物流成本和时间消耗。

智能电网

在电力行业中，强化学习可以用于实现智能电网的自主调度和优化运行。通过训练智能体学习电力需求和供应的平衡策略，可以提高电力系统的稳定性和经济性。

强化学习在贸易金融中的实践

风险评估

强化学习可以帮助金融机构更准确地评估贸易活动的风险水平，从而制定合理的信贷政策和风险控制措施。

欺诈检测

通过训练智能模型学习正常贸易行为的特征，强化学习可以自动识别出异常或欺诈性的贸易活动，提高金融安全性。

量化交易

强化学习可以用于开发量化交易策略，帮助投资者在复杂多变的金融市场中实现稳定的收益。通过训练交易机器人学习历史数据中的价格规律和交易信号，可以自动化地进行买卖操作。

强化学习在NLP、医疗保健等领域的创新

自然语言处理（NLP）

强化学习可以提升NLP任务的性能，如文本分类、情感分析和问答系统等。通过训练模型在大量文本数据中学习语言规则和语义关系，可以实现更准确的文本理解和生成。

医疗保健

在医疗保健领域，强化学习可以用于疾病预测、治疗方案优化和患者管理等任务。例如，利用患者的历史健康数据训练智能模型，可以预测疾病的发展趋势并提前采取干预措施。此外，强化学习还可以帮助医生制定个性化的治疗方案，提高患者的治疗效果和生活质量。

基于行为的智能

人工智能早期工作研究的课题涉及直接编写程序来模拟智能。

但是，新的想法认为人类的智能并非基于复杂是程序执行，而是在经历了世代进化后形成的简单的刺激——反应功能。这种关于“智能”的理论称为“基于行为的智能”，因为“智能的”刺激——反应功能似乎是一些行为的后果，这些行为导致某些个体在其他个体遇难时得以幸免并能繁衍后代。

基于行为的智能似乎能回答人工智能范畴的若干问题，例如，为什么基于冯.诺依曼结构的机器在计算能力上能轻易地胜过人类，却难以展现常识性的判断力。因此，基于行为的智能有希望成为人工智能研究中的一个重要影响因素。基于行为的技术已经应用在：人工神经网络领域，训练神经元如何按所期望的表现。

例：AlphaGo就是刺激——反应的很好例子，它下棋并不是穷尽剩余棋盘的算法，而是学习一流棋手的反应，做出智能的下棋行为。

最新消息：AlphaGo zero以100:0战胜AlphaGo

分组讨论

能够通过对机器进行编程来展现其智能行为的那种推测能力被认为是弱人工智能（weak AI），但是，机器能够通过编程而获得智力——亦即意识——的那种推测能力，则被认为是强人工智能（strong AI）强人工智能引发了广泛的争论。

反对者认为，机器在本质上与人类不同，它永远不能像人类那样感受爱、是否有情感，是不是自私？判断对错，以及考虑自我。

然而，支持者认为，人类的头脑是由许多小的部件构成，每个部件都不是人，没有意识，但是当它们结合在一起就成了人，为什么同样的现象就不可能出现在机器身上呢？

分组讨论

如果在AI驾驶模式下，Tesla出了交通事故，人类驾驶员是否有责任？
人工智能领域的研究者应该对他们研究成果的利用方式承担多少责任？
科学家所开发的AI如果产生了意想不到的后果，怎么办？

分组讨论

如果一个病人的的大脑坏掉，我们在坏掉前拷贝出它所有的意识，
然后用人工神经细胞一点点换掉病人的大脑，再重新模拟他的意识，
那个病人还是同一个人吗？那个病人还算是人吗？谈谈你的看法。

分组讨论

有些人把技术的进步看成是给与人类的一份厚礼——将人类从枯燥的、普通的任务中解放出来，为了更愉悦的生活方式打开大门。但对于同一个现象，另一些人则把它看做是剥夺公民就业机会、把财富引向权势人物的祸根。

对于那些看上去“不如”机器的人，怎么对待？他们被机器夺去工作的后果是什么？历史上有很多财富和权力分配不均而引起的革命。如果今天正在进步的技术加固了这种差异，那将产生灾难性的后果。

你们对新技术持有怎样的态度？