一、基本概念：

二、RL，监督学习，非监督学习的异同.

三、关于马尔科夫：

① 马尔科夫性.

② 马尔科夫过程（MP）

③ 马尔科夫决策过程（MDP）

a. 策略的定义.

b. 累积回报

c. 值函数
- 状态值函数（定义、具体形式、离散形式）
- 动作值函数（定义、具体形式、离散形式）
- 最优值函数.

四、总结.

一、基本概念.

智能体.

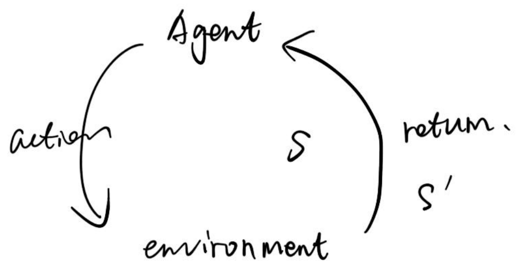
环境.

动作.（随机性）.

即时回报 R.

状态

状态转移概率 $P_{sas'}$（随机性）



二、RL，监督学习、无监督学习的异同

① 数据层面

RL：动态产生

Y | label/ 静态的

② 学习目标：

RL：累积收益最大

Y | label：minimize loss

③ 训练：

RL：链式，有相关性

Y | label：独立同分布

三、马尔科夫.

①.马尔科夫性质.(描述状态). 静态

$$P(s_{t+1} \mid s_t) = P(s_{t+1} \mid s_1, s_2, \ldots s_t)$$

仅与当前状态 s_t 有关,而与以前状态无关

②.马尔科夫过程(动态.描述状态转移)

二元组 (S, P)

$$P = \begin{pmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & & \vdots \\ P_{n1} & \cdots & P_{nn} \end{pmatrix}$$

③ 马尔科夫决策过程（动态）.
（MDP）

交互：产生动作，奖励，考虑进入序列，即为MDP

五元组（S，A，P，R，γ）

S → 状态集.
A → 动作集
P → 状态转移概率
R → 即时回报
γ → 折扣因子.

强化学习目标：

给定一个马尔科夫决策过程，寻找最优策略

$$\pi(a|s)=P(A_t=a|S_t=s)$$

⇒ 随机性策略.
　 确定性策略.

(a) 累积回报：

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

(b) 值函数定义：

R与S、A有关，S、A为随机变量。

⇒ R：随机变量 ⇒ G：随机变量，即不是一个确定值。

但期望是确定值，故可定义为描述状态 S_t 时的价值。

代入

$$\Rightarrow V_\pi(s) = E_\pi\left(\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}\right)$$

与π有关。

解释：给定s后，

只与π有关，故可用

来衡量π的好坏。

同理（评价动作）

$$\Rightarrow Q_\pi(s,a) = E_\pi\left(\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right)$$

(c) 值函数的具体形式.（Bellman）.

$$V_\pi(s) \overset{\text{定义}}{=} E_\pi(G_t \mid S_t = s)$$

$$= E_\pi(R_{t+1} + \gamma R_{t+2} + \cdots \mid S_t = s)$$

$$= E_\pi(R_{t+1} + \gamma G_{t+1} \mid S_t = s)$$

我的理解.

观测值.

$$= E_{\pi, S_{t+1} S_{t+2} \ldots}(R_{t+1} + \gamma G_{t+1} \mid S_t = s)$$

实数.

$$= E_\pi(R_{t+1} \mid S_t = s) + E_{\pi, S_{t+1} S_{t+2} \ldots}(\gamma G_{t+1} \mid S_t = s)$$

$$= E_\pi(R_{t+1} \mid S_t = s) + E_\pi(\gamma V_{t+1} \mid S_t = s)$$

$$= E_\pi(R_{t+1} + \gamma V_{t+1} \mid S_t = s)$$

同理:

$$Q_\pi(s,a) = E_\pi(R_{t+1} + \gamma Q_\pi(S_{t+1}, A_{t+1})) \mid S_t = s, A_t = a)$$

(d) 离散形式

$E(X)=\sum x\cdot p(x)$

$(V_\pi(S_t)=E_\pi(R_{t+1}+\gamma V_{t+1} \mid S_t=s))$

$$\Rightarrow v_\pi(s) = E_\pi(R_{t+1}+\gamma V_{t+1} \mid S_t=s)$$

涉及状态值概念

$$=\sum_{a\in A}\pi(a|s)(r_{t+1}+\gamma V_{t+1})$$

(从定义就能解释)

$$=\sum_{a\in A}\pi(a|s)\Big(r_{t+1}+\gamma\sum_{s'\in S}P_{sas'}v(s')\Big)$$

(下一刻状态价值不就是所有状态求和吗！)

$(Q_\pi(s,a)=E_\pi(R_{t+1}+\gamma Q(S_{t+1},A_{t+1}) \mid S_t=s,A_t=a))$

$$\Rightarrow q(s,a) = \sum_{a\in A}\pi(a|s)\Big(r_{t+1}+\gamma\sum_{s'\in S}P_{sas'}\sum_{a'\in A}\pi(a'|s')q_\pi(s',a')\Big)$$

(e) 最优值函数：取max.

Q^* V^*

四. 总结.

目的：使G最大，学习策略.

如何量化？值函数

$\Downarrow$

最优值函数

$\Downarrow$

去逼近最优值函数

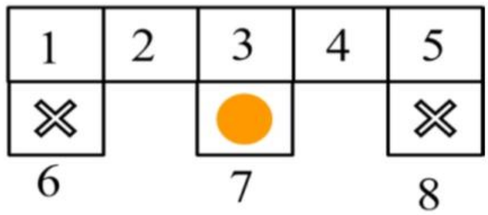
马尔科夫决策过程由元组（S，A，P，R，γ）描述，其中：

S 为有限的状态集
A 为有限的动作集
P 为状态转移概率
R 为回报函数

γ为折扣因子，用来计算累积回报。

2.3 基于 gym 的 MDP 实例讲解

以机器人找金币为例子，设定网格 6、8 是死亡区域，网格 7 是金币区域，构建其 MDP 框架(马尔科夫决策过程)。机器人的初始位置为网格世界中的任意 1 个状态，机器人从初始状态出发寻找金币，机器人每探索 1 次，进入死亡区域或找到金币，本次探索结束。



如上图，可建模为 MDP 结构：
状态空间：S={1,2,3,4,5,6,7,8}
动作空间：A={东，南，西，北}
状态转移概率：P 为机器人运动方程
回报函数：状态 7 的回报为 1，状态 6、8 的回报为-1，状态 1-5 的回报为 0

下面，我们基于 gym 构建机器人找金币的 gym 环境。

环境类的成员函数是 step（）、reset（）和 render（）

1.reset---重新初始化函数

在强化学习算法中，智能体需要一次次地尝试并累积经验，然后从经验中学到好的动作。每一次尝试我们称之为 1 条轨迹或 1 个 episode，每次尝试都要到达终止状态。一次尝试结束后，智能体需要从头开始，这就需要智能体具有重新初始化的功能。函数 reset（）就是用来做这个的。

2.render--图像引擎函数

一个仿真环境必不可少的两部分是物理引擎和图像引擎。物理引擎模拟环境中物体的运动规律；图像引擎用来显示环境中的物体图像，其实，对于强化学习算法而言，可以没有 render 函数，但是，为了便于直观显示当前环境中物体的状态，图像引擎还是有必要的。另外，加入图像引擎可以方便我们调试代码。

3.step--物理引擎函数

其输入是动作 a，输出是：下一步状态、立即回报、是否终止、调试项。它描述了智能体与

环境交互的所有信息，是环境文件中最重要的函数。在本函数中，一般利用智能体的运动学模型和动力学模型计算下一步的状态