# Introduction to Reinforcement Learning

## Basics

**何新卫**

信息学院,
华中农业大学

2024 年 2 月 26 日

# Part I

# CHAPTER1: INTRODUCTION

# 几个例子


**图.** Caption

- 羚羊生存问题：刚出生的羚羊面临生存挑战，需要不断试错，学会站立和奔跑。通过尝试不同的策略，以达到生存的目的。

# 几个例子

**图.** Caption

- 大学新生表现：大学每年录取新生，尽管资质相近，但最终学习成绩差异很大。如何更好地理解自己的行为与结果之间的关系，从而有针对性地调整学习策略。

# 几个例子

**图.** Caption

- 《狂飙电视剧》高启强读孙子兵法：通过研究兵法和战略，在实践中逐步调整和优化这些策略，以应对不同的情况和挑战，人生从消极状态转向开挂模式。

# 几个例子


**图.** Caption

- **机器人控制**：一个机器人被安置在一个未知的环境中，需要通过学习来完成特定的任务，比如捡起一个物体并将其放置在另一个位置。机器人可以通过观察环境反馈并尝试不同的动作，逐步改进策略，使得完成任务的效率最大化。

# 几个例子


**图.** Caption

- 自动驾驶汽车：自动驾驶汽车需要通过对道路、交通信号灯等环境信息的感知，并学习如何进行轨迹规划和决策，以达到安全、高效地驾驶目标。汽车可以试验不同的行驶策略，并通过奖励或惩罚信号不断优化自己的驾驶行为。

# 几个例子

**图.** Caption

- 股票交易：通过观察市场变化和历史数据，并根据不同的交易策略进行投资，该算法可根据交易结果对策略进行调整和优化，以实现最大化收益。
- 等等。

  如何找到一个好的策略 $\pi(a|s)$，如何评估策略，如何优化策略？是本课程解决的重点内容。

# 本课程考核要求以及推荐参考书

**先修课程**：1) 微积分 A，310300001017，5 学分；2) 线性代数 A，310300001024，3 学分；3) 概率论与数理统计 A，310300001021，4 学分。
**考核要求**：

- 课程平时测评：20%
- 四次实验：20%
- 期末考试（待定）：60%

推荐参考书：Richard Sutton，Andrew Barto 著，俞凯等译，《强化学习》（第二版），北京: 电子工业出版社，2019 年 9 月

# 什么是强化学习

机器学习三大主流范式：1) 有监督学习 (supervised learning)；2) 无监督学习 (unsupervised learning)；3) 强化学习 (reinforcement learning)。

- 有监督学习: 基于 标注数据（如图片 + 标签），数据满足 独立同分布假设，训练模型，将正确的标签信息传递给神经网络。
- 如图像分类、目标检测、图像语义分割等任务。

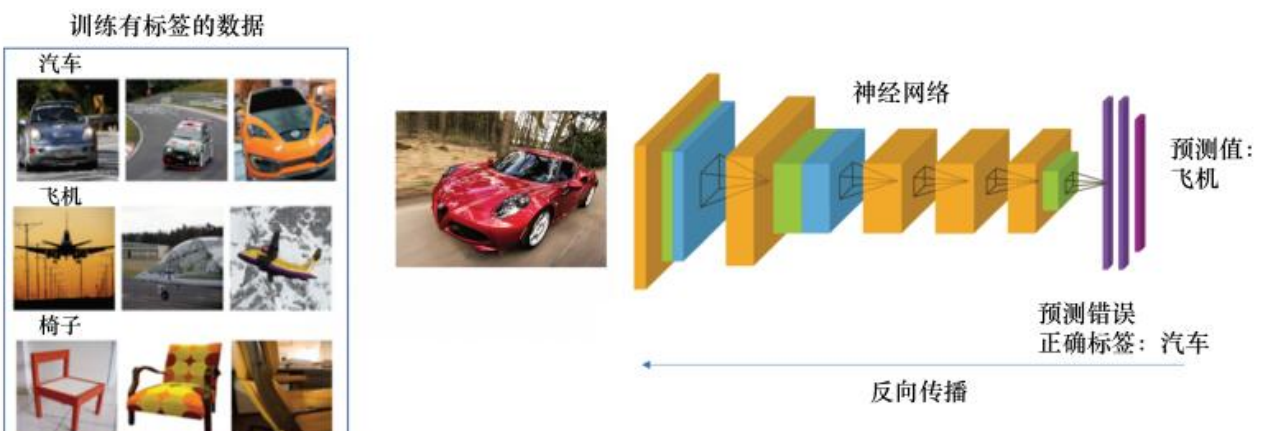


**图. 有监督学习**

# 什么是强化学习

机器学习三大主流范式：1) 有监督学习 (supervised learning)；2）无监督学习 (unsupervised learning)；3）强化学习 (reinforcement learning)。

- 无监督学习: 基于 无标注数据 （如图片），挖掘数据中的模式 (pattern)。
- 例如聚类、降维、异常检测等。

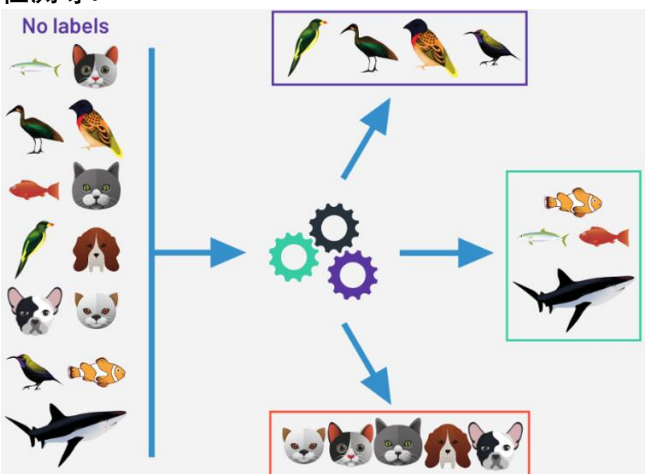


**图.** 无监督学习

# 什么是强化学习

机器学习三大主流范式：1) 有监督学习；2）无监督学习；3）强化学习。

- 强化学习: **智能体 (agent)** 在复杂、不确定**环境 (environment)**，通过不断交互，最大化它能获得的奖励。
- 智能体（agent)：强化学习的主体。由谁做动作或决策，谁就是智能体。比如在超级玛丽游戏中，玛丽奥就是智能体。在自动驾驶的应用中，无人车就是智能体。
- 环境（environment）是与智能体交互的对象，可以抽象地理解为交互过程中的规则或机理。在超级玛丽中，游戏程序就是环境。在围棋、象棋的例子中，游戏规则就是环境。在无人驾驶应用中，真实的物理世界则是环境。
- 一个小例子：超级马里奥。

# 什么是强化学习

## 一个小例子：超级马里奥

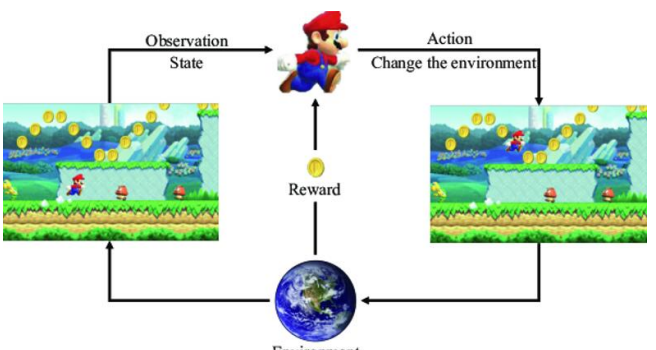


**图.** 例子背后的思考

- 目标：马里奥（智能体）在游戏环境中进行序列动作决策，从而最大化奖励，赢得游戏。

# 什么是强化学习

## 一个小例子：雅达利（Atari）游戏 Breakout


**图.** breakout 游戏界面

- 在玩游戏的过程中，我们可以发现智能体得到的观测（observation）不是独立同分布的，上一帧与下一帧间其实有非常强的连续性。我们得到的数据是相关的时间序列数据，不满足独立同分布。
- 我们并没有立刻获得反馈，游戏没有告诉我们哪个动作是正确动作。比如我们现在把木板往右移，这只会使得球往上或者往左去一点儿，我们并不会得到立刻的反馈。

# 什么是强化学习

思考

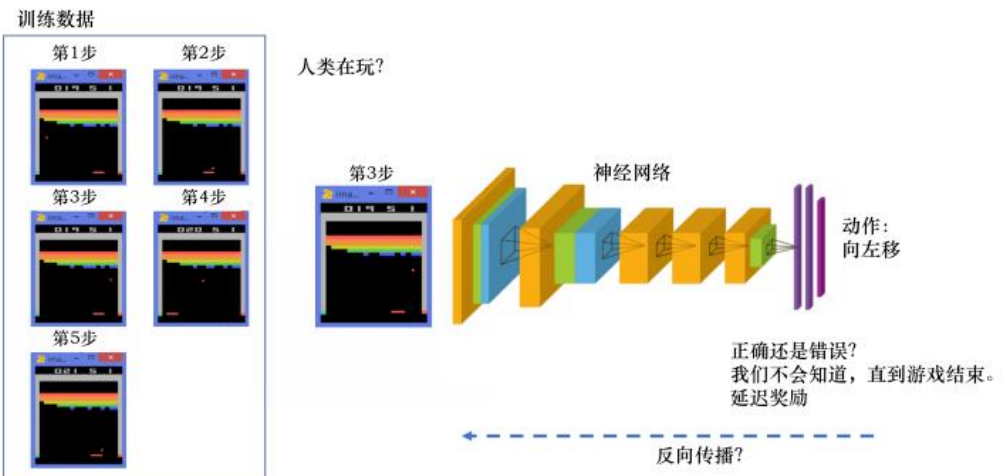


**图.** 强化学习打砖块

我们没有标签来说明现在这个动作是正确还是错误的，必须等到游戏结束才可能知道，这个游戏可能 10s 后才结束。现在这个动作到底对最后游戏是否能赢有无帮助，我们其实是不清楚的。这里我们就面临 延迟奖励（delayed reward） 的问题，延迟奖励使得训练网络非常困难。

# 强化学习特点

## 思考

- 智能体（或者学习器）从环境中拿到的样本是时间序列数据，不满最独立同分布假设
- 智能体没有监督标签，只有奖励信号，需要自己去发现那些动作带来最多的奖励
- 奖励会有延迟，智能体需要发掘具有更多潜在收益的动作
- 智能体的动作会影响获取的数据
- 智能体的学习过程是一个不断试错探索的过程（trial-and-error exploration）。

强化学习得到的模型可以有超人类的表现。监督学习获取的监督数据，其实是人来标注的，比如 ImageNet 的图片的标签都是人类标注的。因此我们可以确定监督学习算法的上限（upper bound）就是人类的表现，标注结果决定了它的表现永远不可能超越人类。但是对于强化学习，它在环境里面自己探索，有非常大的潜力，它可以获得超越人类的能力的表现，比如 DeepMind 的 AlphaGo 这样一个强化学习的算法可以把人类顶尖的棋手打败。

# 强化学习示意图

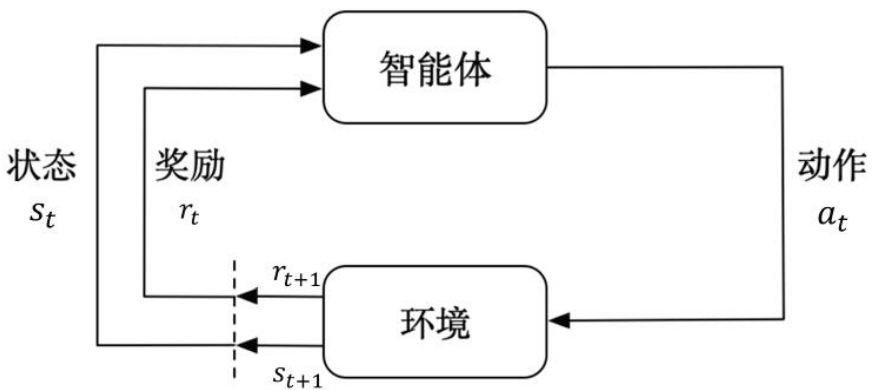


**图.** 强化学习示意图

过程描述：智能体在环境里面获取某个状态后，它会利用该状态输出一个动作（action），这个动作也称为决策（decision）。然后这个动作会在环境之中被执行，环境会根据智能体采取的动作，输出下一个状态以及当前这个动作带来的奖励。智能体的目的就是尽可能多地从环境中获取奖励。

# 强化学习历史

- 标准强化学习：早期需要对状态（例如图像）设计特征，利用特征训练策略网络或者价值评估网络
- 深度强化学习（deep reinforcement learning）：深度学习 + 强化学习，端到端训练，直接输入状态，输出动作概率或者拟合价值函数等。

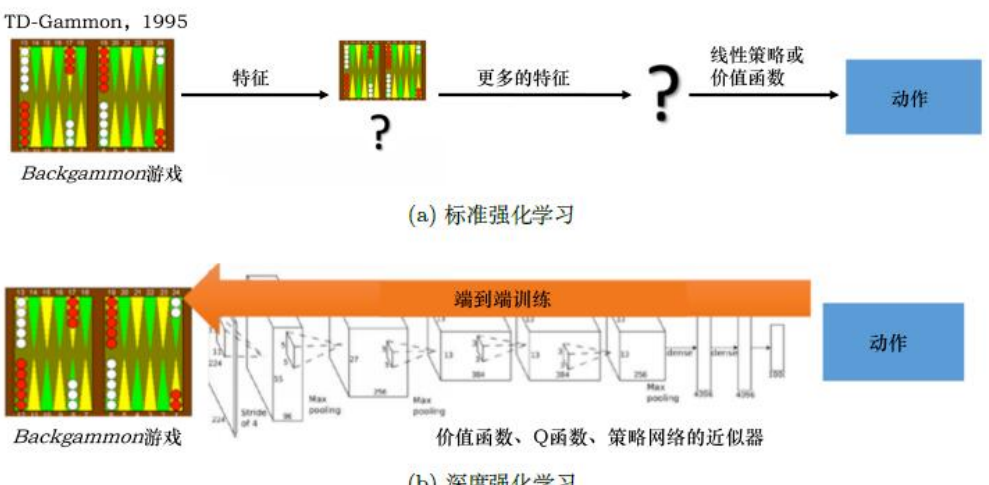


**图.** 标准强化学习和深度强化学习区别

# 强化学习应用举例

为什么强化学习在这几年有很多的应用，比如玩游戏以及机器人的一些应用，并且可以击败人类的顶尖棋手呢？这有如下几点原因: 1) 算力（computation power），有了更多的 GPU，可以更快地做更多的试错尝试; 2) 通过不同尝试，智能体在环境里面获得了很多信息，然后可以在环境里面取得很大的奖励; 3) 通过端到端训练把特征提取和价值估计或者决策一起优化，这样就可以得到一个更强的决策网络。


**图.** 强化学习例子

# 总结

- 强化学习定义和特点
- 机器学习三大主流范式及区别

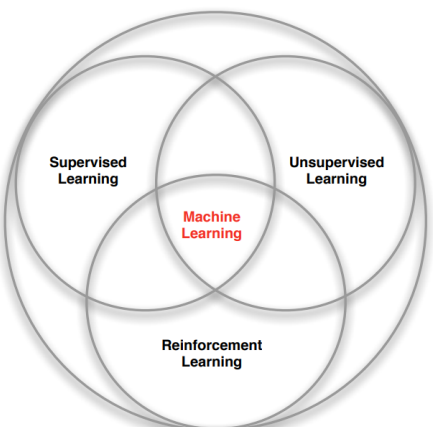


**图.** Machine Learning Paradigms. Slide from https://www.davidsilver.uk/wp-content/uploads/2020/03/intro_RL.pdf

# 序列决策（SEQUENTIAL DECISION MAKING）

**智能体和环境**：强化学习研究的问题是智能体与环境交互的问题，图左边的智能体一直在与图右边的环境进行交互。智能体把它的动作输出给环境，环境取得这个动作后会进行下一步，把下一步的观测与这个动作带来的奖励返还给智能体。这样的交互会产生很多观测，智能体的目的是从这些观测之中学到能最大化奖励的策略。

**奖励**：奖励是由**环境**给的一种标量的反馈信号（scalar feedback signal），这种信号可显示智能体在某一步采取某个策略的表现如何。强化学习的目的就是最大化智能体可以获得的奖励，智能体在环境里面存在的目的就是最大化它的期望的累积奖励（expected cumulative reward）。

**奖励类型和环境相关**：1）比如一个象棋选手，他的目的是赢棋，在最后棋局结束的时候，他就会得到一个正奖励（赢）或者负奖励（输）。2）在股票管理里面，奖励由股票获取的奖励与损失决定。3）在玩雅达利游戏的时候，奖励就是增加或减少的游戏的分数，奖励本身的稀疏程度决定了游戏的难度。

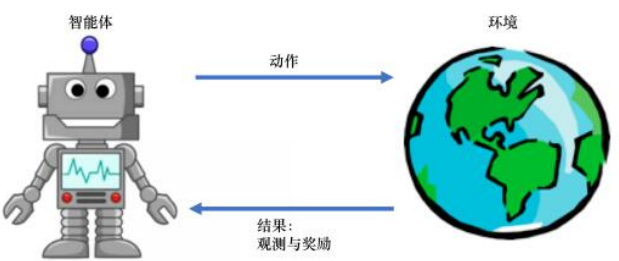


**图.** 智能体与环境

# 序列决策（SEQUENTIAL DECISION MAKING）

智能体的目的就是选取一系列的动作来最大化奖励，所以这些选取的动作必须有长期的影响。但在这个过程里面，智能体的奖励其实是被延迟了的，就是我们现在选取的某一步动作，可能要等到很久后才知道这一步到底产生了什么样的影响。
在与环境的交互过程中，智能体会获得很多观测。针对每一个观测，智能体会采取一个动作，也会得到一个奖励。所以历史是观测、动作、奖励的序列。

$$H_t = o_1, a_1, r_1, ..., o_t, a_t, r_t$$

智能体在采取当前动作的时候会依赖于它之前得到的历史，所以我们可以把整个游戏的状态看成关于这个历史的函数：

$$s_t = f(H_t)$$

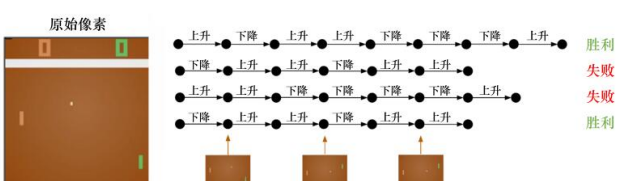


**图.** 玩 Pong 游戏

# 重要术语

## 状态（STATES）和观测（OBSERVATIONS）

- 状态 $s$ 是对环境的完整描述，不会隐藏环境的信息
- 观测 $o$ 是对状态的部分描述，可能会遗漏信息.

**图.** 状态 $s$ 可以是是智能体在环境中某个时间下的环境观测.

- 在深度强化学习中，我们几乎总是用实值向量、矩阵或高阶张量来表示状态和观测。
- 例如，视觉观测可以通过其像素值的 RGB 矩阵来表示；机器人的状态可能通过其关节角度和速度来表示。
- 当智能体可以观测到环境的完整状态时，我们称环境是完全观测的。当智能体只能看到部分观测时，我们称环境是部分观测的。

# 重要术语

## 状态空间 (STATE SPACE)

- 状态空间（state space）是指所有可能存在状态的集合，记作花体字母 $S$。
- 状态空间可以是离散的，也可以是连续的。状态空间可以是有限集合，也可以是无限可数集合。在超级玛丽、星际争霸、无人驾驶这些例子中，状态空间是无限集合，存在无穷多种可能的状态。围棋、五子棋、中国象棋这些游戏中，状态空间是离散有限集合，可以枚举出所有可能存在的状态（也就是棋盘上的格局）。

# 重要术语

## 动作空间（ACTION SPACE）

- 不同的环境允许不同类型的动作。在给定的环境中，所有有效动作的集合通常被称为**动作空间**。
- **离散动作空间（discrete action space）**：智能体只能选择有限数量的动作，例如，像雅达利游戏和围棋（Go）等环境中，是离散动作空间。
- 例：走迷宫机器人如果只有往东、往南、往西、往北这 4 种移动方式。
- **连续动作空间（continuous action space）**：其他环境，比如智能体控制物理世界中的机器人。在连续空间中，动作是实值向量。
- 例：如果机器人可以向 360 中的任意角度进行移动，则其动作空间为连续动作空间

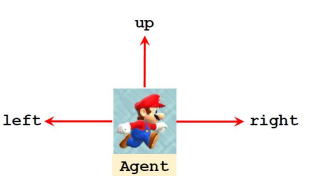


**图.** 动作空间

# 重要术语

## 策略（POLICY）

策略是智能体用来决定采取何种行动的规则。其实是一个函数，**用于把输入的状态映射为动作**。

- **确定性策略（deterministic policy）**：它可以是确定性的，此时通常用 $\mu$ 进行表示:

$$a_t = \mu(s_t),$$

- **随机策略（stochastic policy）**，这种情况下通常以 $\pi$ 表示:

$$a_t \sim \pi(\cdot|s_t).$$

因为策略实质上是智能体的思维过程，所以将" 策略" 一词替换为" 智能体" 并不罕见，例如说" 智能体正在尝试最大化奖励。"

- 在深度强化学习中，我们处理的是参数化策略：策略的输出是可计算的函数，它们依赖于一组参数（例如神经网络的权重和偏置），我们可以通过某种优化算法来调整这些参数，以改变行为。
- 通常用 $\theta$ or $\phi$ 表示这种策略的参数，然后将参数写成策略符号的下标，以突出它们之间的联系:

$$\begin{aligned} a_t &= \mu_\theta(s_t) \\ a_t &\sim \pi_\theta(\cdot|s_t). \end{aligned} \tag{1}$$

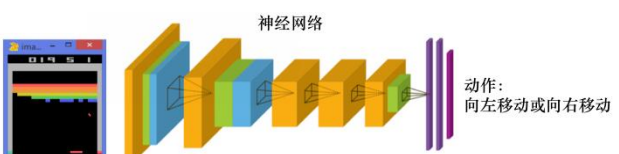


**图.** 策略函数

## 策略总数

- 假设动作空间是离散的 $A$, 状态空间也是离散的 $S$, 且每个状态的 valid actions 是动作空间任意一个。则 $\pi(a|s)$ for all s, 策略总数是

$$|A|^{|S|}$$

- 2x1 的 gridworld, 每个 cell 可执行 up、down、left、right 四个动作，则确定性策略的总数目为 $4^2$ (Hint: $|A|^{|S|}$), 分别如下：策略 1：$s = 0 \rightarrow up$，$s = 1 \rightarrow up$
  策略 2：$s = 0 \rightarrow up$，$s = 1 \rightarrow down$
  策略 3：$s = 0 \rightarrow up$，$s = 1 \rightarrow left$
  策略 4：$s = 0 \rightarrow up$，$s = 1 \rightarrow right$,
  ...,
  策略 13：$s = 0 \rightarrow right$，$s = 1 \rightarrow up$
  策略 14：$s = 0 \rightarrow right$，$s = 1 \rightarrow down$
  策略 15：$s = 0 \rightarrow right$，$s = 1 \rightarrow left$
  策略 16：$s = 0 \rightarrow right$，$s = 1 \rightarrow right$
- 高中数学知识：4 个包代表 4 个状态，每个包里有四个小球，代表可以执行的四个方向，分别从四个包里摸一个小球，生成一个策略。问总过有几种策略（或者组合）。

## 例题: MARIO

在马里奥游戏中，使用大小为 224x224x3 (RGB channels) 的图像表示，以及三种可能的动作, 可以定义多少种独特的策略和状态?

# 作业题目

## 1: GRIDWORLD

An agent moving in the 4 x 4 grid world.

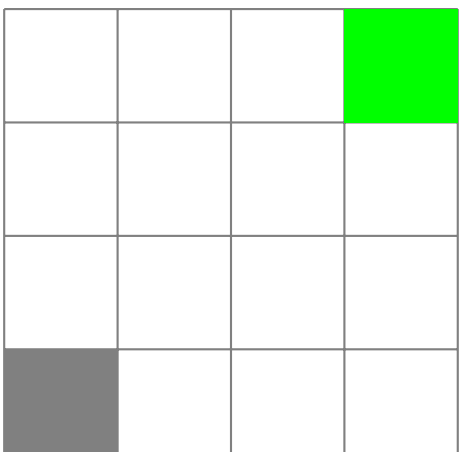

**图.** Possible actions in each cell: $\uparrow, \downarrow, \leftarrow, \rightarrow$. Cell with gray entrance and green means ending.

- The number of States?
- The number of Actions?
- The number of policies?
- Please give one policy?
- Please encode each state.

# 作业题目

## 2: 概念性习题

1）在强化学习中，观测和状态的区别 2）强化学习和有监督学习的差别

## 例题: MARIO

在马里奥游戏中，使用大小为 224x224x3 (RGB channels) 的图像表示，以及三种可能的动作, 可以定义多少种独特的策略和状态?

- 在马里奥游戏场景中有三种可能的动作，策略数量可以根据每个状态的可能动作总数来计算。由于在每个状态下有三种可用的动作，因此策略的总数将是 $(3^N)$, 其中 $(N = 256^{224\times224\times3})$ 是状态空间个数。

# 重要术语

## 轨迹（TRAJECTORIES）

一个 **轨迹** $\tau$ 是智能体在环境中交互生成的一个状态-动作序列, $\tau = (s_0, a_0, s_1, a_1, ...)$. 环境中的第一个状态是, $s_0$, 是从一个初始状态分布中随机采样得到,
轨迹常被称为 **回合（episodes）** 或者预演（rollouts）.

- 一个 episode 可以代表一场游戏的结束、迷宫旅行的结束，或者任何重复性交互的结束
- 尽管 episode 的结束可以终止于不同的结果（或者 reward），比如赢了或者输了，都可以视为同一个状态，即终止状态（terminal state）
- episode 之间可以视为独立的
- 两类任务：1）episodic task，即有终止状态; continuing task，即没有结束（如机器人的一生）。

# 价值函数（VALUE FUNCTION）

什么状态是最佳的状态？

**状态的价值**：我们用价值函数来对当前状态进行评估。价值函数用于评估智能体进入某个状态后，可以对后面的奖励带来多大的影响。价值函数值越大，说明智能体进入这个状态越有利。价值函数的值是对未来奖励的预测，我们用它来评估状态的好坏。

**折扣因子（discount factor）**：我们希望在尽可能短的时间里面得到尽可能多的奖励。比如现在给我们两个选择：10 天后给我们 100 块钱或者现在给我们 100 块钱。因此，我们可以把折扣因子放到价值函数的定义里面，价值函数的定义为：$V_\pi(s) \doteq E_\pi[G_t|s_t = s] = E_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \Big| s_t = s], \quad s \in S$ 期望 $E_\pi$ 的下标是 $\pi$ 函数，$\pi$ 函数的值可反映在我们使用策略 的时候，到底可以得到多少奖励

什么动作是最佳的动作？

**状态-动作的价值** Q 函数里面包含两个变量，即状态和动作。其定义为：

$Q_\pi(s,a) \doteq E_\pi[G_t|s_t = s, a_t = a] = E_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \Big| s_t = s, a_t = a]$ 为当我们得到 $Q$ 函数后，进入某个状态要采取的最优动作可以通过 $Q$ 函数得到。

# 模型

- 模型表示智能体对环境的状态进行理解，它决定了环境中世界的运行方式。
- **下一步的状态**取决于**当前的状态**以及**当前采取的动作**。它由状态转移概率和奖励函数两个部分组成。状态转移概率即：
  $P^a_{ss'} = p(s_{t+1} = s' | s_t = s, a_t = a)$
- 奖励函数是指我们在当前状态采取了某个动作，可以得到多大的奖励
  $R(s, a) = \mathbb{E}[r_{t+1} | s_t = s, a_t = a]$

# 强化学习智能体的类型

## 基于价值和基于策略

基于价值的智能体（value-based agent）显式地学习价值函数，隐式地学习它的策略。策略是其从学到的价值函数里面推算出来的。

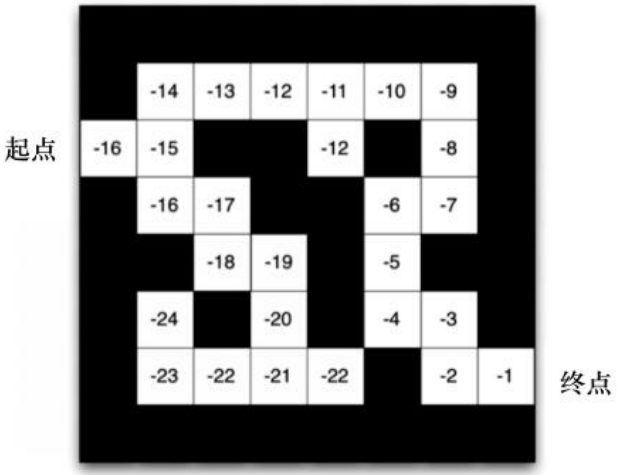


**图.** 使用基于价值的强化学习方法得到的结果

基于策略的智能体（policy-based agent）直接学习策略，我们给它一个状态，它就会输出对应动作的概率。基于策略的智能体并没有学习价值函数。

# 强化学习智能体的类型

## 有模型和免模型

- 有模型 (model-based) 强化学习智能体: 有模型（model-based）强化学习智能体通过学习状态的转移来采取动作, 它通过学习价值函数和策略函数进行决策。
  - 对真实环境建模
  - 可以在虚拟环境中预测要发生的事情，并采取对自己有利的策略
  - 马尔可夫模型等，属于此类智能体
- 免模型（model-free） 强化学习智能体的模型：没有去直接估计状态的转移，也没有得到环境的具体转移变量。
  - 不对环境建模，直接通过与真实环境交互来学习最有策略
  - 属于数据驱动型，需要大量采样来估计状态、动作和奖励函数，从而优化动作策略
  - 大部分主流方法都是免模型

# 重要话题：探索（EXPLORATION）与利用 (EXPLOITATION)

- 探索即我们去探索环境，通过尝试不同的动作来得到最佳的策略（带来最大奖励的策略）。
- 利用即我们不去尝试新的动作，而是采取已知的可以带来很大奖励的动作。

在刚开始的时候，强化学习智能体不知道它采取了某个动作后会发生什么，所以它只能通过试错去探索，所以探索就是通过试错来理解采取的动作到底可不可以带来好的奖励。利用是指我们直接采取已知的可以带来很好奖励的动作。所以这里就面临一个权衡问题，即怎么通过牺牲一些短期的奖励来理解动作，从而学习到更好的策略。

# 总结

- 说一下免模型和有模型的区别
- 基于策略和基于价值的强化学习的区别
- 举一个生活的例子，说明探索与利用的均衡问题
- 给一个游戏例子，举例说明出其中涉及到的奖励、状态、动作空间、策略分别指代什么？