

# 多智能体系统与强化学习

主讲人：高阳、杨林、杨天培

<https://reinforcement-learning-2025.github.io/>

# 课程信息

南京大学智能科学与技术学院

# 任课老师



**高阳，教授、博导，**  
中国人工智能学会会  
士、中国计算机学会  
会士。主要研究强化  
学习、多智能体系统、  
大模型、具身智能。



**杨林，助理教授，  
特聘研究员，博导。**  
主要研究计算机系  
统建模与优化。



**杨天培，助理教授，  
特聘研究员，博导。**  
主要研究AI Agent，强  
化学习理论和应用。

# 参考教材

## Reinforcement Learning, second edition: An Introduction (可参考中文版)

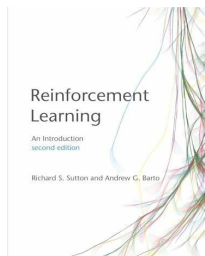
– Richard S. Sutton, Andrew G. Barto

<https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf>

学者信息

Richard S. Sutton: 阿尔伯塔大学计算科学系教授、强化学习和人工智能实验室首席研究员

Andrew G. Barto: 马萨诸塞大学阿默斯特分校信息与计算机科学学院教授



## Multi-Agent Reinforcement Learning: Foundations and Modern Approaches

– Stefano V. Albrecht, Filippos, Christianos, Lukas Schäfer

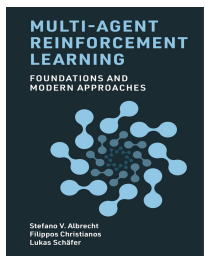
<https://www.marl-book.com/download/marl-book.pdf>

学者信息

Stefano V. Albrecht: 爱丁堡大学信息学学院副教授

Filippos, Christianos: 爱丁堡大学和自主代理研究小组的多代理强化学习博士

Lukas Schäfer: 微软研究院, 人工智能研究员



## Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations

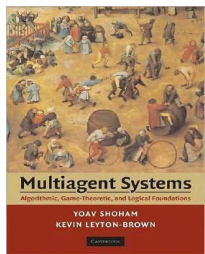
– Yoav Shoham, Kevin Leyton-Brown

<https://www.masfoundations.org/mas.pdf>

学者信息

Yoav Shoham: 斯坦福大学计算机科学系, 人工智能教授

Kevin Leyton-Brown: 不列颠哥伦比亚大学计算机科学学院教授, 加拿大皇家学会院士



# 课程内容

## 部分一：强化学习基础（4个主题）

- 基础理论与基本概念、蒙特卡洛与时序差分学习、函数逼近、策略梯度与深度强化学习、在线与离线强化学习

## 部分二：多智能体系统基础（3个主题）

- 多智能体系统简介、博弈论与纳什均衡、经典多智能体算法

## 部分三：多智能体强化学习（3~4个主题）

- 中心化训练与分散执行、合作多智能体算法、竞争与混合场景算法

## 部分四：高级主题与应用（2个主题）

- 大语言模型中的强化学习、多智能体强化学习前沿(邀请报告)

**每个主题大约2-4小时的课堂授课**

# 课程评估

## 课程考评方式

- 考勤：10分。每学期2次点名。
- 课程作业：20分。10次课后的小实验练习。
- 课程实践：20分。共2题(全选)，小组大作业，每组3人。
- 课程项目：50分。4题中选1题，每组5人。评价标准为系统和论文、专利等。

提交网址：采用课程自主代码训练平台(DODO)，  
实现代码提交、验证。

# 课程资源

## 上交张伟楠

课程网页: <https://wnzhang.net/teaching/sjtu-rl-2024/index.html>

课件下载: <https://wnzhang.net/teaching/sjtu-rl-2024/slides/12-13-marl.pdf>

视频网站: [https://www.bilibili.com/video/BV1Qopze2E1M/?spm\\_id\\_from=333.337.search-card.all.click&vd\\_source=de2e9219f23447529ee6fd993443a984](https://www.bilibili.com/video/BV1Qopze2E1M/?spm_id_from=333.337.search-card.all.click&vd_source=de2e9219f23447529ee6fd993443a984)

## 西湖大学WindyLab

课件下载: <https://github.com/MathFoundationRL/Book-Mathematical-Foundation-of-Reinforcement-Learning>

视频网站: [https://space.bilibili.com/2044042934?spm\\_id\\_from=333.1369.opus.module\\_author\\_name.click](https://space.bilibili.com/2044042934?spm_id_from=333.1369.opus.module_author_name.click)

## 斯坦福大学

视频网站: <https://web.stanford.edu/class/cs234/>

## 王树森(Meta)

视频网站: <https://www.youtube.com/watch?v=vmkRMvhCW5c>

# 第一讲：强化学习基础

强化学习问题和范式

高 阳



# 大 纲

起源

MDP模型

动态规划

# 大 纲

起源

MDP模型

动态规划

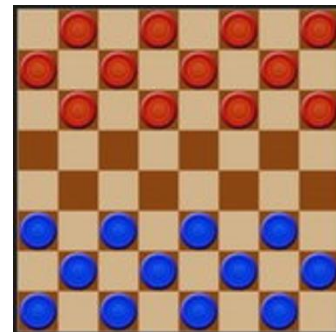
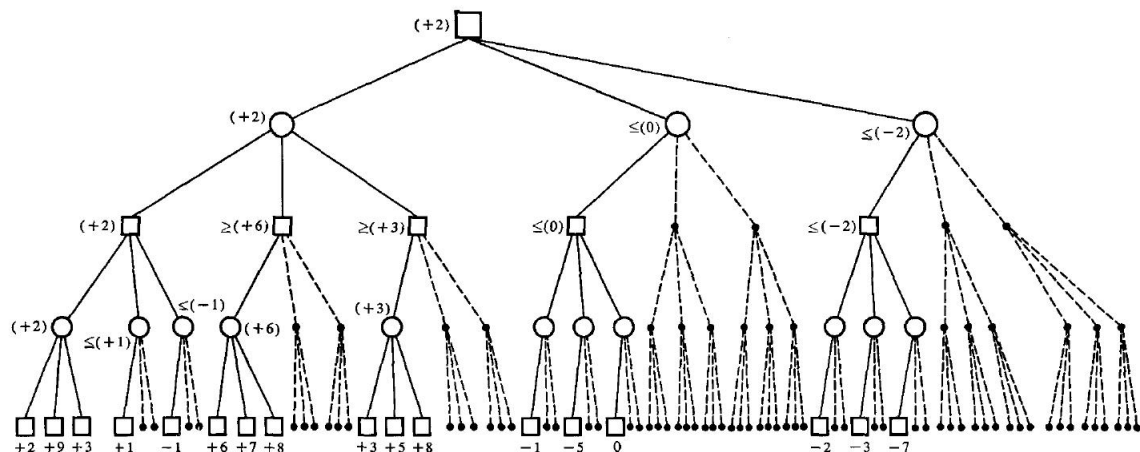
# 起源

## □ 什么是学习

- ✓ 获取新的知识、行为和认知能力
- ✓ 将所获得的知识 and 技能嵌入到系统中
- ✓ 随系统自身的运行，导致系统性能的不不断提升
- ✓ 不同于统计机器学习(概念学习)

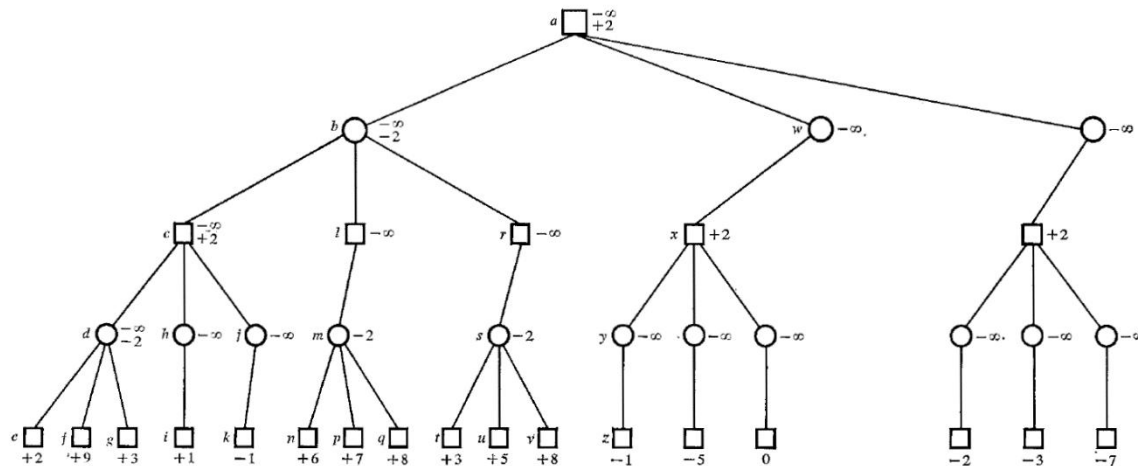


# 最早的“人机大战”

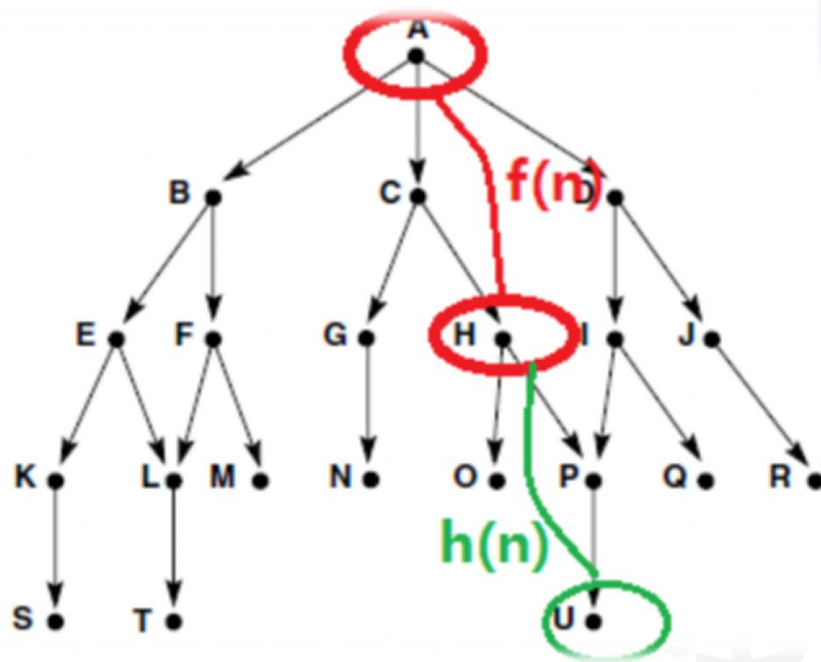


评估函数

Alpha-beta剪枝



# 启发式搜索



当存在相同启发式估值时如何选择



$$g(n)=f(n)+h(n)$$

A\*算法的产生

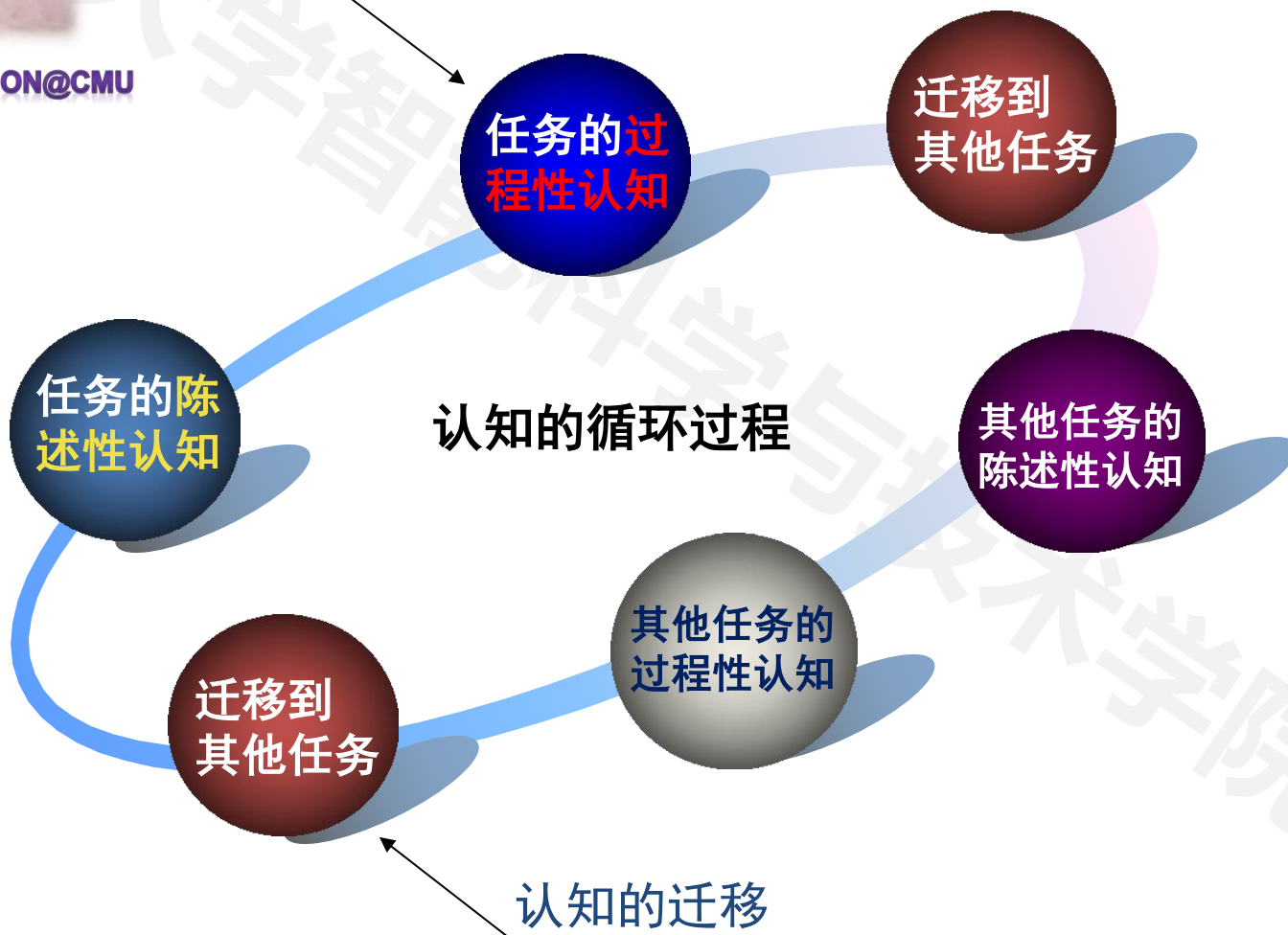
P. E. Hart, N. J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths in graphs. IEEE Trans. Syst. Sci. and Cybernetics, SSC-4(2):100-107, 1968



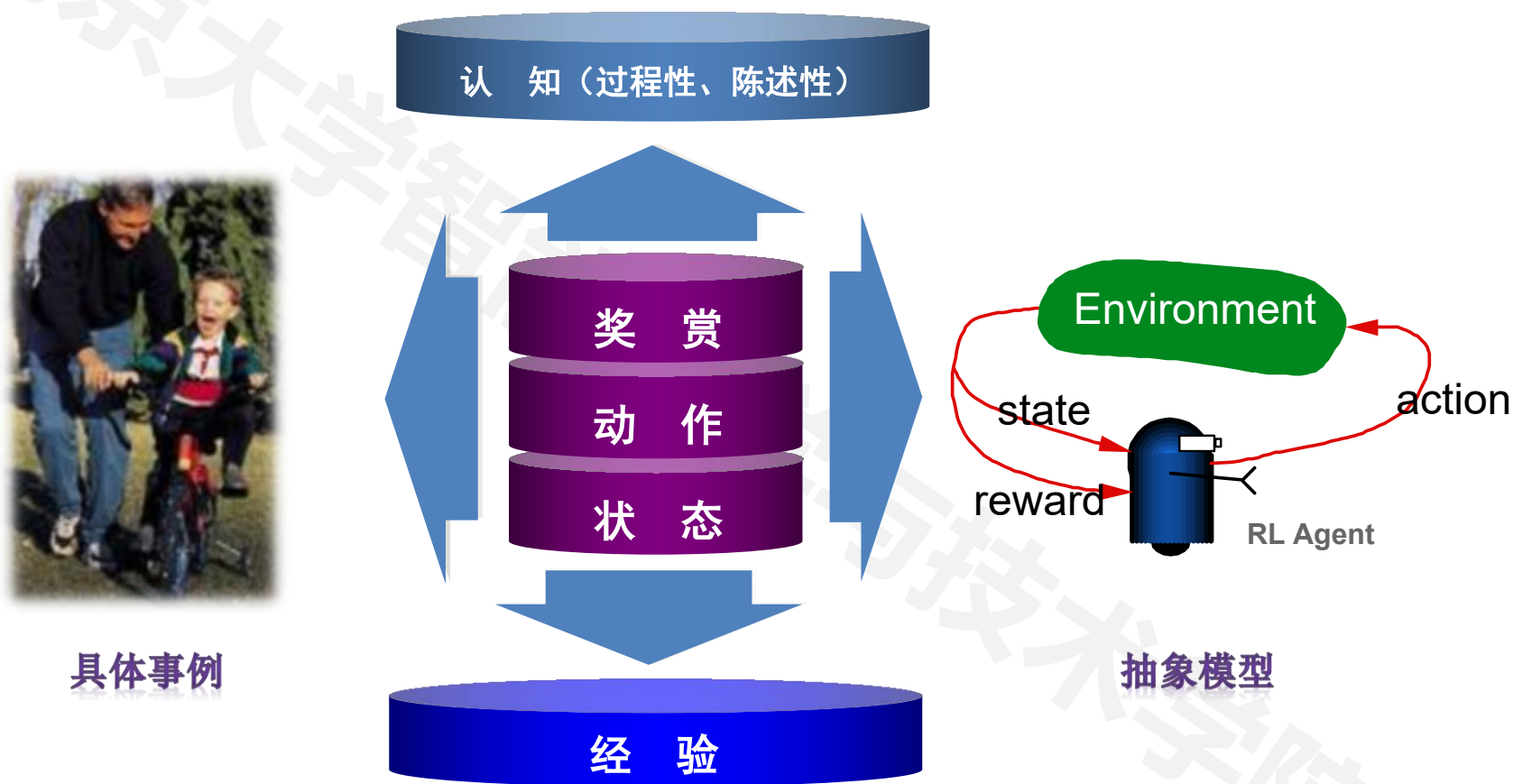
JOHN R. ANDERSON@CMU

# 从认知的角度

认知的强化



# 强化学习问题

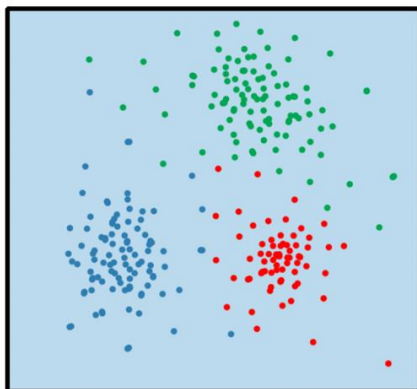


强化学习的本质：奖惩和试错(Trial and Error)

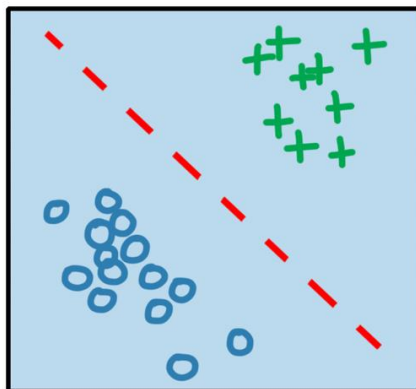
# 三类机器学习的方法

machine learning

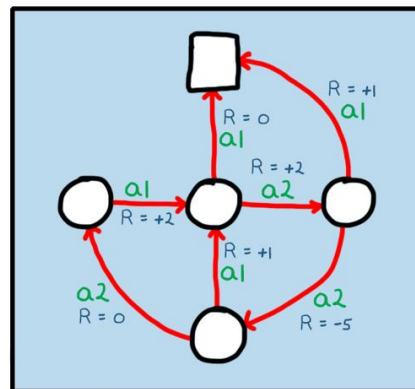
unsupervised  
learning



supervised  
learning



reinforcement  
learning





# 三类机器学习的方法

监督学习 (supervised learning)	无监督学习 (unsupervised learning)
输入数据为带标签数据	输入数据为不带标签数据
需要训练数据集进行模型训练	只需要输入数据
通常用作预测	通常用于分析
分类和回归等任务	聚类等

强化学习不依赖于静态数据集，而是在动态环境中运行，并从收集的  
经验中学习。数据点或经验是在训练过程中通过环境与智能体之间的  
试错交互收集的。

# 概念学习 VS 交互学习

## □ 概念学习

- ✓ 给定正例/反例，学习目标概念

## □ 交互学习

- ✓ 通过交互的方式学习一个任务
  - ✓ 系统(或外部环境)存在若干个“状态”
  - ✓ 学习算法/动作会影响“状态”的分布
  - ✓ 潜在的Exploration和Exploitation折衷



# 挑 战

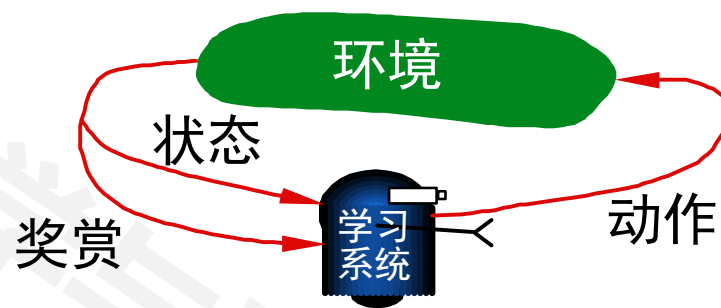
## □ 不确定性

✓ 环境、动作、反馈、模型

## □ 学习的目标

✓ 概念 → 决策

✓ 最大化长期奖赏



↓  
✓ **M**arkov **D**ecision **P**rocess

# 大 纲


起源

MDP模型

动态规划

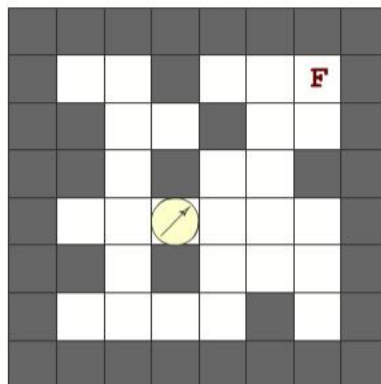
# 一个例子

Local perception

0	1	0
0		0
0	1	0

Current situation

[10001000]



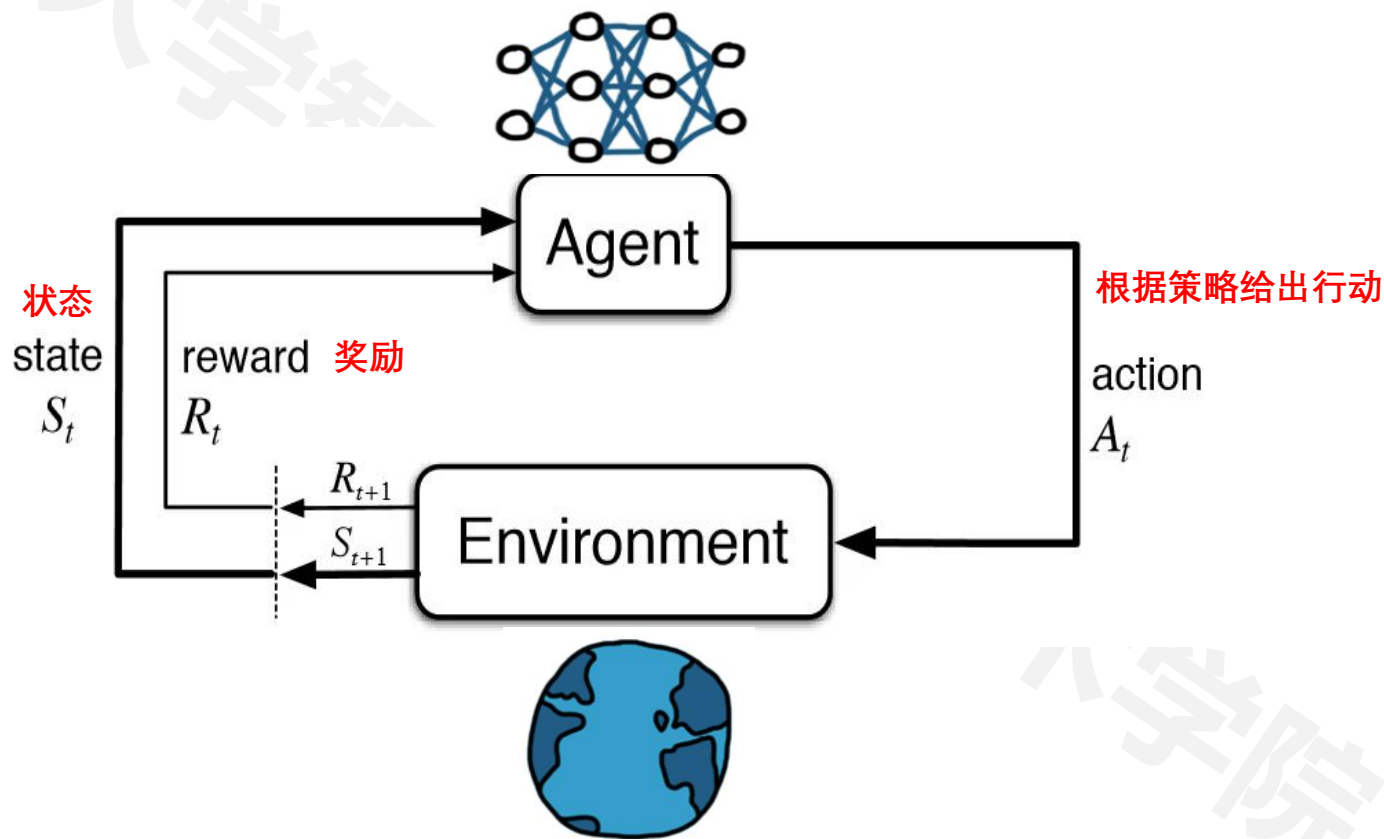
Action

10001000

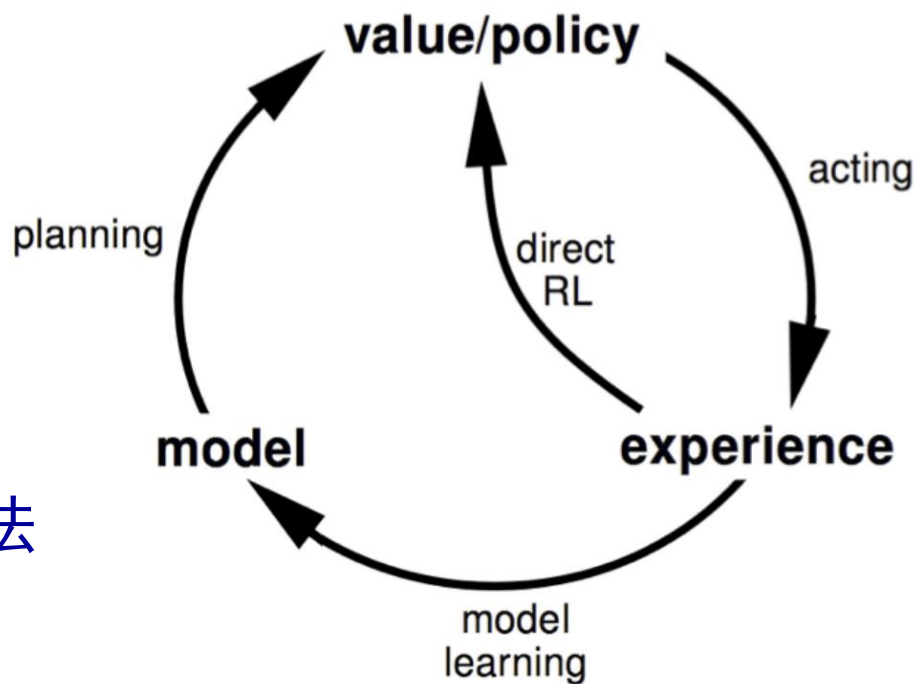
	Condition	Action	Payoff	Best classifier
Classifier list	[0010#010]	[nw]	0.7	
	[1##0100#]	[ w]	0.5	
	[000#0101]	[nw]	0.8	
	[0#000#01]	[ew]	0.5	
	[0#100001]	[ n]	0.9	
	[0010#0#0]	[sw]	0.3	
	[#1#01###]	[ w]	0.9	

# 强化学习要素

## 强化学习过程抽象



# 强化学习要素



## 基于模型的智能体决策方法

通过模型对未来的状态和收益进行  
预测和规划

# 数学模型 - MDP

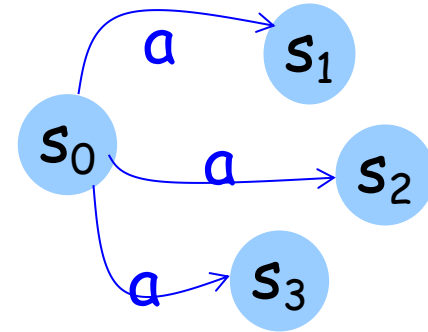
**M**arkov **D**ecision **P**rocess

S- set of **s**tates, 状态集合

A- set of **a**ctions, 动作集合

$\delta$  - transition **p**robability, 状态转移概率

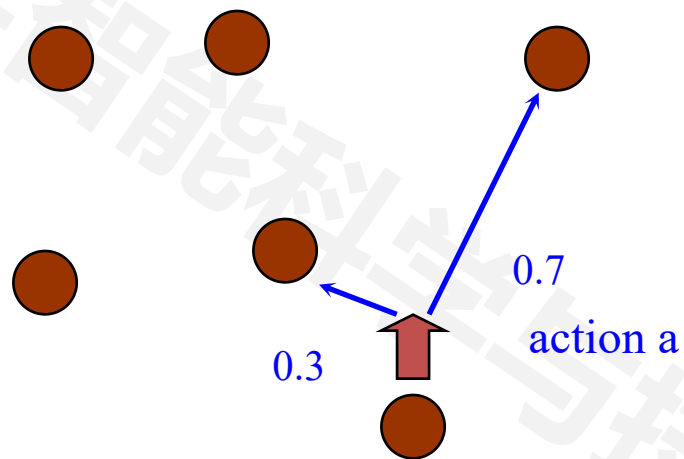
R – immediate **r**eward function, 即时奖赏函数





# MDP模型 – 状态和动作

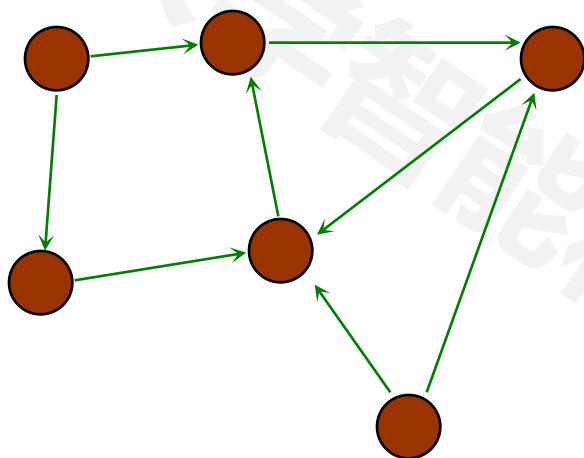
环境 = 状态集合



状态之间的转移  $\delta(s, a, s')$

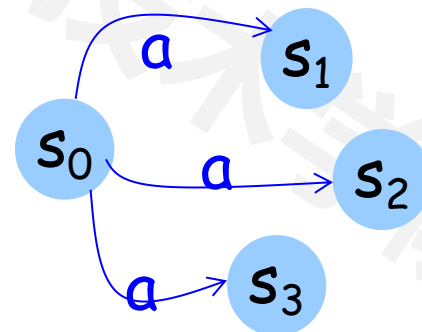
# MDP模型 – 奖赏

$R(s,a)$  = 在状态 $s$ ，采用 $a$ 动作获得的奖赏

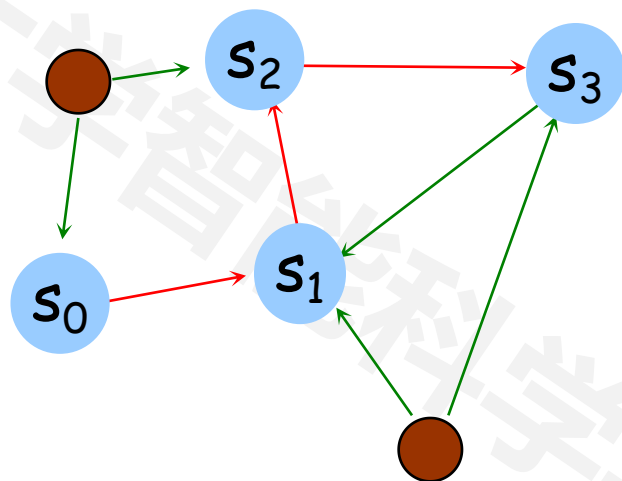


举例:

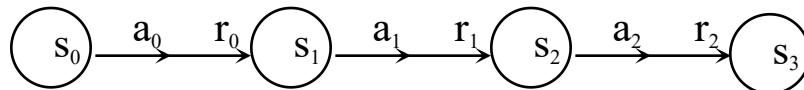
$R(s,a) =$     -1 with probability 0.5  
                  +10 with probability 0.35  
                  +20 with probability 0.15



# MDP模型 – 轨迹



在一次Episode中，所获得的经验或轨迹(trjectory)



# MDP模型 – 动作选择

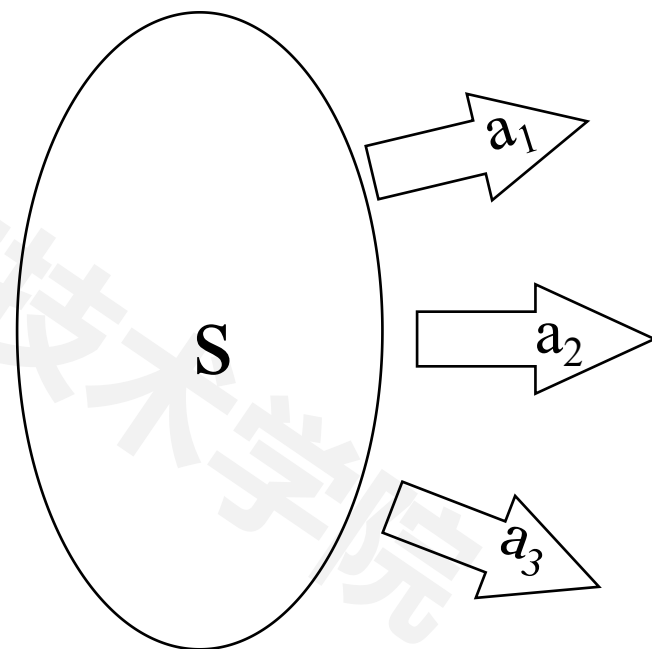
## □ 目标

- ✓ 最大化期望奖赏(单状态下)

单状态学习问题

## □ 策略

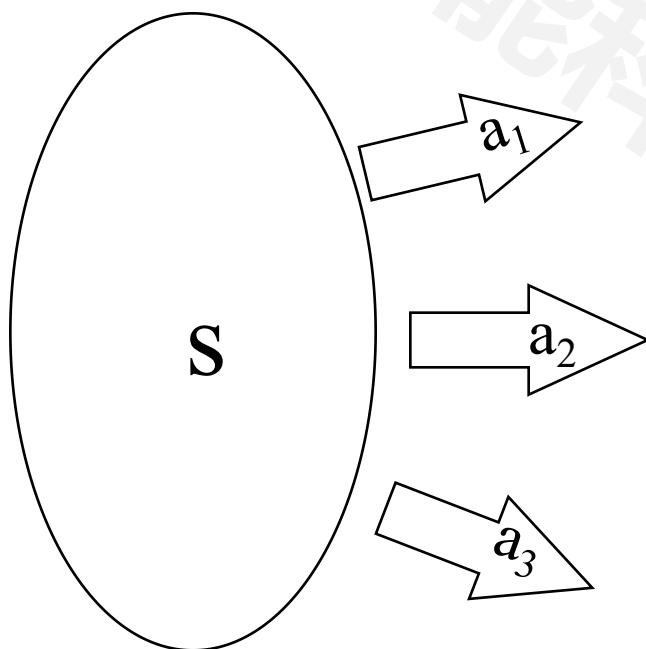
- ✓ 状态到动作的映射( $\pi: S \rightarrow A$ )



# 例：N-臂老虎机



单状态学习问题



目标：最大化期望即时奖赏

给定模型：采用贪心动作  
(Greedy action)



困难：模型未知

# MDP模型 – 返回函数

## □ 返回函数 (面向多状态学习问题)

- ✓ 将所有的即时奖赏组合成一个单一值

## □ Modeling Issues

- ✓ 轨迹中早期的奖赏和晚期的奖赏相比，谁更重要？
- ✓ 系统是持续的？还是有终止状态的？

通常返回函数是即时奖赏值的线性组合

# MDP模型 – 返回函数

## □ 有限窗口(Finite Horizon)

$$\text{return} = \sum_{1 \leq i \leq H} R(s_i, a_i)$$

## □ 无穷窗口(Infinite Horizon)

✓ 有折扣  $\text{return} = \sum_{i=0}^{\infty} \gamma^i R(s_i, a_i)$

✓ 无折扣  $\text{return} = \frac{1}{N} \sum_{i=0}^{N-1} R(s_i, a_i) \quad N \rightarrow \infty$

通常返回函数是即时奖赏值的线性组合

# MDP模型 – 动作选择

## □ 目标

- ✓ 最大化期望返回(Return)

## □ 策略

- ✓ 状态到动作的映射( $\pi: S \rightarrow A$ )

## □ 最优策略

- ✓ 如果 $\pi$ 是最优策略，则其从任一状态出发，均是最优的策略

定理：必然存在着一个确定性的最优策略



# 监督学习 VS 强化学习

## □ 监督学习

✓ (正/反例)在样本上的分布是确定的\*。

## □ 强化学习

✓ (状态/奖赏)的分布是策略依赖的(Policy Dependent!!!)

✓ 策略上小的变化都会导致返回值的巨大改变.

\*目前有迁移学习在考虑非独立同分布的学习任务

# MDP模型 – 小结

状态集合,  $|S|=n$ .  $s \in S$

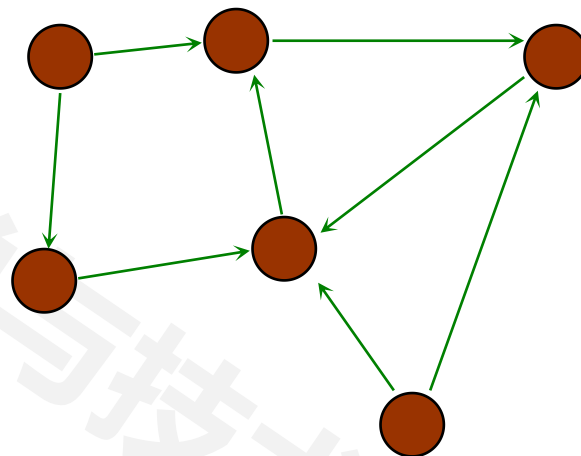
动作集合,  $|A|=k$ .  $a \in A$

转移函数  $\delta(s_1, a, s_2)$

即时奖赏函数  $R(s, a)$

策略  $\pi: S \rightarrow A$

折扣累计返回  $\sum_{i=0}^{\infty} \gamma^i r_i$



# 大 纲

起源

MDP模型

动态规划

# 动态规划

给定一个完全已知的MDP模型

## □ 策略评估(Policy Evaluation)

- ✓ 给定一个策略 $\pi$ , 评估其返回值

## □ 最优控制(Optimal Control)

- ✓ 寻找一个最优策略 $\pi^*$  (从任一状态出发, 其返回值都为最大)

# 动态规划 – 值函数

- $V^\pi(s)$ : 从s状态出发, 采用 $\pi$ 策略, 所获得的期望返回值
- $Q^\pi(s,a)$ : 从s状态出发, 采用a动作, 继而采用 $\pi$ 策略, 所获得的期望返回值
- 最优值函数 $V^*(s)$  and  $Q^*(s,a)$ : 采用最优策略 $\pi^*$ 所获得的期望返回值

定理: 策略 $\pi$  为最优策略当且仅当, 在每一个状态s

$$V^*(s) = \max_{\pi} V^\pi(s)$$

$$V^\pi(s) = \max_a Q^\pi(s,a)$$

# 动态规划 – 策略评估

□ Bellman等式(有折扣无限窗口)

$$\checkmark V^\pi(s) = E_{s' \sim \pi(s)} [R(s, \pi(s)) + \gamma V^\pi(s')]$$

□ 重写

$$\checkmark V^\pi(s) = E[R(s, \pi(s))] + \gamma \sum_{s'} \delta(s, \pi(s), s') V^\pi(s')$$

系统中所有值函数是以上公式构成的公式组，需要进行线性规划求解

# 例 - 策略评估

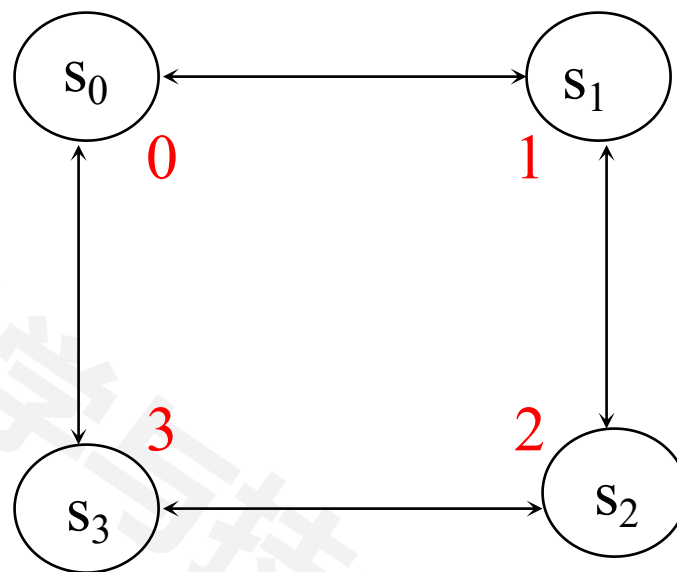
$$A = \{+1, -1\}$$

$$\gamma = 1/2$$

$$\delta(s_i, a) = s_{i+a}$$

$\pi$ : 随机 (一半概率选择+1或者-1动作)

$$\forall a: R(s_i, a) = i$$



$$V^\pi(s_0) = 0 + \gamma [\pi(s_0, +1) V^\pi(s_1) + \pi(s_0, -1) V^\pi(s_3)]$$

.....

# 例 - 策略评估

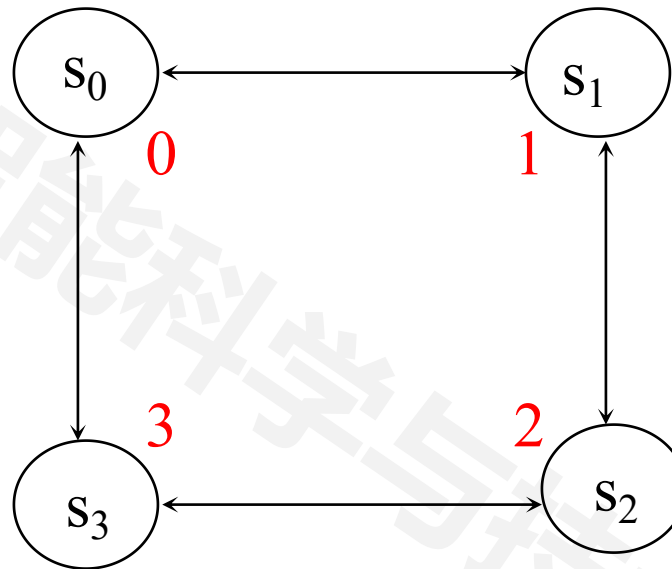
$$A = \{+1, -1\}$$

$$\gamma = 1/2$$

$$\delta(s_i, a) = s_{i+a}$$

$\pi$ : 随机

$$\forall a: R(s_i, a) = i$$



$$V^\pi(s_0) = 0 + (V^\pi(s_1) + V^\pi(s_3))/4$$

.....

$$V^\pi(s_0) = 5/3$$

$$V^\pi(s_1) = 7/3$$

$$V^\pi(s_2) = 11/3$$

$$V^\pi(s_3) = 13/3$$



# 动态规划 – 最优控制

□ Bellman等式(有折扣无限窗口)

$$\checkmark V^\pi(s) = E_{s' \sim \pi(s)} [R(s, \pi(s)) + \gamma V^\pi(s')]$$

□ 重写

$$\checkmark V^\pi(s) = E[R(s, \pi(s))] + \gamma \sum_{s'} \delta(s, \pi(s), s') V^\pi(s')$$

□ 状态-动作对值函数(对任意确定策略 $\pi$ )

$$\checkmark Q^\pi(s, a) = E[R(s, a)] + \gamma \sum_{s'} \delta(s, a, s') V^\pi(s')$$

$$\checkmark \text{其中, } V^\pi(s) = Q^\pi(s, \pi(s, a))$$

# 例 - 最优控制

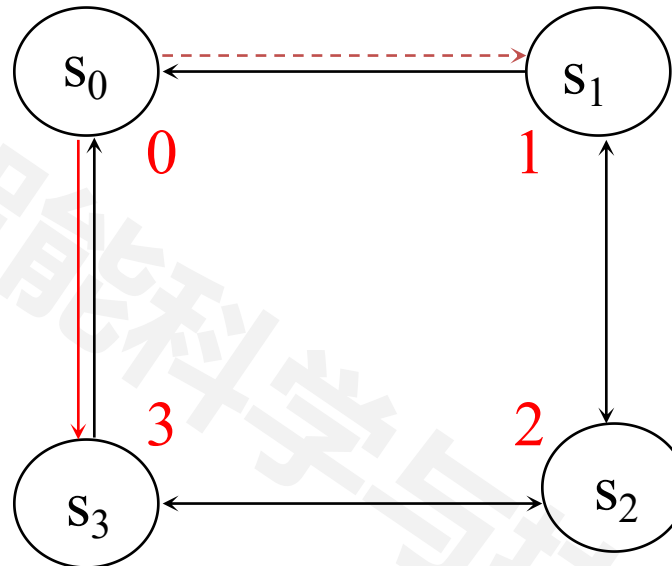
$$A = \{+1, -1\}$$

$$\gamma = 1/2$$

$$\delta(s_i, a) = s_{i+a}$$

$\pi$ : 随机

$$\forall a: R(s_i, a) = i$$



$$Q^\pi(s_0, +1) = 0 + \gamma V^\pi(s_1)$$

$$Q^\pi(s_0, +1) = 7/6$$

$$Q^\pi(s_0, -1) = 13/6$$

$$V^\pi(s_1) = 7/3$$

$$V^\pi(s_2) = 11/3$$

$$V^\pi(s_3) = 13/3$$

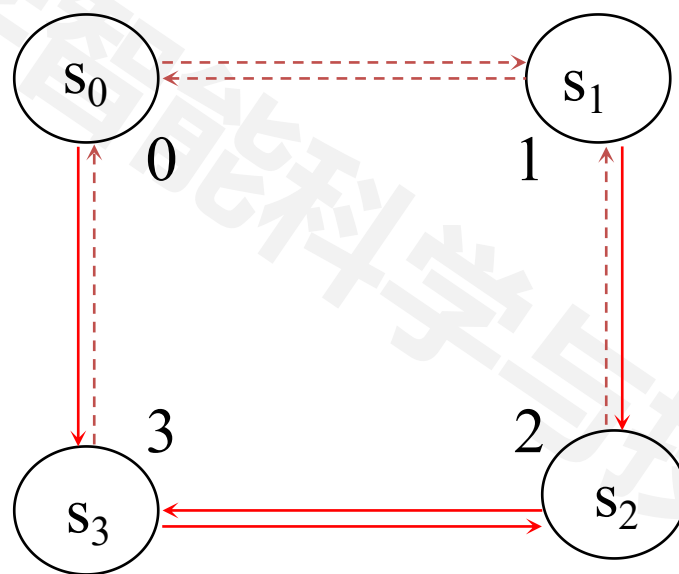
# 例 - 最优控制

$$A = \{+1, -1\}$$

$$\gamma = 1/2$$

$$\delta(s_i, a) = s_{i+a}$$

$$R(s_i, a) = i$$



$\pi$ : 根据状态-动作值函数进行修改

# 动态规划 - 最优控制

## □ 贪心策略

✓  $\pi(s) = \operatorname{argmax}_a Q^\pi(s, a)$

## □ $\epsilon$ -贪心策略

✓ 以  $1 - \epsilon$  概率选择,  $\pi(s) = \operatorname{argmax}_a Q^\pi(s, a)$

✓ 以  $\epsilon$  概率选择其他动作

# 动态规划 - 计算最优策略

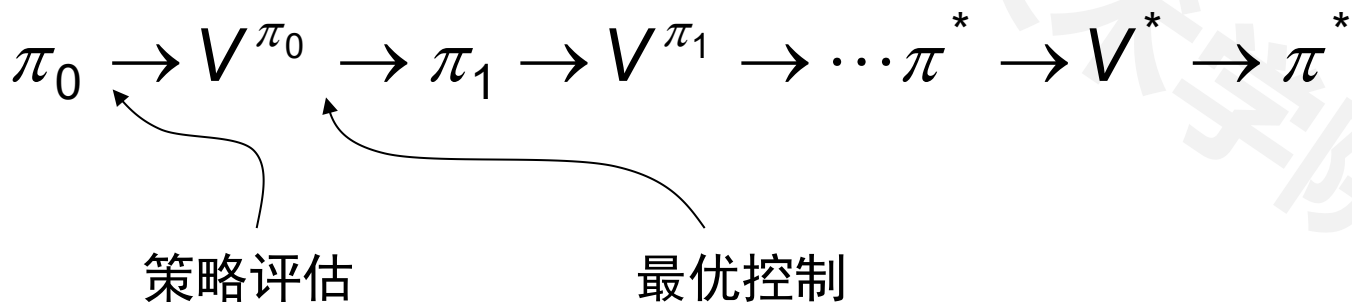
1. 线性规划

2. 策略评估

$$V^{i+1}(s) \leftarrow \max_a \{R(s, a) + \gamma \sum_{s'} \delta(s, a, s') V^i(s')\}$$

3. 最优控制

$$\pi_i(s) = \arg \max_a \{Q^{\pi_{i-1}}(s, a)\}$$



# 思考和讨论

1. 理解强化学习范式和概念学习的不同
2. 理解离散型MDP模型中各个要素
3. 掌握动态规划方法

谢谢！