WDS

DataScience@Web

第1部分

强化学习基础知识介绍

汇报人：商小雨

2021.7.6

2021 WDS暑期讨论班

目录

- 人工智能与Agent
- 强化学习定义
- 重要概念介绍
- 汇报安排

目录

- 人工智能与Agent
- 强化学习定义
- 重要概念介绍
- 汇报安排

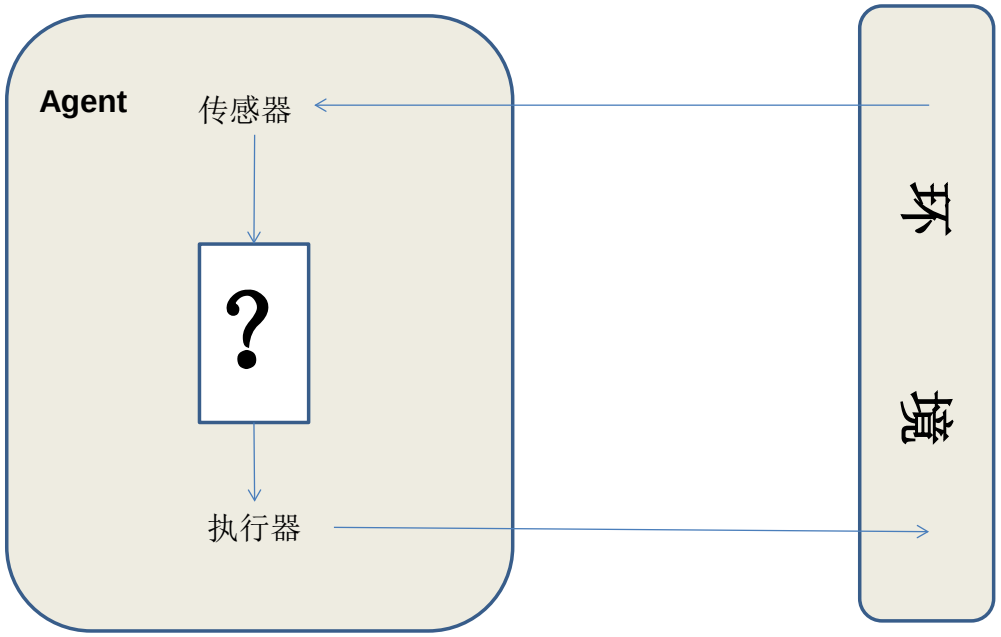
人工智能

- 目的：（科学上）不但要理解思维，研究思维的规律，构建思维的模型，（工程上）而且进一步要制作能思考的人造物。
- 定义：

像人一样思考	理性地思考
“使计算机思考的令人激动的新成就，……按完整的字面意思就是：有头脑的机器”（Haugeland, 1985） “与人类思维相关的活动，诸如决策、问题求解、学习等活动（的自动化）”（Bellman, 1978）	“通过使用计算模型来研究智力”（Charniak & McDermott, 1985） “使感知、推理和行动成为可能的计算的研究”（Winston, 1992）
像人一样行动	**理性地行动**
“创造能执行一些功能的机器的技艺，当由人来执行这些功能时需要智能”（Kurzweil, 1990） “研究如何使计算机能做那些目前人比计算机更擅长的事情”（Rich & Knight, 1991）	“计算智能是研究智能Agent的设计”（Poole等人, 1998） “AI关心人工制品中的智能行为”（Rich & Knight, 1991）

理性Agent

- Agent：源于拉丁语agere，意为“去做”，Agent就是能够行动的某种东西。
- Agent是以感知序列为输入，以动作作为输出的函数。能够通过**传感器**感知**环境**，通过**执行器**的动作作用于环境。



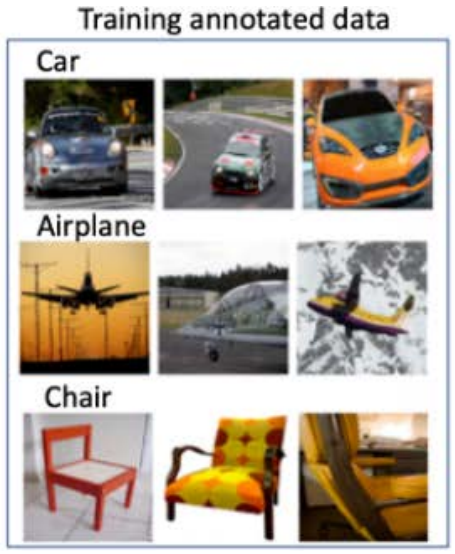
理性Agent

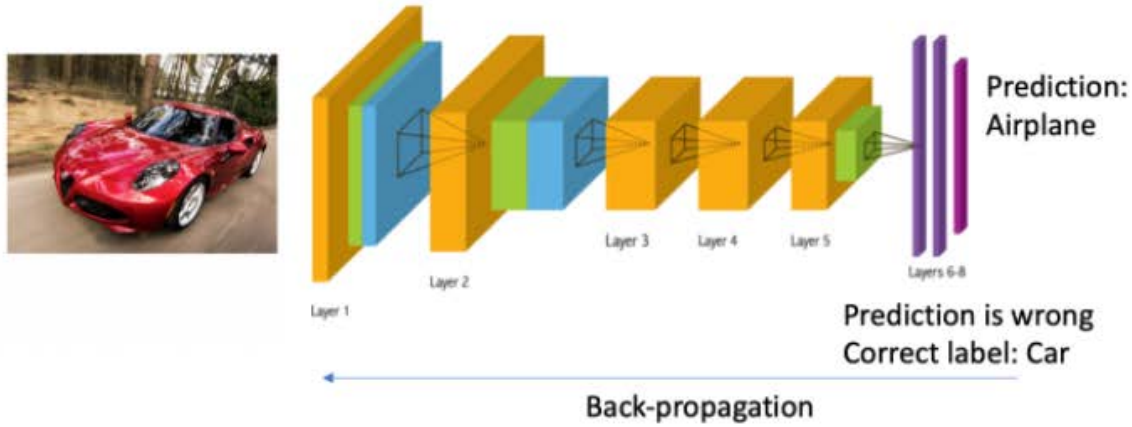
- 一个理性的 Agent 以达到**最好结果**为行动目标，如果有不确定性的时候，以获得**最大期望值的结果**为行动目标。
- 更具有一般性
  - 相较于“理性地思考”，理性地思考只是实现理性的几种可能的机制之一
  - 相较于“像人一样思考/行动”，基于数学而非经验主义

图片分类

监督学习

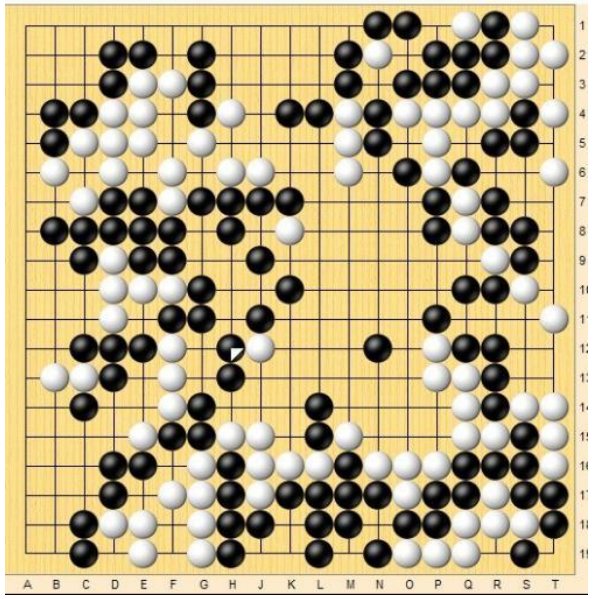
- 两个假设
  - 输入的数据（标注的数据）是没有关联的
  - 告诉 Agent 正确的标签是什么，让 Agent 通过正确的标签修正自己的预测





下棋问题

- 监督学习实现
  - 训练数据：大量的棋盘状态
  - 标签：对应的最佳落子位置
- 实现难点
  - 获取带标签数据的人力成本较大
  - 难以对大量的棋局给出精确一致的评价



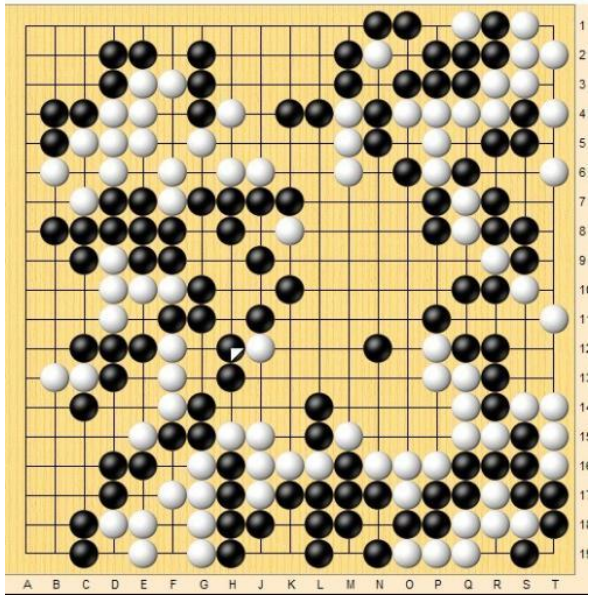
问题分析

- 延迟奖励（Delayed Reward）

  棋局结束 Agent 才能知道之前的落子动作是否有帮助

- 试错探索（Trial-and-error Exploration）

  需要 Agent 自己去尝试不同行为，以此发现哪些行为可以得到最多的奖励



解决方法

- 特点分析
  - 延迟奖励（Delayed Reward）
  - 试错探索（Trial-and-error Exploration）

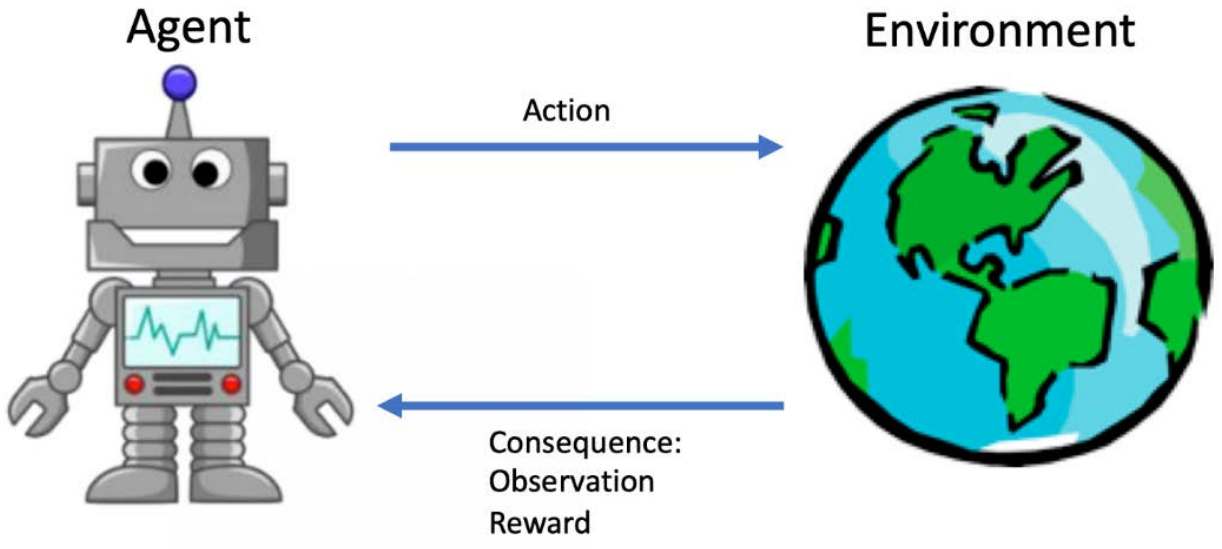
- 解决方法
  - 通过大量的模拟数据（**试错**）以及最后的结果（**奖励**）去倒退每一步棋的好坏，从而学习出最佳的下棋策略
  - 只提供感知信息和偶尔获得的反馈，一个 Agent 如何在未知的环境中变得熟练

目录

- 人工智能与Agent
- 强化学习定义
- 重要概念介绍
- 汇报安排

强化学习（Reinforcement Learning）

- Agent 被置于一个环境中，而且必须学会在其中游刃有余。
- 强化学习（Reinforcement Learning），指一类从与环境的交互中不断学习的问题以及解决这类问题的方法。



强化学习的特征

- 延迟奖励：Agent 会从环境中获得延迟的奖励
- 试错探索：需要 Agent 自己探索环境来获取对环境的理解
- 数据具有时间关联，而不是独立同分布
- 任务环境是连贯的

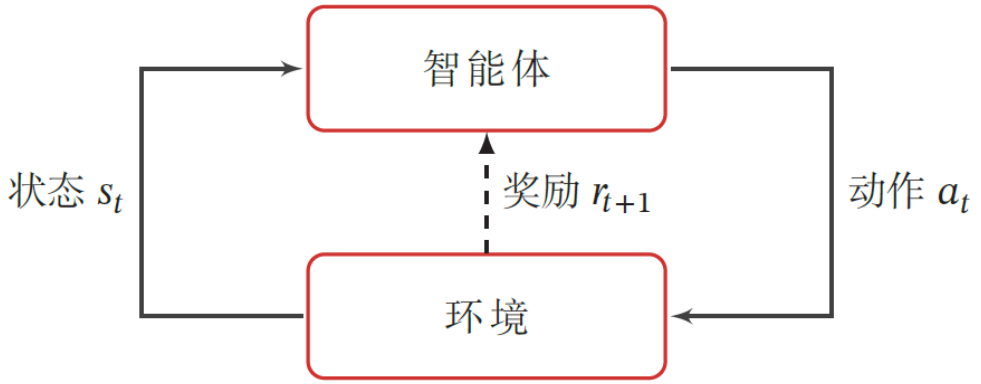
与其他方法的区别

- 监督学习 vs. 强化学习
  - 监督学习：需要一定数量的带标签的数据，描述各种情况以及该情况下应采取的正确动作。
  - 强化学习：不需要给出“正确”策略作为监督信息，只需要给出策略的（延迟）奖励，并通过调整策略来使得期望回报最大化。
- 无监督学习 vs. 强化学习
  - 无监督学习：通常是关于发现隐藏在未标记数据集合中的结构。
  - 强化学习：试图最大化奖励，而不是试图找到隐藏结构。
- 强化学习更侧重于从**交互**中进行**目的导向**的学习

目录

- 人工智能与Agent
- 强化学习定义
- 重要概念介绍
- 汇报安排

基本要素



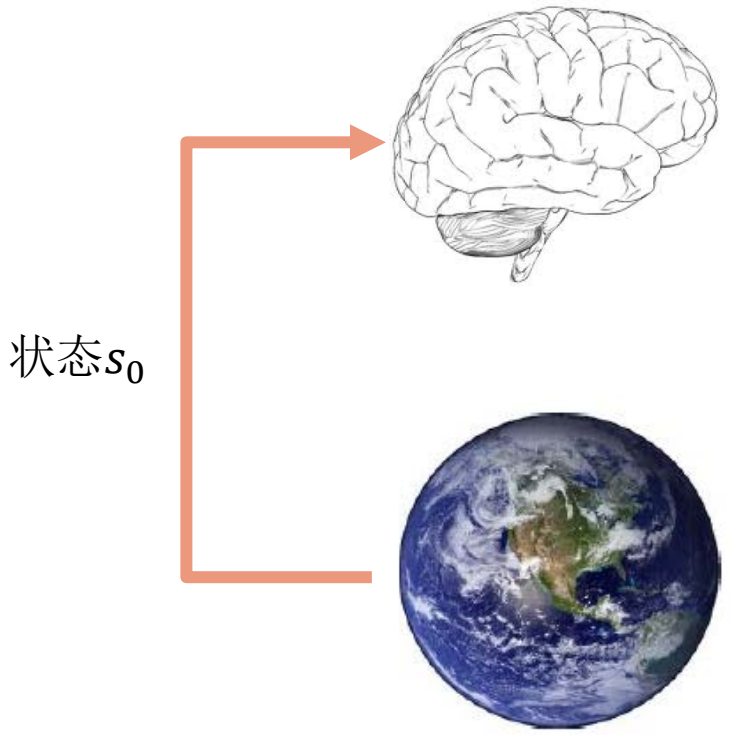
两个进行交互的对象：Agent 和环境

- Agent：感知外界**环境状态**和反馈的**奖励**，根据外界环境状态做出不同的**动作**，根据外界环境的奖励来调整**策略**。
- 环境：受 Agent **动作**的影响而改变其**状态**，并反馈给 Agent 相应的**奖励**。

名称	符号	描述
状态	s	对环境的描述，可以是离散的或连续的，状态空间为 $\mathcal{S}$
动作	a	对Agent行为的描述，可以是离散的或连续的，动作空间为$\mathcal{A}$
策略	$\pi(a\|s)$	Agent 根据环境状态 s 来决定下一步动作 a 的函数
状态转移概率	$p(s'\|s,a)$	Agent根据当前状态 s 做出动作 a 之后，环境在下一个时刻转变为状态 s' 的概率
即时奖励	$r(s,a,s')$	Agent根据当前状态 s 做出动作 a 之后，环境反馈给Agent的奖励。该奖励常和下一个时刻的状态 s' 有关

交互过程描述

- 将 Agent 与环境的交互看做离散的时间序列
  s_0 ： Agent 感知到初始环境 s_0

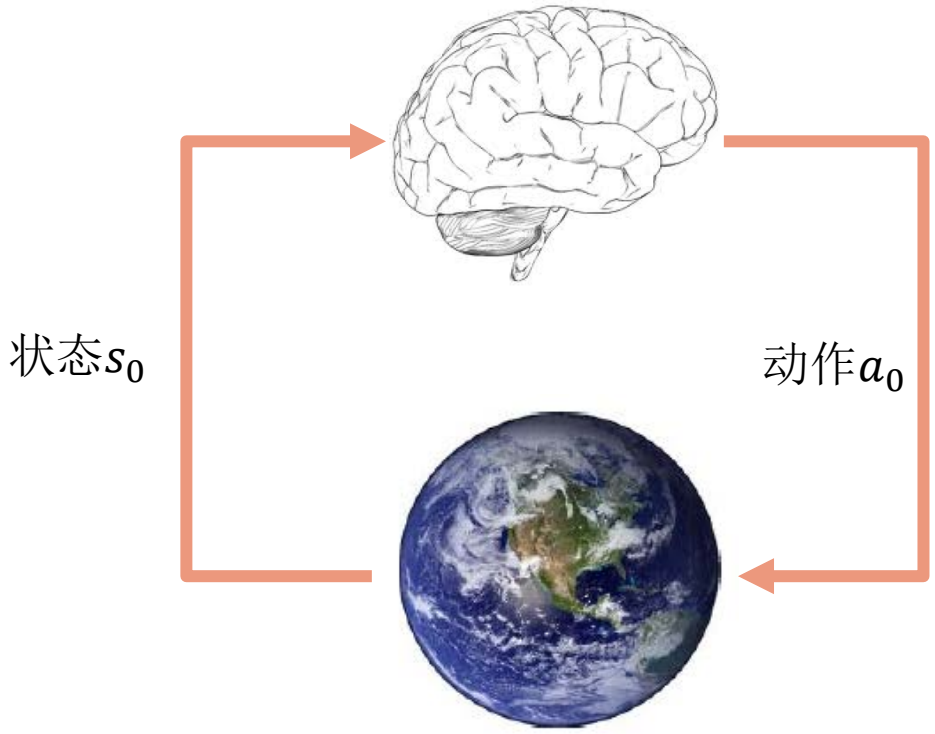


交互过程描述

- 将 Agent 与环境的交互看做离散的时间序列

  s_0 ： Agent 感知到初始环境 s_0

  a_0 ： Agent 根据状态 s_0 做出相应的动作 a_0



交互过程描述

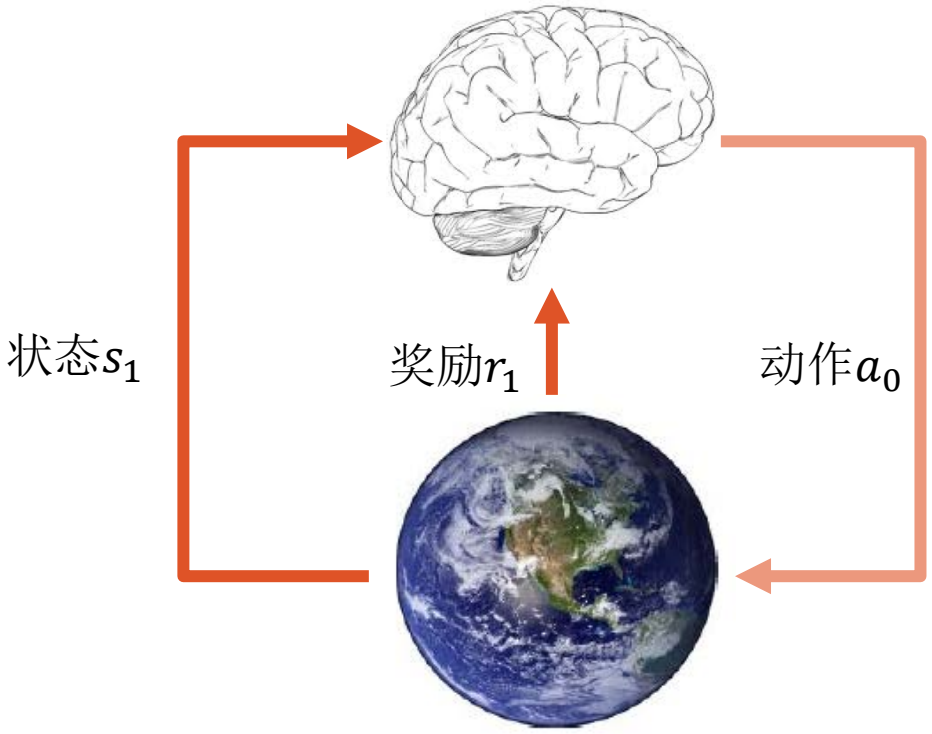
- 将 Agent 与环境的交互看做离散的时间序列

  s_0：Agent 感知到初始环境 s_0

  a_0：Agent 根据状态 s_0 做出相应的动作 a_0

  s_1：因为动作 a_0，环境发生改变到新状态 s_1

  r_1：环境反馈给 Agent 一个即时奖励 r_1



交互过程描述

- 将 Agent 与环境的交互看做离散的时间序列

  s_0：Agent 感知到初始环境 s_0

  a_0：Agent 根据状态 s_0 做出相应的动作 a_0

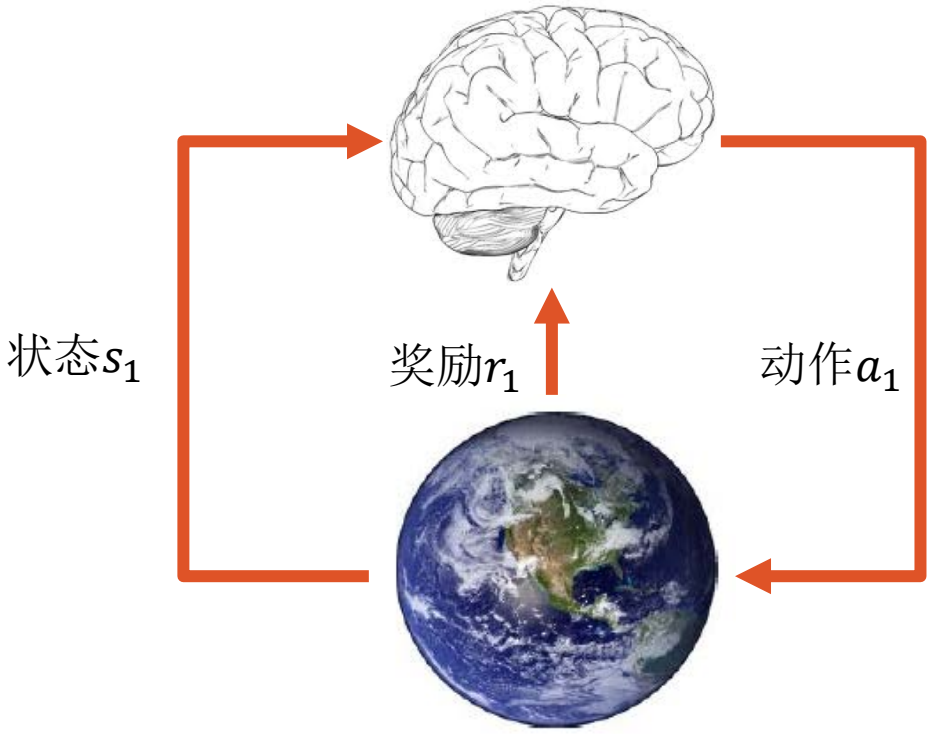
  s_1：因为动作 a_0，环境发生改变到新状态 s_1

  r_1：环境反馈给 Agent 一个即时奖励 r_1

  a_1：Agent 根据状态 s_1 做出相应动作 a_1

  ……

- 轨迹 $\tau = s_0, a_0, s_1, r_1, a_1, \ldots, s_{T-1}, a_{T-1}, s_T, r_T$



总回报（Return）

- 总回报（Return）：agent和环境一次交互过程的轨迹 τ 所收集到的累积奖励

$$G(\tau) = \sum_{t=0}^{T} r_{t+1} = \sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1})$$

- $T \neq \infty$：回合式任务（Episodic Task）
- $T = \infty$：持续式任务（Continuing Task），回报可能是无穷大，因此引入折扣率$\gamma \in [0,1]$来降低远期回报的权重

  折扣回报（Discounted Return）：

$$G(\tau) = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$$

目标

- 一个理性的 Agent ， 以达到**最好结果**为行动目标，如果无法达到最好结果，可以以获得**最大期望值的结果**为行动目标。
- Agent的任务是学习一个$\pi\ (a|s)$，要求此策略对 Agent 产生最大的总回报

目标函数（Object Function）

- 一个理性的 Agent，以达到**最好结果**为行动目标，如果无法达到最好结果，可以以获得**最大期望值的结果**为行动目标。
- Agent的任务是学习一个π $(a|s)$，要求此策略对 Agent 产生最大的总回报
- 学习一个策略$\pi_\theta(a|s)$来**最大化期望回报**（Expected Return），即希望agent执行一系列的动作来获得尽可能多的平均回报

$$\mathcal{J} = \mathbb{E}_{\tau \sim p_\theta(\tau)}[G(\tau)]$$

强化学习Agent的主要组成部分

一个强化学习Agent可能包括一个或多个以下组成部分：

- 值函数：每个环境状态或 Agent 动作表现得好不好
- 策略：Agent的行为函数
- 模型：Agent对环境的表示

名称	符号	描述
状态	s	对环境的描述，可以是离散的或连续的，状态空间为 $\mathcal{S}$
动作	a	对Agent行为的描述，可以是离散的或连续的，动作空间为$\mathcal{A}$
策略	$\pi(a\|s)$	Agent 根据环境状态 s 来决定下一步动作 a 的函数
状态转移概率	$p(s'\|s,a)$	Agent根据当前状态 s 做出动作 a 之后，环境在下一个时刻转变为状态 s' 的概率
即时奖励	$r(s,a,s')$	Agent根据当前状态 s 做出动作 a 之后，环境反馈给Agent的奖励。该奖励常和下一个时刻的状态 s' 有关

值函数

策略

模型

目标函数（Object Function）

- 强化学习的目标：学习一个策略 $\pi_\theta(a|s)$ 来最大化期望回报（Expected Return），即希望agent执行一系列的动作来获得尽可能多的平均回报
- $\mathcal{J} = \mathbb{E}_{\tau \sim p_\theta(\tau)}[G(\tau)]$，其中$\theta$为策略函数的参数

$$\mathbb{E}_{\tau \sim p(\tau)}[G(\tau)] = \mathbb{E}_{s \sim p(s_0)}\left[\mathbb{E}_{\tau \sim p(\tau)}\left[\sum_{t=0}^{T-1} \gamma^t r_{t+1} | \tau_{s_0} = s\right]\right] = \mathbb{E}_{s \sim p(s_0)}[V^\pi(s)]$$

从状态s开始，执行策略π得到的总回报的期望

状态值函数（State Value Function）

名称	符号	描述
状态	s	对环境的描述，可以是离散的或连续的，状态空间为S

- 初始状态为 s 时，执行策略 π 得到的期望总回报

$$V^{\pi}(s) = \mathbb{E}_{\tau \sim p(\tau)}\left[\sum_{t=0}^{T-1} \gamma^{t} r_{t+1} | \tau_{s_0} = s\right]$$

状态-动作值函数(State-Action Value Function)

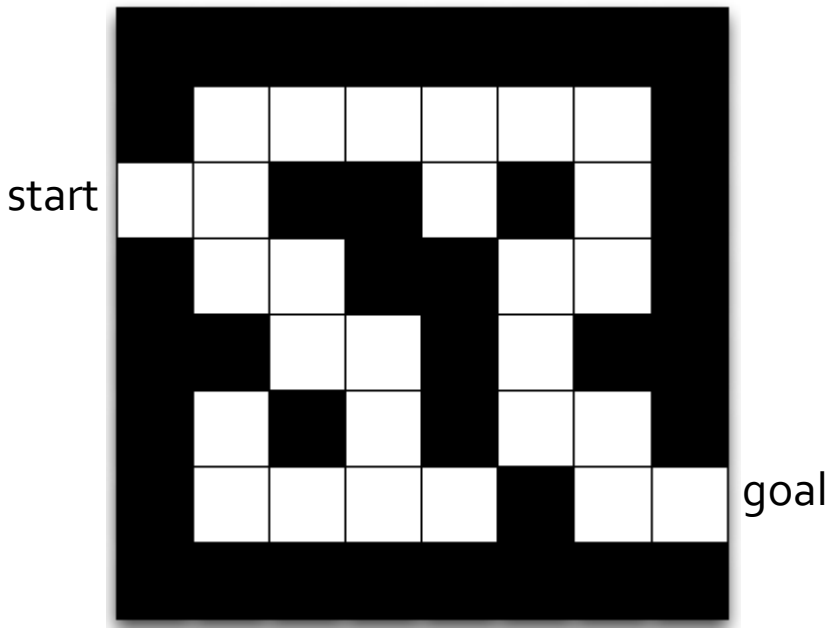
名称	符号	描述
状态	s	对环境的描述，可以是离散的或连续的，状态空间为$\mathcal{S}$
动作	a	对Agent行为的描述，可以是离散的或连续的，动作空间为$\mathcal{A}$

- 状态-动作值函数/Q函数：初始状态为 s，并进行动作 a，然后执行策略 π 得到的期望总回报

$$Q^{\pi}(s,a) = \mathbb{E}_{s' \sim p(s'|s,a)}[r(s,a,s') + \gamma V^{\pi}(s')]$$

基于值函数的Agent

- 走迷宫

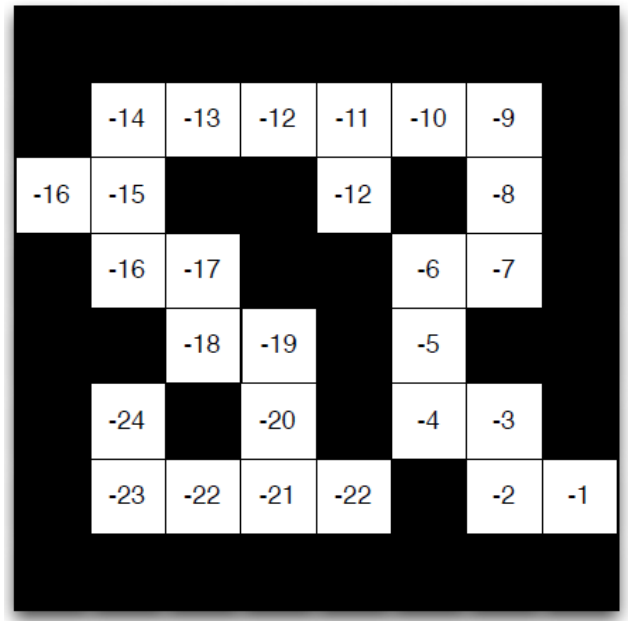


- 奖励：每走一步，得到-1的奖励
- 动作：上下左右
- 状态：Agent的位置
- 状态值函数：表示每个状态会返回的值

基于值函数的Agent

- 走迷宫

start



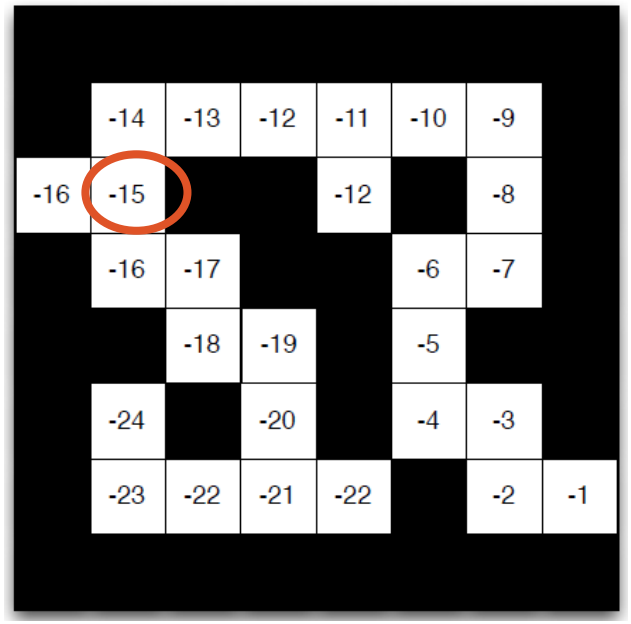
goal

- 奖励：每走一步，得到-1的奖励
- 动作：上下左右
- 状态：Agent的位置
- 状态值函数：表示每个状态会返回的值

基于值函数的Agent

- 走迷宫

start



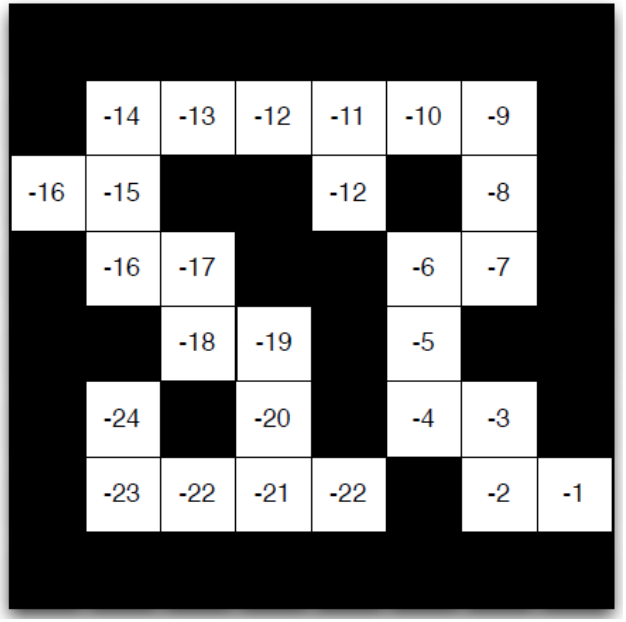
goal

- 奖励：每走一步，得到-1的奖励
- 动作：上下左右
- 状态：Agent的位置
- 状态值函数：表示每个状态会返回的值

基于值函数的Agent

- 显式地学习值函数
- 隐式地学习策略，策略是从学习到的值函数中推算得出

start



goal

- 奖励：每走一步，得到-1的奖励
- 动作：上下左右
- 状态：Agent的位置
- 状态值函数：表示每个状态会返回的值

策略（Policy）

名称	符号	描述
策略	$\pi(a\|s)$	Agent 根据环境状态 s 来决定下一步动作 a 的函数。**输入是状态 s，输出是行为 a 的概率。**

- 确定性策略（deterministic policy）
  - 从状态空间到动作空间的映射函数 $\pi: \mathcal{S} \to \mathcal{A}$
- 随机性策略（stochastic policy）
  - 在给定环境状态时，agent 选择某个动作的概率分布

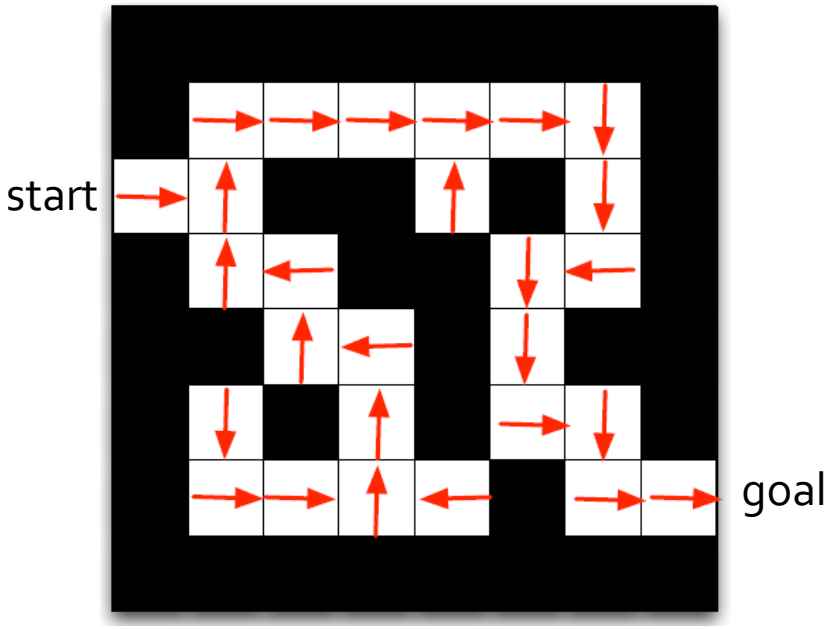
$$\pi(a|s) \triangleq p(a|s),$$

$$\sum_{a \epsilon \mathcal{A}} \pi(a|s) = 1$$

  - 可以更好地探索环境；使动作具有多样性，策略不易被对手预测

基于策略的Agent

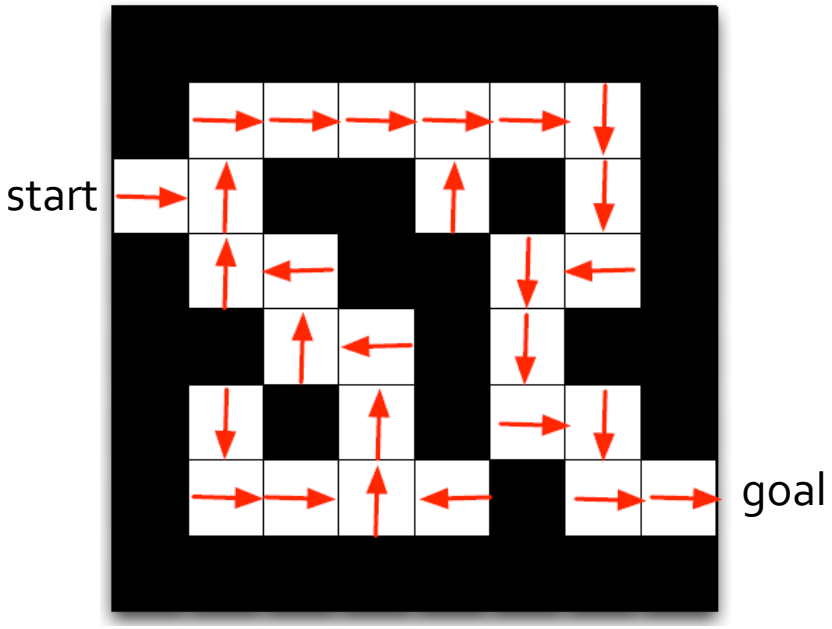
- 走迷宫



- 奖励：每走一步，得到-1的奖励
- 动作：上下左右
- 状态：Agent的位置
- 最佳策略：在每一个状态，会得到一个最佳的行为

基于策略的Agent

- 直接学习策略，给定一个状态，输出执行相应动作的概率
- 不学习值函数

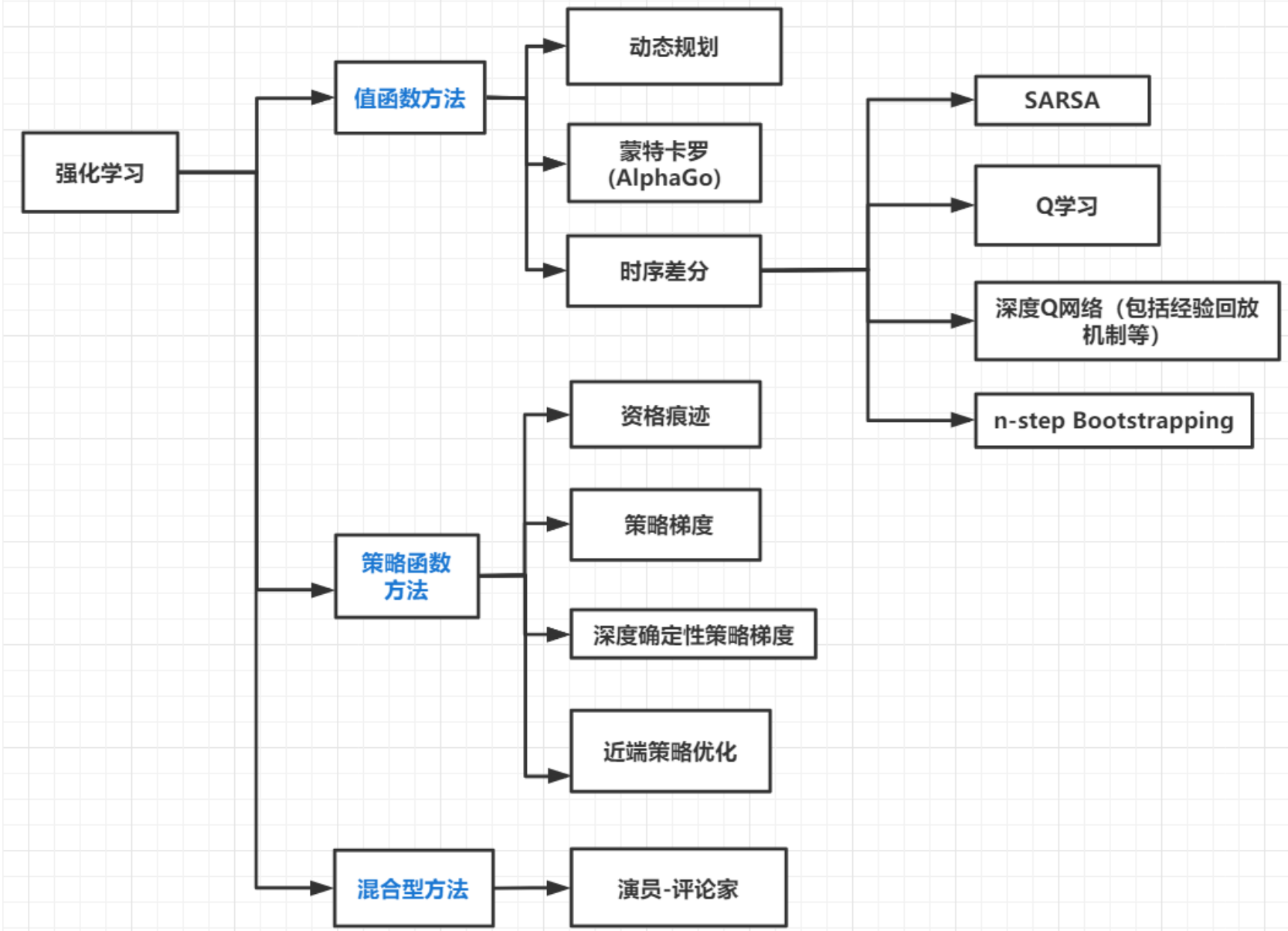


- 奖励：每走一步，得到-1的奖励
- 动作：上下左右
- 状态：Agent的位置
- 最佳策略：在每一个状态，会得到一个最佳的行为

基于值函数 vs. 基于策略

	策略	值函数	实现方式	适用场景
基于值函数	无需指定显式的策略	显式学习值函数	维护一个值表格或者值函数，通过值表格或值函数选取使得值最大的动作	离散的环境
基于策略	需要制定显式的策略	不学习值函数	直接对策略进行优化，使得制定的策略能够获得最大的奖励	集合规模庞大、动作连续的场景
混合型方法	学习策略	学习值函数	Agent根据策略执行动作，值函数会对做出的动作给出相应的值	

导图


强化学习
值函数方法
动态规划
蒙特卡罗
(AlphaGo)
时序差分
SARSA
Q学习
深度Q网络（包括经验回放
机制等）
n-step Bootstrapping
策略函数
方法
资格痕迹
策略梯度
深度确定性策略梯度
近端策略优化
混合型方法
演员-评论家

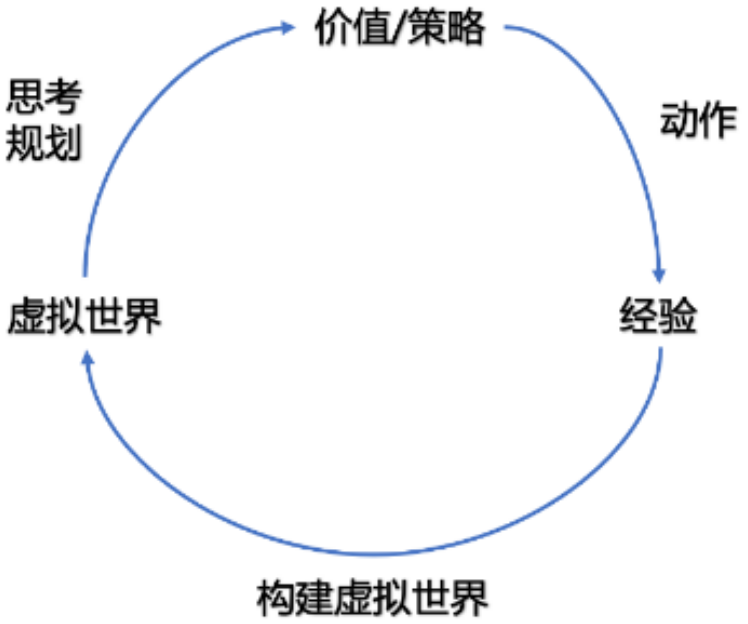
模型（Model）

名称	符号	描述
状态转移概率	$p(s'\|s,a)$	Agent根据当前状态 s 做出动作 a 之后，环境在下一个时刻转变为状态 s' 的概率
即时奖励	$r(s,a,s')$	Agent根据当前状态 s 做出动作 a 之后，环境反馈给Agent的奖励。该奖励常和下一个时刻的状态 s' 有关

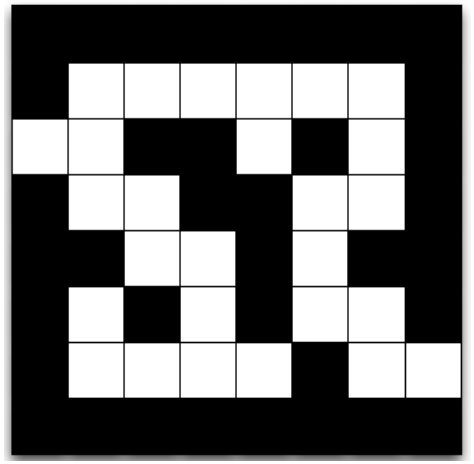
- 模型决定了下一个状态会是什么样，即下一步的状态取决于当前的状态以及当前采取的行为
- 基于模型的 Agent 通过学习状态的转移来采取动作

基于模型的方法

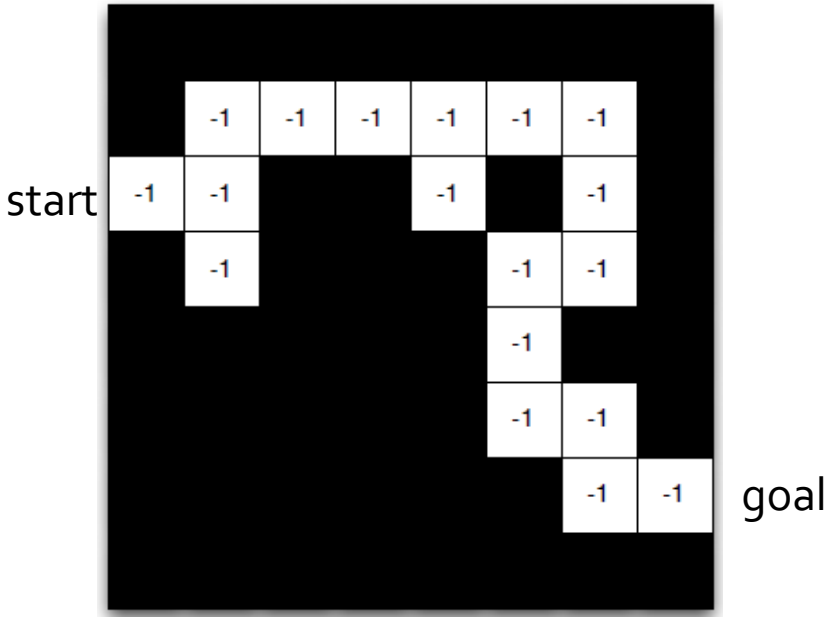
- Model-based：根据环境中的经验，构建虚拟世界，同时在真实环境和虚拟世界中学习。
- Model-free：不对真实环境进行建模，直接与真实环境进行交互以学习得到最优策略。
- 差别仅仅在于是否对真实环境进行建模
- 如何区分：Agent 执行动作前，是否能对下一步的状态和奖励进行预测



基于模型的Agent



- 直接学习策略，给定一个状态，输出执行相应动作的概率
- 不学习值函数



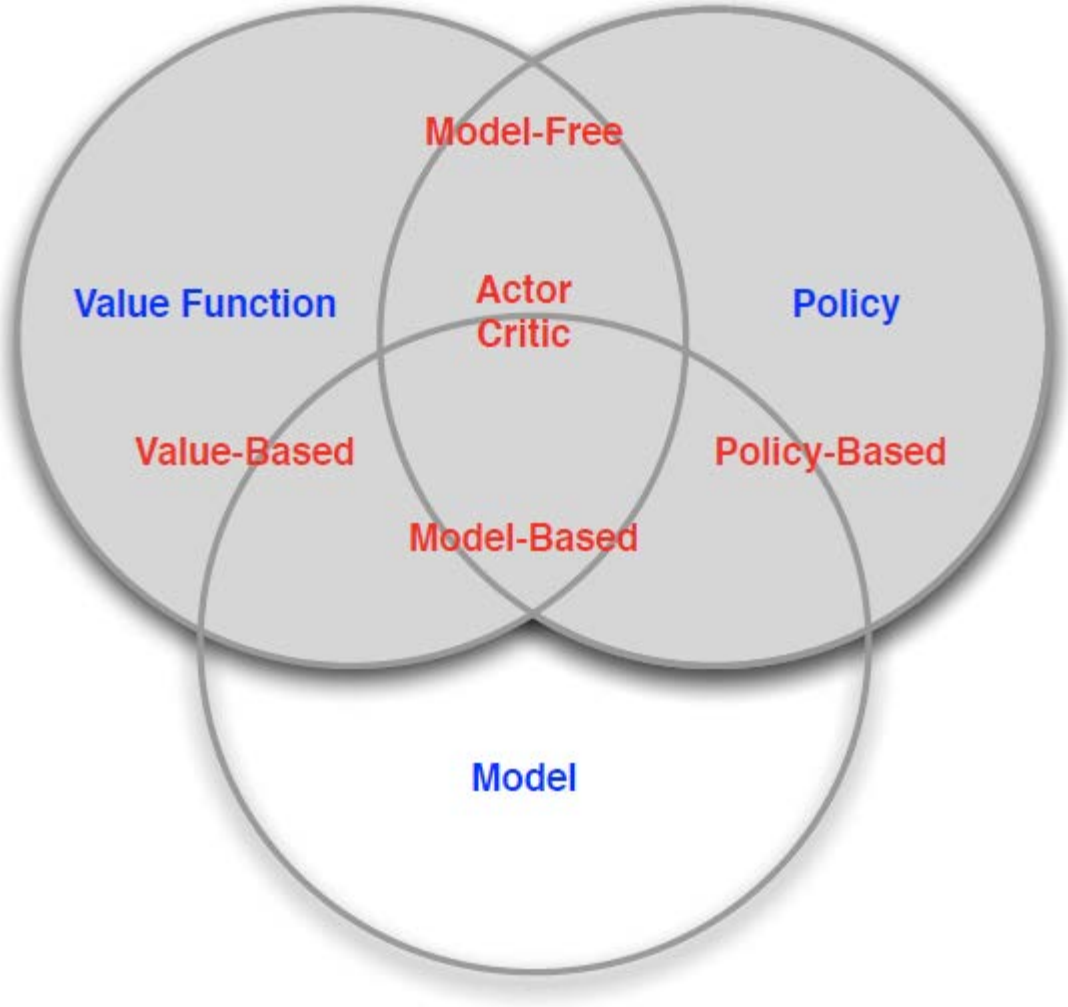
- 奖励：每走一步，得到-1的奖励
- 动作：上下左右
- 状态：Agent的位置
- 最佳策略：在每一个状态，会得到一个最佳的行为

有模型 vs. 无模型

	优点	缺点
Model-based	具有“想象能力”	虚拟世界与真实环境之间存在差异，限制了泛化性
Model-free	泛化性较好	只能一步一步地采取策略，等待真实环境的反馈

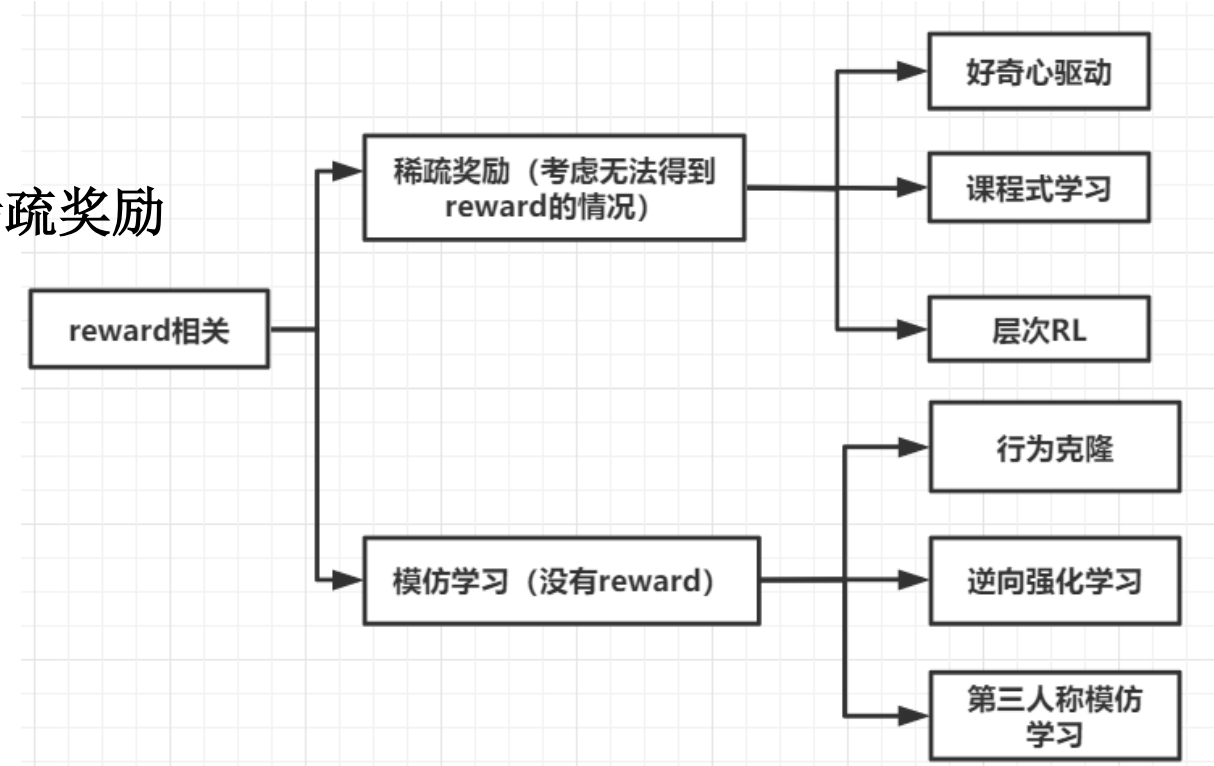
- 目前的强化学习研究中，大部分情况下，环境是**静态的**、**可描述的**，Agent的状态是**离散的**、**可观察的**，不需要状态转移函数和奖励函数，直接采用model-free方法即可。

分类


Model-Free
Value Function
Actor
Critic
Policy
Value-Based
Policy-Based
Model-Based
Model

奖励reward

- 奖励是标量反馈信号，反映出在某个时间步Agent表现得如何
- Agent的工作就是使得奖励最大化
- 问题
  - 大多数情况下都没有办法得到奖励：**稀疏奖励**
  - 没有奖励：**模仿学习**



目录

- 人工智能与Agent
- 强化学习定义
- 重要概念介绍
- 汇报安排

汇报安排

- 网页： index (wds.ac.cn)
- 时间：周二、周四晚上8点
- 地点：腾讯会议

周二会议ID：335 7536 2830

周四会议ID：909 7688 3626

第一部分：基础知识介绍

1. 2021.7.6 强化学习相关基础知识介绍. **商小雨**
2. 2021.7.6 马尔科夫决策过程. **李蕾**

第二部分：值函数方法

1. 2021.7.8 动态规划. **蔡新宇**
2. 2021.7.8 蒙特卡罗（结合AlphaGo）. **李鑫**
3. 2021.7.13 时序差分：SARSA. **张炯**
4. 2021.7.13 时序差分：Q学习. **冯羽茜**
5. 2021.7.15 时序差分：深度Q网络（包括经验回放机制等）. **胡家欣**
6. 2021.7.15 时序差分：n-step Bootstrapping. **戴鑫邦**

第三部分：策略函数方法

1. 2021.7.20 资格痕迹. **吴楚仪**
2. 2021.7.20 策略梯度. **庄玥**
3. 2021.7.22 深度确定性策略梯度. **刘立恒**
4. 2021.7.22 近端策略优化. **王瑛**

第四部分：混合型方法

1. 2021.7.27 演员-评论家. **吴博**

第五部分：reward相关

1. 2021.7.27 稀疏奖励：好奇心驱动. **陈海燕**
2. 2021.7.29 稀疏奖励：课程式学习. **袁书伟**
3. 2021.7.29 稀疏奖励：层次强化学习. **刘伟翼**
4. 2021.8.3 模仿学习：行为克隆. **曹俊**
5. 2021.8.3 模仿学习：逆向强化学习. **耿飚**
6. 2021.8.5 模仿学习：第三人称模仿学习. **任艳杰**

谢 谢