强化学习

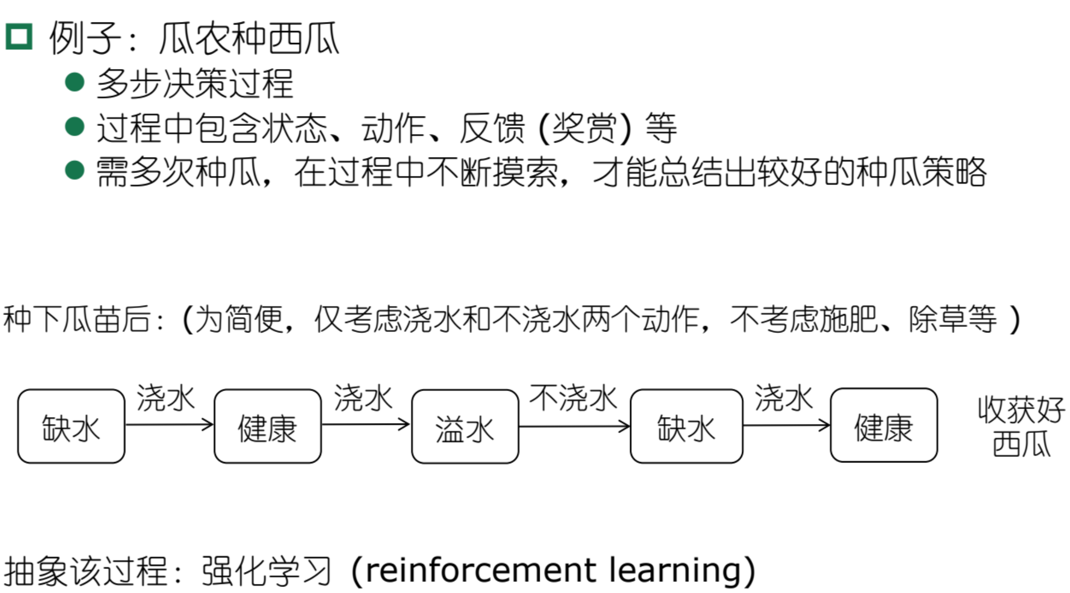
概述

强化学习又称增强学习、加强学习。是一种从环境状态到行为映射的学习，目的是使动作从环境中获得的累积回报值最大。强化学习是机器学习分支之一，介于有监督学习和无监督学习之间。

> 机器学习三大分支
>
> - 无监督学习
> - 有监督学习
> - 机器学习

强化学习问题基本设置

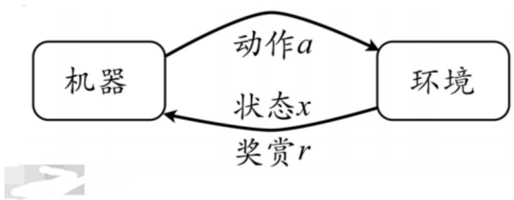
EX:瓜农种西瓜



强化学习常用马尔可夫决策过程（MDP)描述：

- 机器所处环境E
  - 例如种西瓜里，自然界为E
- 状态空间X：$x \in X$ 是继器感知到的环境的描述
  - 瓜苗长势的描述

- 机器能采取的行为空间A
  - 浇水、施肥等
- 策略$\pi: X \to A$(或者 $\pi: X \times A \to R$)
  - 根据瓜苗的状态返回对应的动作。X映射到A
- 潜在的状态转移函数$P: X \times A \times X- > R$
  - 瓜苗当前状态缺水，选择动作浇水，有一定概率恢复健康，一定概率无法恢复
- 潜在的奖赏函数（reward) R$R: X \times A \times X \to R$或者$R: X \times X \to R$
  - 瓜苗健康对应奖赏+1 ，凋零对应奖赏- 10

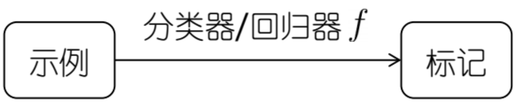
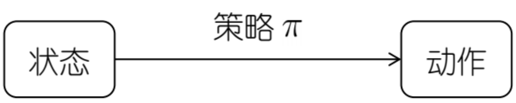
> 
>

此强化学习对应四元组：E=$< X, A, P, R >$

强化学习的目标

- 机器通过在环境中不断尝试从而学到一个策略π 使得长期执行该策略后得到的累积奖赏最大
  - T 步累积奖赏：$\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} r_t]$
    γ 折扣累积奖赏：$\mathbb{E}[\sum_{t=0}^{+\infty} \gamma^t r_{t+1}]$

强化学习和监督学习的对比：

> - 监督学习：给有标记样本
>
> 
>
>
> - 强化学习：没有有标记样本，通过执行动作之后反馈的奖赏来学习
>
> 
>
>
> 强化学习在某种意义上可以认为是具有“延迟标记信息”的监督学习

有模型学习

model-based learning : $E = < X, A, P, R >$的特点：

- X,A,P,R 均已知
- 方便起见，假设状态空间和动作空间均有限

强化学习的目标：找到时累积奖赏最大的策略π

策略评估：使用某策略所带来的累积奖赏

状态值函数：从状态x出发，使用策略π所带来的累积奖赏

$$\begin{cases} V_T^\pi(x) = \mathbb{E}_\pi\left[\frac{1}{T}\sum_{t=1}^{T} r_t | x_0 = x\right], & T\text{ 步累积奖赏；} \\ V_\gamma^\pi(x) = \mathbb{E}_\pi\left[\sum_{t=0}^{+\infty} \gamma^t r_{t+1} | x_0 = x\right], & \gamma\text{ 折扣累积奖赏.} \end{cases}$$

状态-动作值函数：从状态x出发，执行动作a后再使用策略π所带来的累积奖赏

$$\begin{cases} Q_T^\pi(x,a) = \mathbb{E}_\pi\left[\frac{1}{T}\sum_{t=1}^{T} r_t | x_0 = x, a_0 = a\right]; \\ Q_\gamma^\pi(x,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{+\infty} \gamma^t r_{t+1} | x_0 = x, a_0 = a\right]. \end{cases}$$

- 给定π，值函数的计算：值函数具有简单的递归形式
  - T 步累积奖赏：

$$\begin{aligned} V_T^\pi(x) &= \mathbb{E}_\pi\left[\frac{1}{T}\sum_{t=1}^{T} r_t | x_0 = x\right] \\ &= \mathbb{E}_\pi\left[\frac{1}{T} r_1 + \frac{T-1}{T}\frac{1}{T-1}\sum_{t=2}^{T} r_t | x_0 = x\right] \quad \text{（全概率公式）} \\ &= \sum_{a\in A}\pi(x,a)\sum_{x'\in X} P_{x\to x'}^a\left(\frac{1}{T}R_{x\to x'}^a + \frac{T-1}{T}\mathbb{E}_\pi\left[\frac{1}{T-1}\sum_{t=1}^{T-1} r_t | x_0 = x'\right]\right) \\ &= \sum_{a\in A}\pi(x,a)\sum_{x'\in X} P_{x\to x'}^a\left(\frac{1}{T}R_{x\to x'}^a + \frac{T-1}{T}V_{T-1}^\pi(x')\right). \quad \text{Bellman等式} \end{aligned}$$

  - γ折扣累积奖赏：

$$V_\gamma^\pi(x) = \sum_{a\in A}\pi(x,a)\sum_{x'\in X} P_{x\to x'}^a\left(R_{x\to x'}^a + \gamma V_\gamma^\pi(x')\right).$$

策略改进：将非最优策略改进为最优策略

策略迭代：求解最优策略的方法

- 随机策略作为初始策略
- 策略评估+策略改进+策略评估+策略改进....
- 直到策略收敛
-

输入: MDP 四元组 $E = \langle X, A, P, R \rangle$;
累积奖赏参数 T.

过程:

$\forall x \in X:\ V(x) = 0,\ \pi(x,a) = \frac{1}{|A(x)|}$;
loop
  for $t = 1, 2, \ldots$ **do**
    $\forall x \in X: V'(x) = \sum_{a \in A} \pi(x,a) \sum_{x' \in X} P^a_{x \to x'} \left(\frac{1}{t} R^a_{x \to x'} + \frac{t-1}{t} V(x') \right)$;
    if $t = T + 1$ **then**
      break
    else
      $V = V'$
    end if
  end for
  $\forall x \in X: \pi'(x) = \arg\max_{a \in A} Q(x,a)$;
  if $\forall x: \pi'(x) = \pi(x)$ **then**
    break
  else
    $\pi = \pi'$
  end if
end loop

输出: 最优策略 π

无模型学习

model-free learning :更加符合实际情况

- 转移概率，奖赏函数未知
- 甚至环境中的状态数目未知
- 假定状态空间有限

面临困难:

- 策略无法评估
- 无法通过值函数计算状态-动作值函数
- 机器只能从一个初始状态开始探索环境

解决方法:

- 多次采样
- 直接评估每一对状态-动作的值函数
- 探索过程中逐渐发现各个状态

- 蒙特卡罗强化学习：采样轨迹，用样本均值近似期望
  - 策略评估：蒙特卡罗法
    - 从某状态出发，执行某策略
    - 对轨迹中出现的每对状态-动作，记录其后的奖赏之和
    - 采样多条轨迹，每个状态-动作对的累积奖赏取平均
  - 策略改进：换入当前最优动作

一条轨迹：

$$< x_0, a_0, r_1, x_1, a_1, r_2, \ldots, x_{T-1}, a_{T-1}, r_T, x_T >$$

- 蒙特卡罗强化学习可能遇到的问题：轨迹的单一性

- 解决问题的办法
  - ϵ-贪心法
    - 同策略：被评估与被改进的是同一个策略
    - 异策略：被评估与被改进的是同一个策略（用重要性采样技术）

$$\pi^{\epsilon}(x) = \begin{cases} \pi(x), & \text{以概率 } 1-\epsilon; \\ A \text{ 中以均匀概率选取的动作}, & \text{以概率 } \epsilon. \end{cases}$$

- 同策略蒙特卡罗强化学习算法

输入： 环境 E;
动作空间 A;
起始状态 x_0;
策略执行步数 T.

过程:

$Q(x,a)=0,\ \mathrm{count}(x,a)=0,\ \pi(x,a)=\frac{1}{|A(x)|}$;
for $s=1,2,\ldots$ **do**
  在 E 中执行策略 π 产生轨迹
  $< x_0, a_0, r_1, x_1, a_1, r_2, \ldots, x_{T-1}, a_{T-1}, r_T, x_T >$;
  for $t=0,1,\ldots,T-1$ **do**
    $R=\frac{1}{T-t}\sum_{i=t+1}^{T} r_i$;
    $Q(x_t,a_t)=\frac{Q(x_t,a_t)\times \mathrm{count}(x_t,a_t)+R}{\mathrm{count}(x_t,a_t)+1}$;
    $\mathrm{count}(x_t,a_t)=\mathrm{count}(x_t,a_t)+1$
  end for
  对所有已见状态 x：

$$\pi(x,a)=\begin{cases} \arg\max_{a'} Q(x,a'), & \text{以概率 } 1-\epsilon; \\ \text{以均匀概率从 } A \text{ 中选取动作}, & \text{以概率 } \epsilon\ . \end{cases}$$

end for

输出： 策略 π

值函数近似

值函数不再是关于状态的“表格值函数” (tabular value function)

- 值函数近似
  - 将值函数表达为状态的线性函数

$$V_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^{\top}\boldsymbol{x}$$

状态向量

参数向量

  - 用最小二乘误差来度量学到的值函数与真实的值函数 V^{π} 之间的近似程度

$$\varepsilon_{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{x}\sim\pi}\left[\left(V^{\pi}(\boldsymbol{x}) - V_{\boldsymbol{\theta}}(\boldsymbol{x})\right)^2\right].$$

  - 用梯度下降法更新参数向量，求解优化问题

对强化学习的理解

有模型学习：

- 强化学习任务可归结为基于动态规划的寻优问题
- 与监督学习不同，这里并未涉及到泛化能力，而是为每一个状态找到最好的动作。

无模型学习：

蒙特卡罗强化学习的缺点：低效

- 求平均时以批处理式进行
- 在一个完整的采样轨迹完成后才对状态-动作值函数进行更新
- 克服缺点的办法：时序差分 (temporal difference，TD) 学习

如何处理环境中的未知因素

- 蒙特卡罗强化学习
- 时序差分学习

如何处理连续状态空间

- 值函数近似

如何提速强化学习过程

- 直接模仿学习

- 逆强化学习