强化学习

Reinforcement learning

第一节

强化学习导论

张世周

Outlines

- **1.1** 课程要求
- **1.2** 强化学习概述
- **1.3** 关于强化学习的相关问题
- **1.4**强化学习智能体内部
- **1.5**强化学习中的一些问题

1.1 课程要求

- 课程安排
- **9-16周：周一、周三11-12节（19：00---20：40）**
- 教学西楼**B**座教西**B212**

- **szzhang@nwpu.edu.cn**
- **https://teacher.nwpu.edu.cn/szzhang**

1.1 课程要求

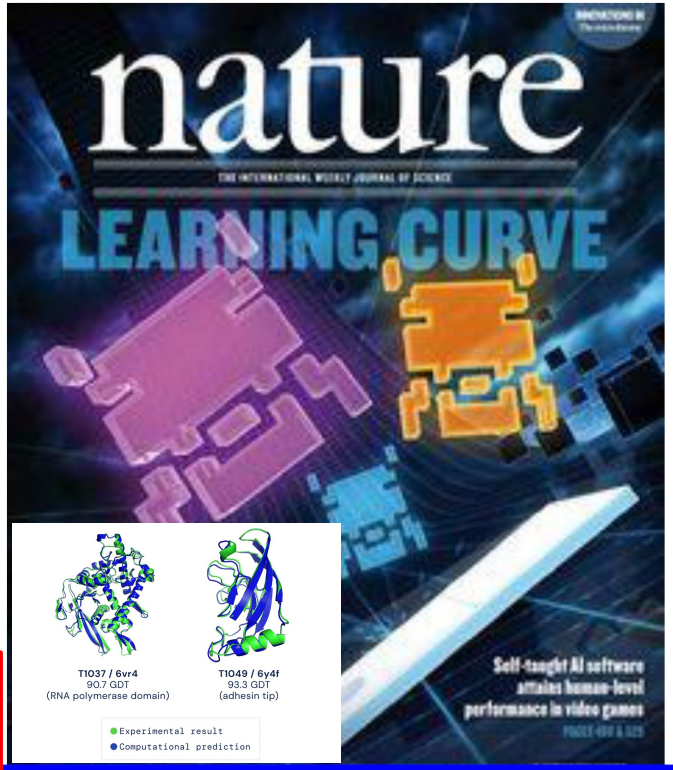
- 考试评价
- 10%课堂表现+40%课后作业+50%闭卷考试

1.1 课程要求

- 参考教材：
- 《强化学习》第二版，Richard Sutton & Andrew Barto著，俞凯 等译，中国工信出版集团，电子工业出版社
- https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf
- 《强化学习》网络公开课，by David Silver
- 《深度强化学习》系列讲座，李宏毅
- <Algorithms for Reinforcement learning>, Szepesvari
- https://sites.ualberta.ca/~szepesva/rlbook.html

强化学习问题描述

- DQN在绝大部分Atari游戏上战胜人类玩家
- AlphaGo先后战胜人类围棋冠军李世石、柯洁
- AlphaFold2在国际蛋白质结构预测大赛中取得绝对优势（40->92.4），在AI和结构生物学届“一石激起千层浪”



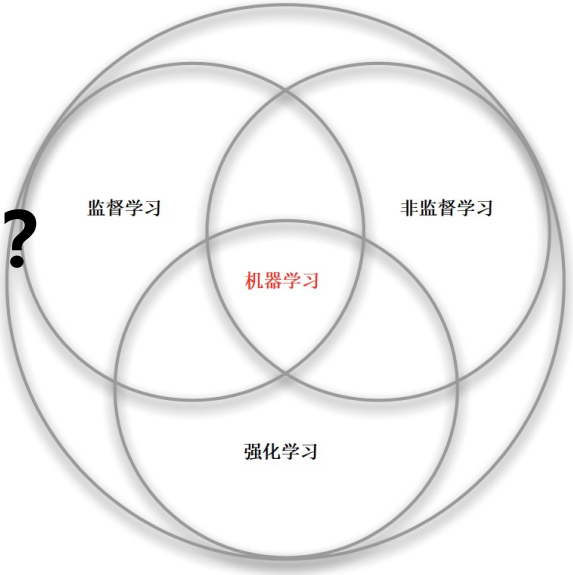
Deep Reinforcement Learning: AI = RL + DL

潜心研究基础算法，掌握“卡脖子”核心技术
探索学科交叉，尝试用所学知识去解决世界前沿问题
---立大志，做大事

强化学习问题描述

强化学习就是学习“做什么（即如何把当前的情境映射成动作）才能使得数值化的收益信号最大化”---Sutton,Barto

强化学习是机器学习中的一个领域，强调如何基于环境而行动，以取得最大化的预期利益。---Wikipedia



是什么使强化学习不同于其他两种机器学习范式?

强化学习的特点：

1、强化学习没有监督者，只有收益信号

2、收益是延时的

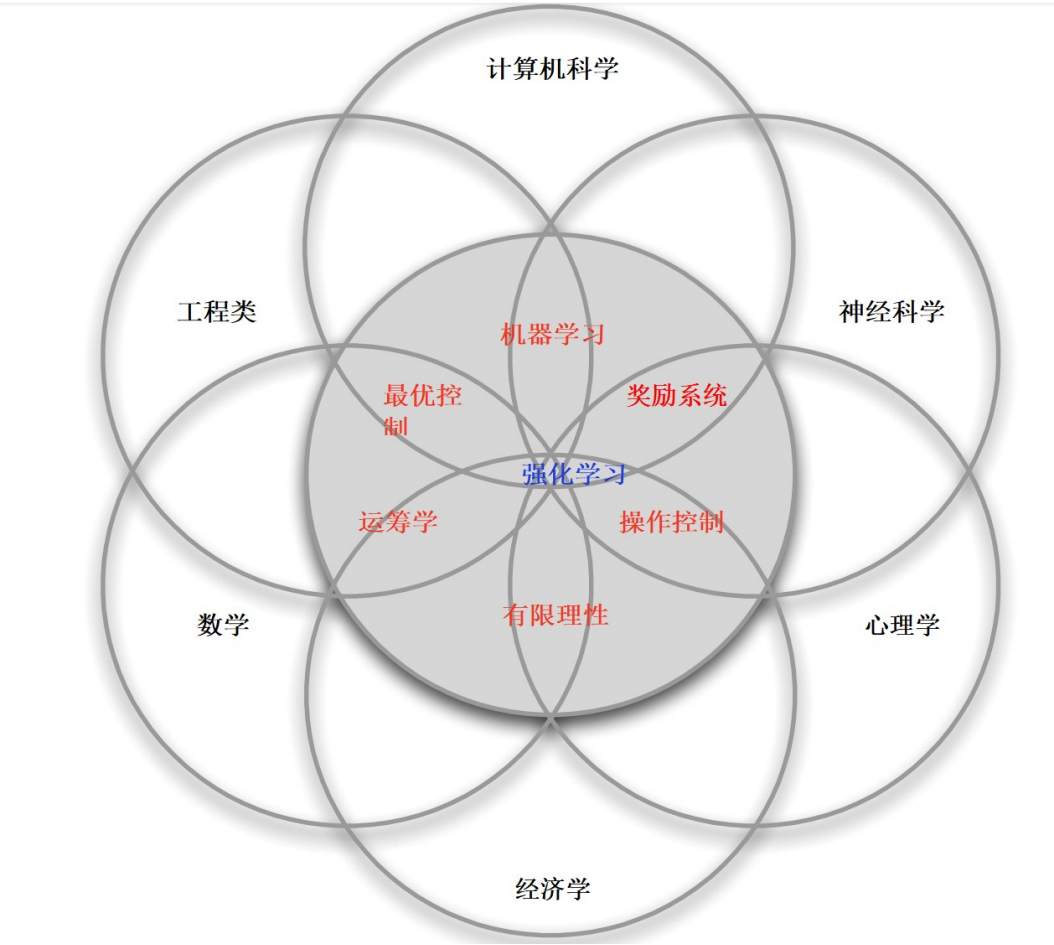
3、时间是非常重要的 (数据是连续的，不是独立同分布的数据)

4、智能体的行动会影响之后其接收的数据

交互式学习

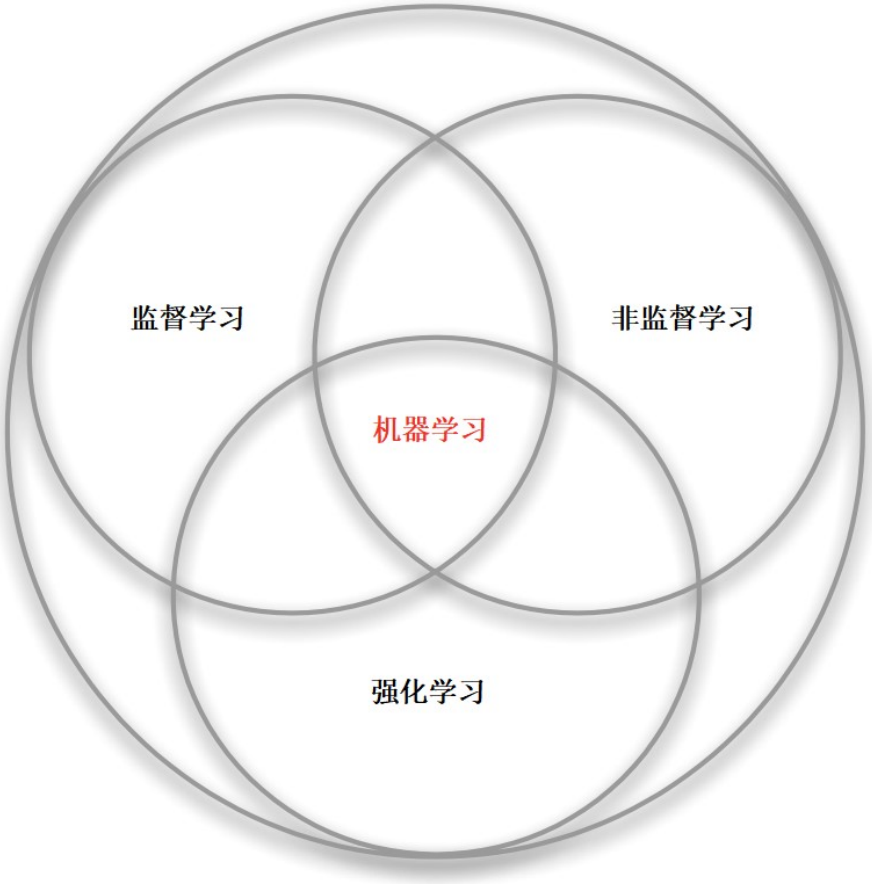
1.2 强化学习概述

强化学习的应用领域



1.2 强化学习概述

机器学习的分支



1.2 强化学习概述

是什么使强化学习不同于其他两种机器学习范式？

强化学习的特点：

1、强化学习没有监督者，只有收益信号

2、收益是延时的

3、时间是非常重要的 (数据是连续的，不是独立同分布的数据)

4、智能体的行动会影响之后其接收的数据

1.2 强化学习概述

关于强化学习的例子：

1、直升机特技表演

2、在西洋双陆棋比赛中击败世界冠军

3、管理投资组合

4、控制电站

5、控制一个人型机器人行走

6、在Atari游戏中，得分超过人类选手

1.2 强化学习概述

直升机特技表演

1.2 强化学习概述

赛车游戏

1.2 强化学习概述

Atari

1.2 强化学习概述

收益（reward）

1、收益 R_t 是一个标量的反馈信号

2、其用来表明智能体在步骤t中的表现如何

3、智能体的工作就是最大化累计收益

强化学习建立在奖励假说之上

奖励假说(Reward Hypothesis)定义：

所有的目标可以被最大化累计期望收益来描述

你同意这种说法吗？

1.2 强化学习概述

关于收益的一些例子：

1、直升机做飞行特技动作：在做了一个特技动作后获得一定的收益，但坠毁后则得到负值很大的收益。

2、在下西洋双陆棋时：在击败对手后获得收益而在被击败时则得到负的收益。

3、管理投资组合：在赚钱时获得收益，亏钱时则失去收益。

4、控制电站运行：发电时会获得收益，但发生安全事故时会失去收益。

5、控制人型机器人行走：在走出正常步伐时获得收益，跌倒时会失去收益 。

1.2 强化学习概述

关于序贯决策问题（sequential decision making）

目标：选择动作去最大化总的收益

特点：

1、动作可能有一个长期效应

2、延时收益

3、牺牲眼前的收益去获取长远的收益可能会使总的收益更高

例子：

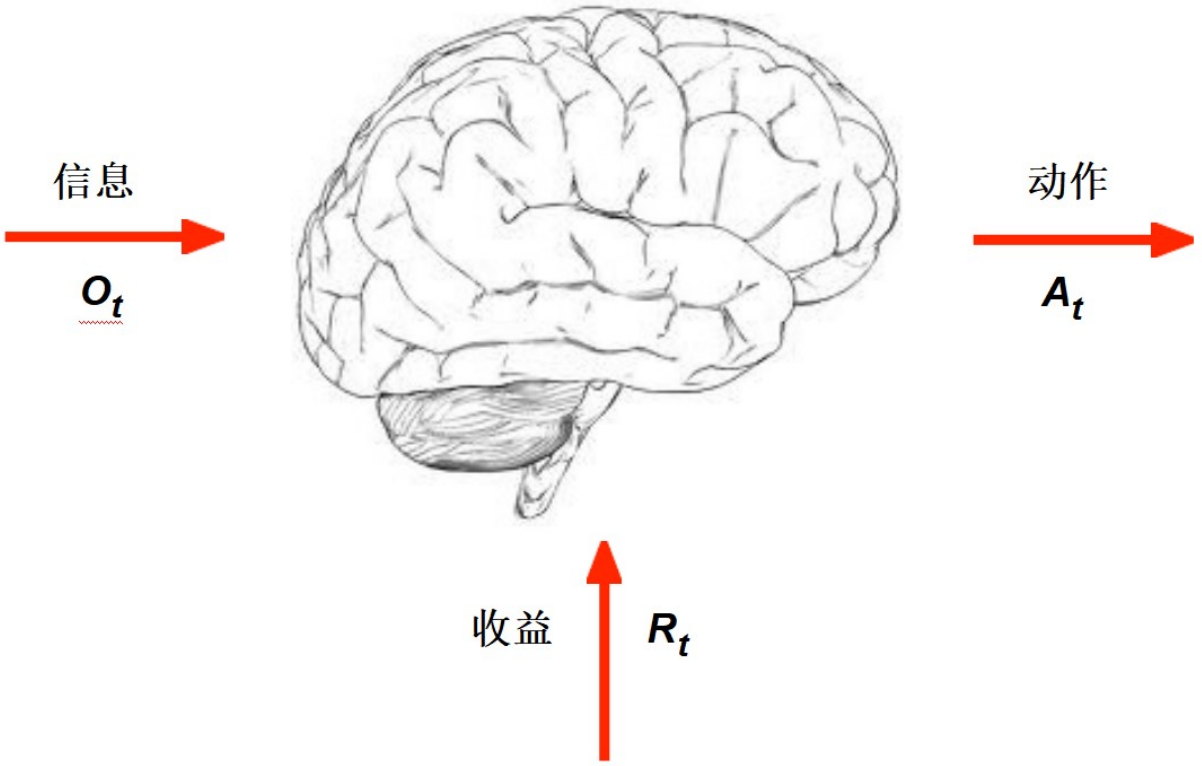
1、一笔金融投资可能会在几个月后才能看到收益

2、给做特技的直升机加油可以避免之后因缺油导致的坠毁

3、阻挡住对手的一步棋可以使自己从现在起获得优势

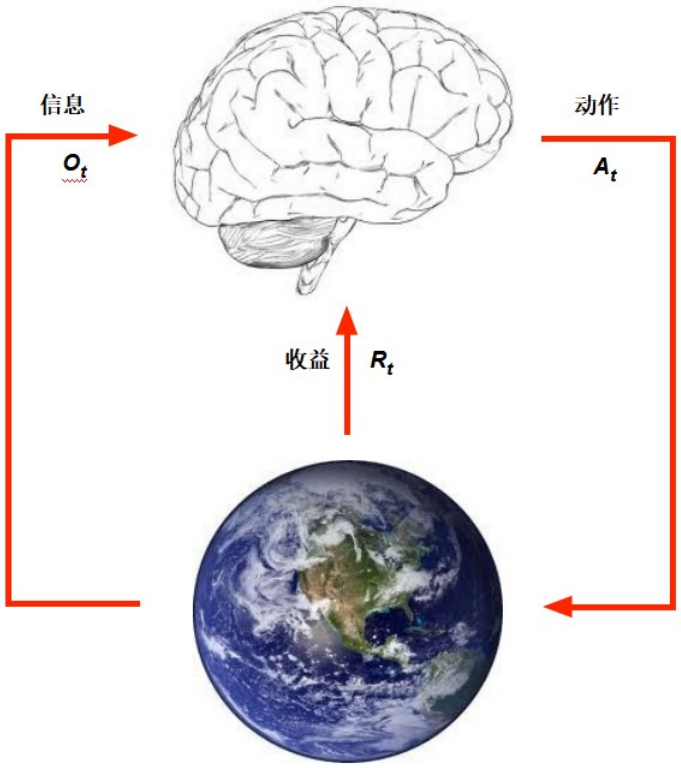
1.3 关于强化学习的相关问题

智能体和环境



1.3 关于强化学习的相关问题

智能体和环境



每个步骤t智能体都会：

1、执行动作At

2、获取信息 Ot

3、获取标量收益 Rt

环境都会:

1、接受动作 At

2、释放信息 Ot+1

3、给智能体收益 Rt+1

随着环境的改变t会增加

1.3 关于强化学习的相关问题

历史和状态

历史包含过去智能体获取的信息，收益和执行的动作

$$H_t = O_1, R_1, A_1, \ldots, A_{t-1}, O_t, R_t$$

i.e. 到时间**t**的所有可观测变量

i.e. 机器人或具身**AI**的传官数据流

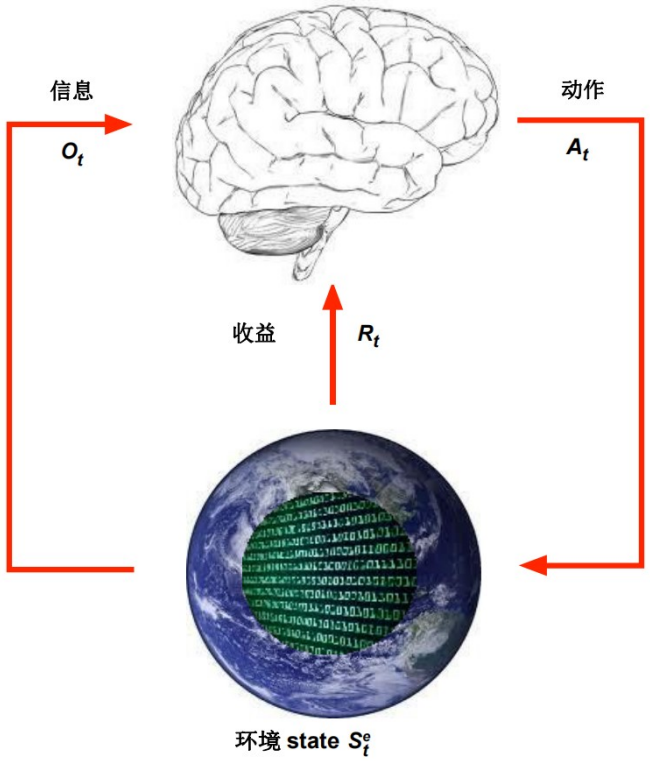
智能体选择的动作和环境将给出的信息和收益取决于历史

状态是用来决定接下来发生什么的信息。通常，状态和历史构成以下这个函数：

$$S_t = f(H_t)$$

1.3 关于强化学习的相关问题

环境状态



环境状态有以下几个特征：

1、环境状态 S_t^e 是环境的私有状态

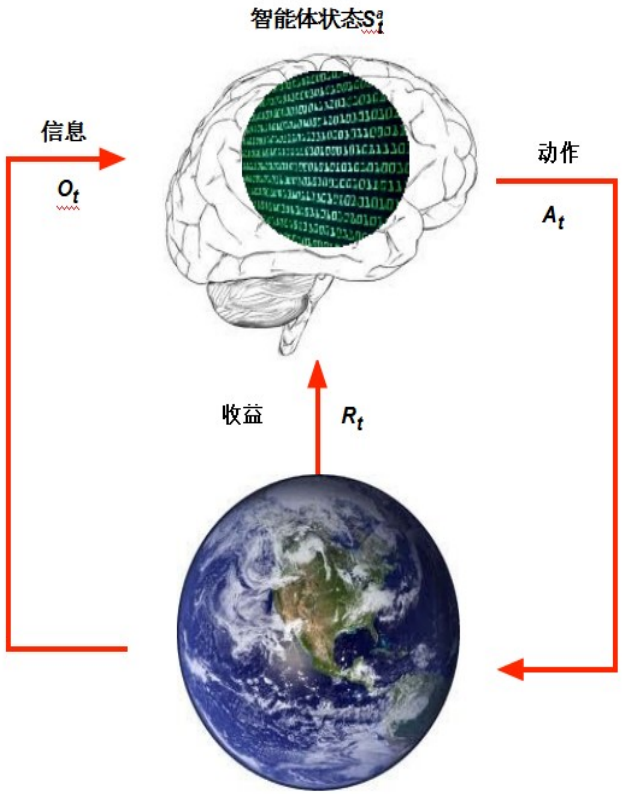
2、环境所采用数据，用以决定智能体的下一个观测/奖励

3、环境状态对智能体来说通常是不可见的

4、即使 S_t^e 是可见的，它也有可能（一般）包含无用的信息

1.3 关于强化学习的相关问题

智能体状态



智能体状态有以下几个特征：

1、智能体状态 S_t^a 是智能体的内部状态

2、是智能体用来决定下一个动作所采用的信息

3、是强化学习算法中使用的数据

它与history构成以下的这个函数：

$$S_t^a = f(H_t)$$

1.3 关于强化学习的相关问题

信息状态

信息状态 (又称马尔可夫状态) 包含历史中所有有用的信息

定义：

如果一个状态 S_t 拥有马尔可夫性，则当且仅当满足以下公式

$$P[S_{t+1} \mid S_t] = P[S_{t+1} \mid S_1, \ldots, S_t]$$

“未来在给定当前状态的条件下，和过去无关”

$$H_{1:t} \rightarrow S_t \rightarrow H_{t+1:\infty}$$

1、一旦已知当前的状态，过去的信息就可以被丢弃。

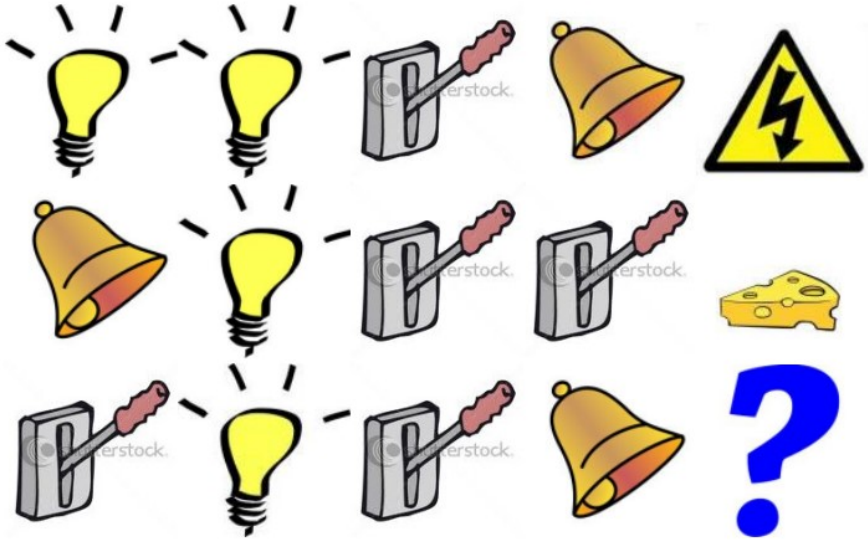
2、状态中包含了预测未来的足够的信息

3、环境状态 S_t^e 是具有马尔可夫性的

4、历史 H_t 也是具有马尔可夫性的

1.3 关于强化学习的相关问题

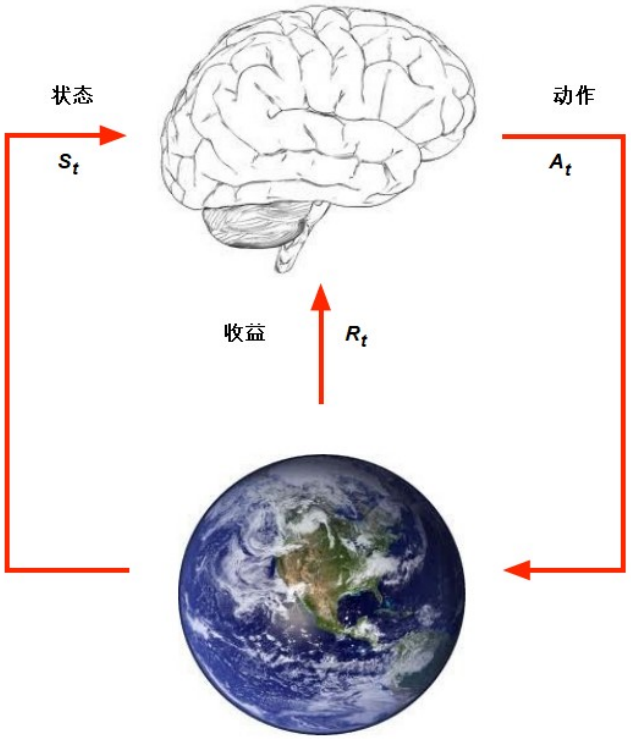
实验鼠



- 如果老鼠的状态取决于顺序中的最后3项会怎样？
- 如果老鼠的状态取决于铃铛，灯和开关的数量会怎样
- 如果老鼠的状态取决于完整序列会怎样？

1.3 关于强化学习的相关问题

完全可观测环境



完全可观测：智能体可以直接观测环境状态

$$O_t = S_t^a = S_t^e$$

- 智能体状态 = 环境状态 = 信息状态
- 通常，这是一个马尔可夫决策过程(MDP)
- (下一节会详细介绍这个马尔可夫决策过程)

1.3 关于强化学习的相关问题

部分可观测环境

部分可观测： 智能体只能获取环境的一部分信息

- 靠摄像头获取视觉信息的机器人无法知道其绝对位置
- 贸易智能体只能获取价格等有关信息
- 一个打扑克的智能体只能观察到打出去的公共牌

在部分可观察环境下智能体状态 != 环境状态

通常这是一个部分可观测马尔可夫决策过程 (POMDP)

智能体必须构造自己的状态表示 S_t^{a}，分情况有以下几种：

- 状态=history： $S_t^{\mathrm{a}} = H_{\mathrm{t}}$
- 循环神经网络： $S_t^a = \sigma(S_{t-1}^{\mathrm{a}} W_s + O_t W_o)$
- Beliefs of environment state： $S_t^{\mathrm{a}} = (P[S_t^e = s^1], \ldots, P[S_t^e = s^n])$

1.4 强化学习智能体内部

RL智能体的主要组成部分

一个RL智能体可能包含以下一个或多个组成部分：

- 策略Policy：决定智能体行为的机制
- 价值函数Value function：评价某一状态或者行为的好坏
- 模型Model：智能体对环境的建模

1.4 强化学习智能体内部

策略policy

- 策略是决定智能体行为的机制
- 是从状态到行为的一个映射，即在某一个状态下选择某一个行为
- 策略可以是确定性的，如 $a = \pi(s)$,即在某个状态下一定采取这个行为
- 策略也可以是不确定性的,如 $\pi(a|s) = P[A_t = a | S_t = s]$,即在某个状态下执行某种行为的（试探**Exploration**和开发**Exploitation**）

1.4 强化学习智能体内部

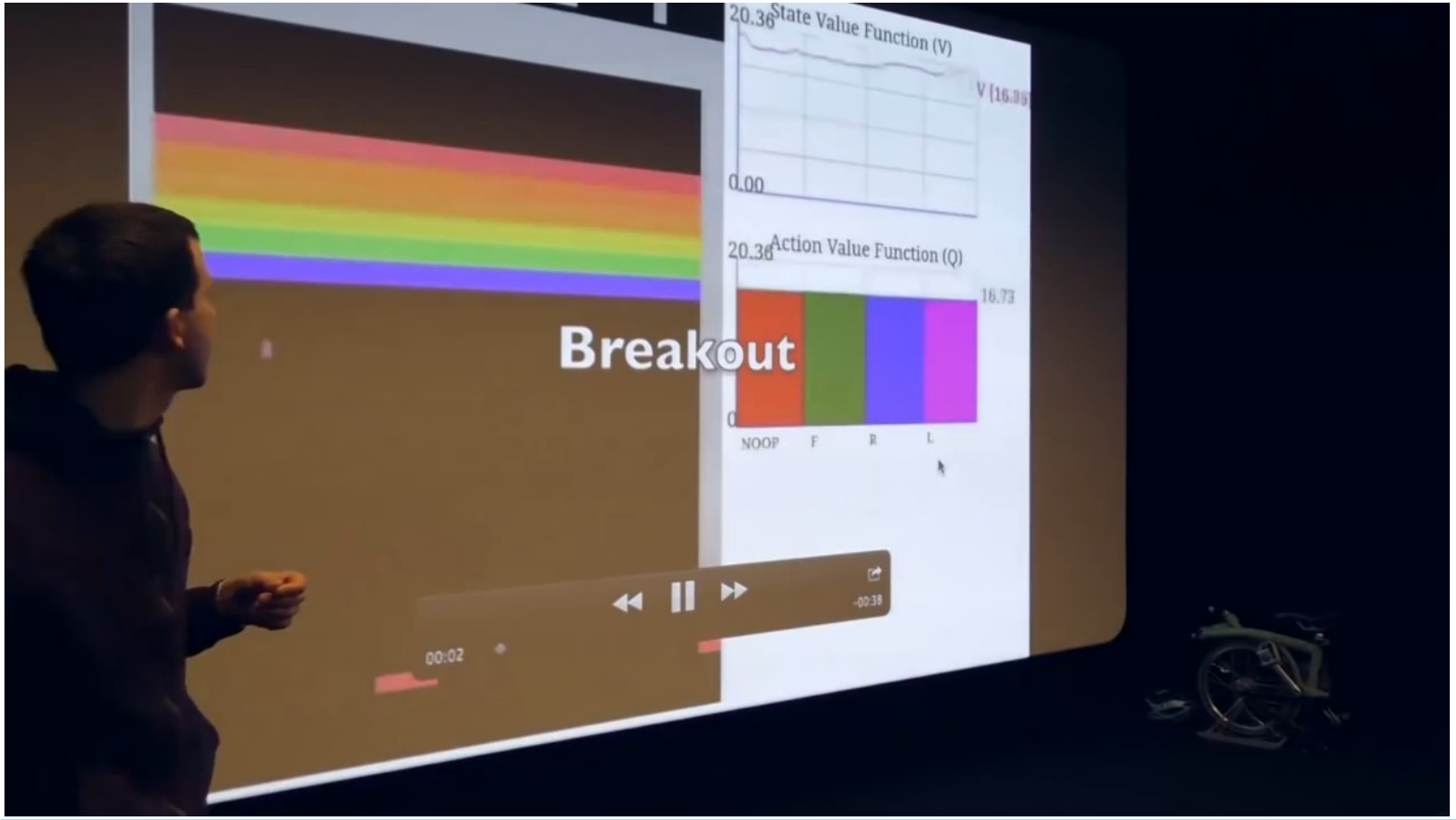
价值函数Value Function

- 价值函数是一个未来奖励的预测，用来评价当前状态的好坏程度。当面对两个不同的状态时，个体可以用一个**Value**值来评估这两个状态可能获得的最终奖励区别，继而指导选择不同的行为，即制定不同的策略。

- 价值函数表达：$v_\pi = E_\pi[R_{t+1} + \lambda R_{t+2} + \lambda^2 R_{t+3} \ldots | S_t = S]$ 在下一讲会对这个公式进行展开解释。

1.4 强化学习智能体内部

在Atari游戏中的价值函数



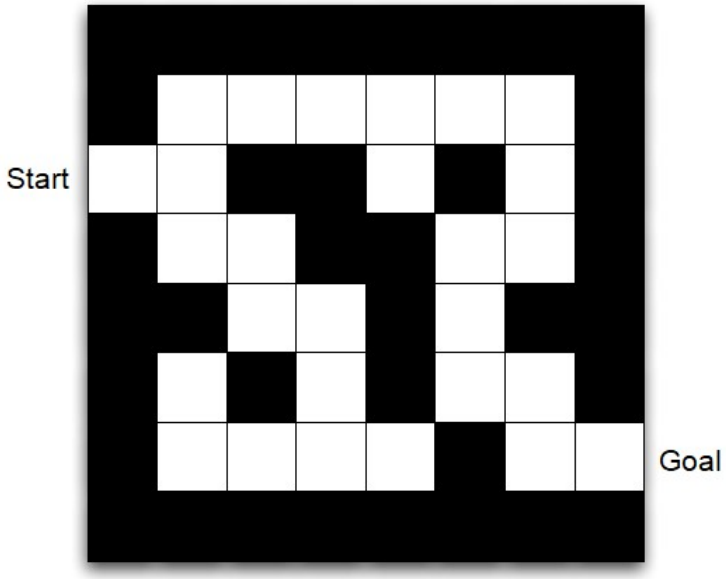
1.4 强化学习智能体内部

模型model

- 模型是智能体对环境的一个建模，它体现了智能体是如何思考环境运行机制的，智能体希望模型能模拟环境与个体的交互机制,这样就可以不需要真正的环境了。
- 模型可以预测环境接下来会发生什么,具体来说包含两部分,即预测下一个可能的状态发生的概率 $P_{ss'}^{a}=P[S_{t+1}=s'|S_t=s,A_t=a]$,和下一个可能获得的即时奖励 $R_s^a=E[R_{t+1}|S_t=s,A_t=a]$ 。
- 模型仅针对智能体而言，环境实际运行机制不称为模型，而称为环境动力学(dynamics of environment)，它能够明确确定个体下一个状态和所得的即时奖励。

1.4 强化学习智能体内部

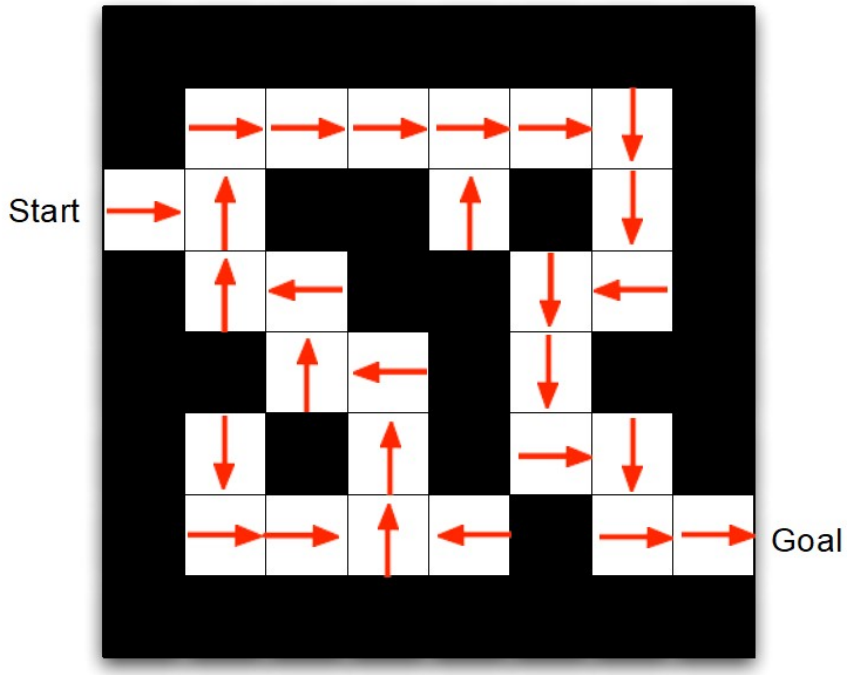
迷宫游戏



- 收益：每走一步收益减一
- 动作区间：上下左右
- 状态：智能体的位置

1.4 强化学习智能体内部

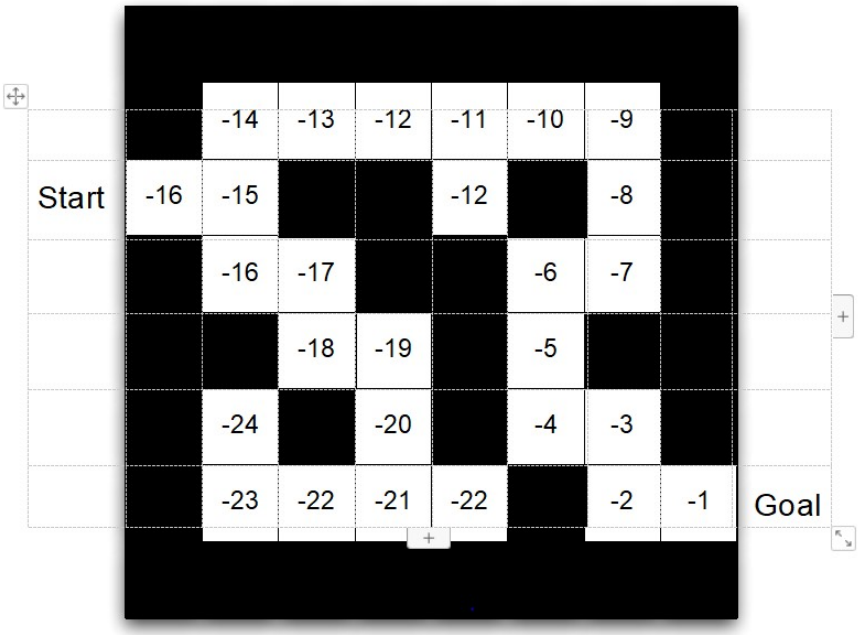
迷宫游戏中的策略



- 箭头代表每个状态s的策略 $\pi(s)$

1.4 强化学习智能体内部

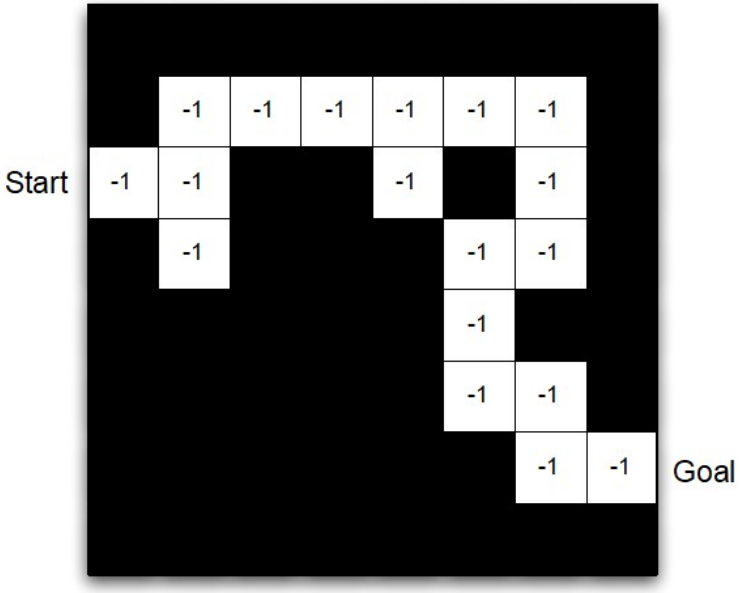
迷宫游戏中的价值函数



- 数字代表每个状态s的价值 $v_\pi(s)$

1.4 强化学习智能体内部

迷宫游戏中的模型



- 智能体可能在内部对环境构建了一个模型
- 动力学：模型会研究动作是怎么改变状态的
- 收益：每个状态获得的收益
- 模型可能是不完美的

- 网格布局表示状态转移模型 $P^{a}_{ss'}$
- 数字代表从每个状态s中的即时收益 R^{a}_{s}

RL智能体的分类 (1)

1、仅基于价值函数的智能体

在这样的智能体中，有对状态的价值估计函数，但是没有直接的策略函数，策略函数由价值函数间接得到。

2、仅直接基于策略的智能体

这样的智能体中行为直接由策略函数产生，个体并不维护一个对各状态价值的估计函数。

3、执行者-评判者（Actor-Critic）形式

智能体既有价值函数、也有策略函数。两者相互结合解决问题。

1.4 强化学习智能体内部

RL智能体的分类 (2)

1、不基于模型的智能体（model-free）

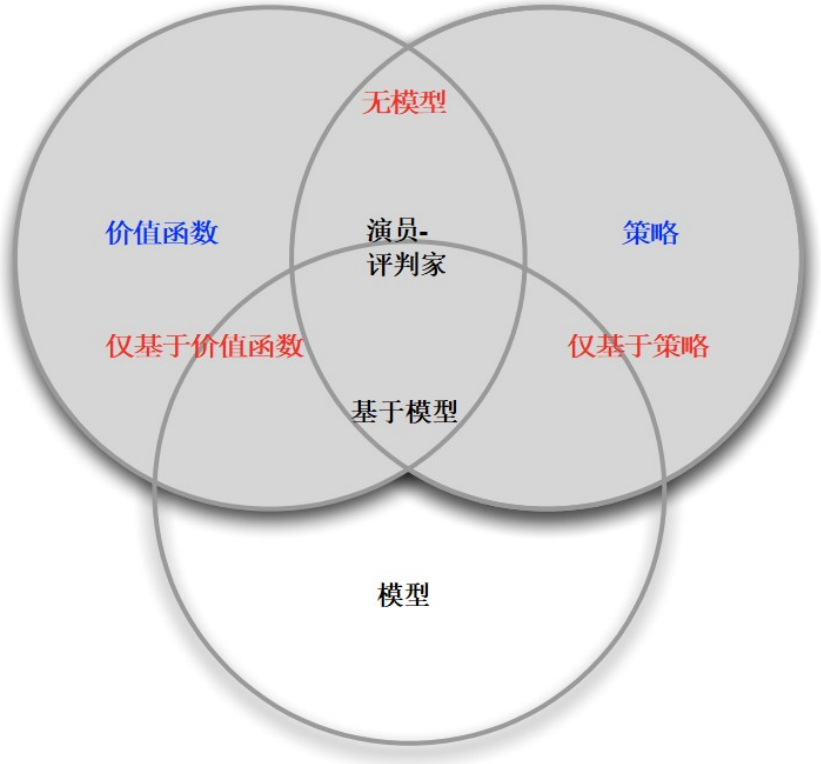
这类智能体并不想了解环境如何工作，而仅聚焦于价值和/或策略函数

2、基于模型的智能体(model-based)

智能体尝试建立一个描述环境运作过程的模型，以此来指导价值或策略函数的更新。

1.4 强化学习智能体内部

RL智能体分类图（术语）



1.5 强化学习中的一些问题

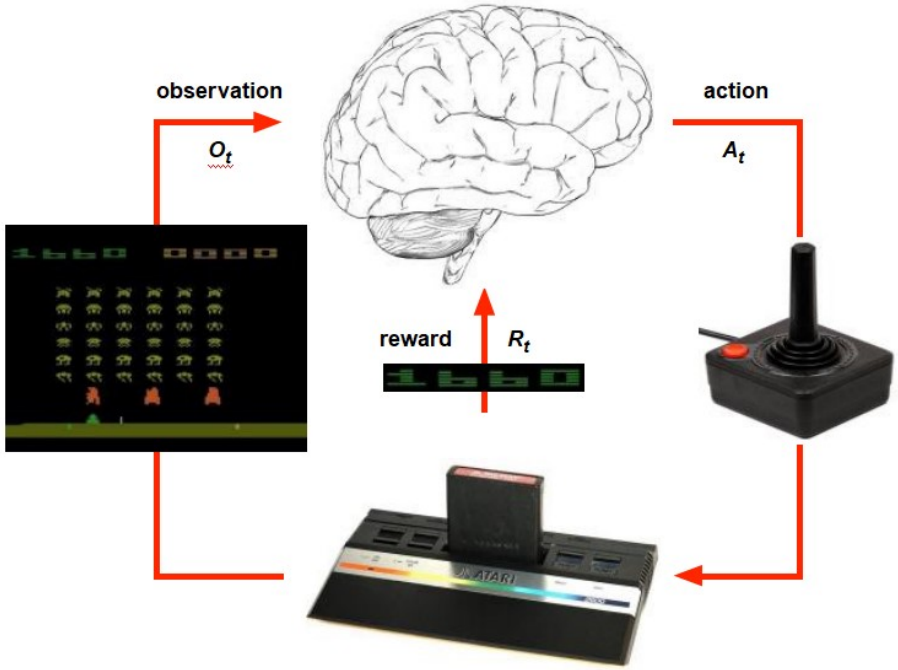
在序贯决策中两个基本的问题：学习和规划

- 学习（Reinforcement Learning）：
  - 环境最初是未知的
  - 智能体与环境进行交互
  - 智能体改善其行为策略（学习）
- 规划（Planning）：
  - 环境如何工作对于智能体是已知的
  - 智能体并不与环境发生实际的交互（利用模型计算）
  - 利用其构建的模型进行计算，在此基础上改善其行为策略

- 一个常用的强化学习问题解决思路是，先学习环境如何工作，也就是了解环境工作的方式,即学习得到一个模型，然后利用这个模型进行规划

1.5 强化学习中的一些问题

Atari游戏中的强化学习

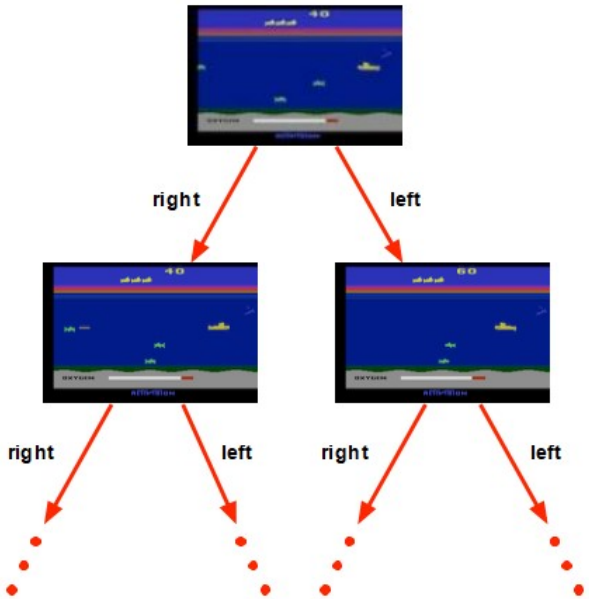


- 游戏的规则是事先不知道的
- 直接从与游戏互动中学习
- 直接从屏幕上读取信息，通过操纵杆来控制动作

1.5 强化学习中的一些问题

Atari 游戏中的规划

- 游戏的规则是事先知道的
- 可以查询模拟器
  - 智能体拥有一个完美的模型
- 如果我在状态s下选择动作a
  - 他会知道下一个状态是什么
  - 他会知道得分是多少
- 提前规划以找到最优决策
  - 右边的树型结构就是一个例子



1.5 强化学习中的一些问题

试探和开发 (1)

Exploration and Exploitation

- 强化学习就像一个不停试错的学习过程（trial-and-error）
- 智能体应该寻找一个最优的策略
- 寻找最优策略的依据是来自它对环境的经验
- 在寻找的过程中不能失去太多的收益
- 多臂赌博机（Multi-armed bandits）

1.5 强化学习中的一些问题

试探和开发 (2)

- 试探是为了发现是否还有更好的获得收益的方式
- 开发是把当前已知的获得最大化收益的方法好好利用
- 这两者对于强化学习来说同样重要，缺一不可

1.5 强化学习中的一些问题

举例

饭店选择

开发：去你最喜欢的饭店

试探：　去尝试一家新的饭店

投放网络广告

开发：选择最成功的广告

试探：　换一个广告

石油勘探

开发：　选择已知产油多的地方

试探：　换一个新的地方

玩游戏

开发：　选择你认为最好的打法

试探：　换一种打法

1.5 强化学习中的一些问题

预测和控制

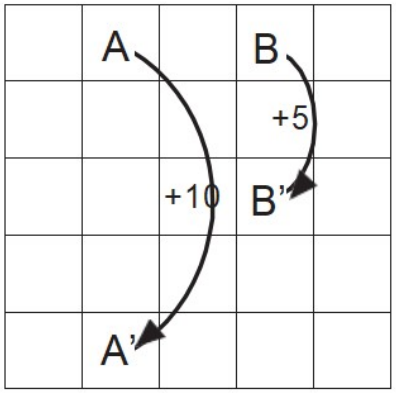
在强化学习里，我们经常需要先解决关于预测（prediction）的问题，而后在此基础上解决关于控制（Control）的问题。

预测：给定一个策略，对未来进行评估（策略评估）

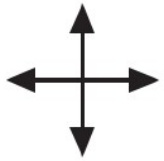
控制：找到最优的策略，对未来进行优化（策略迭代）

1.5 强化学习中的一些问题

用Gridworld的例子来说明预测



(a)



Actions

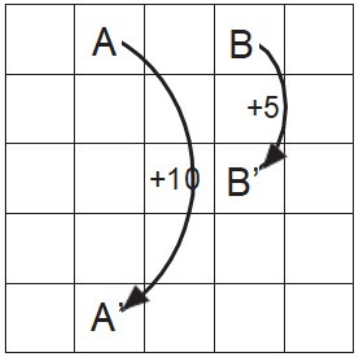
3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

(b)

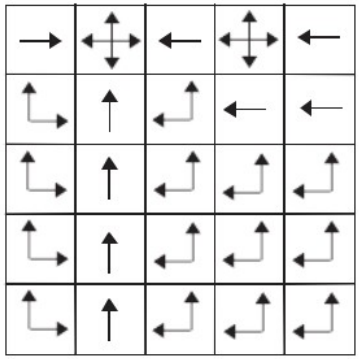
现在给出了从A到A'的奖励以及从B到B'的奖励，在"随机选择4个方向进行移动"的策略下，如何得知每一个位置的价值

1.5 强化学习中的一些问题

用Gridworld的例子来说明控制



22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7



a) gridworld　　b) V　　c) π

同样的条件，在所有可能的策略下最优的价值函数是什么？最优策略是什么？

1.5 强化学习中的一些问题

课程回顾

第一部分：初级强化学习

1介绍强化学习

2马尔可夫决策过程

3动态规划法

4无模型预测

5无模型控制

第二部分：强化学习实践

1价值函数逼近

2策略梯度法

3整合学习和规划

4试探和开发

5案例研究-游戏AI

The End