蒙特卡罗强化学习

汇报人：李鑫

2021.7.8

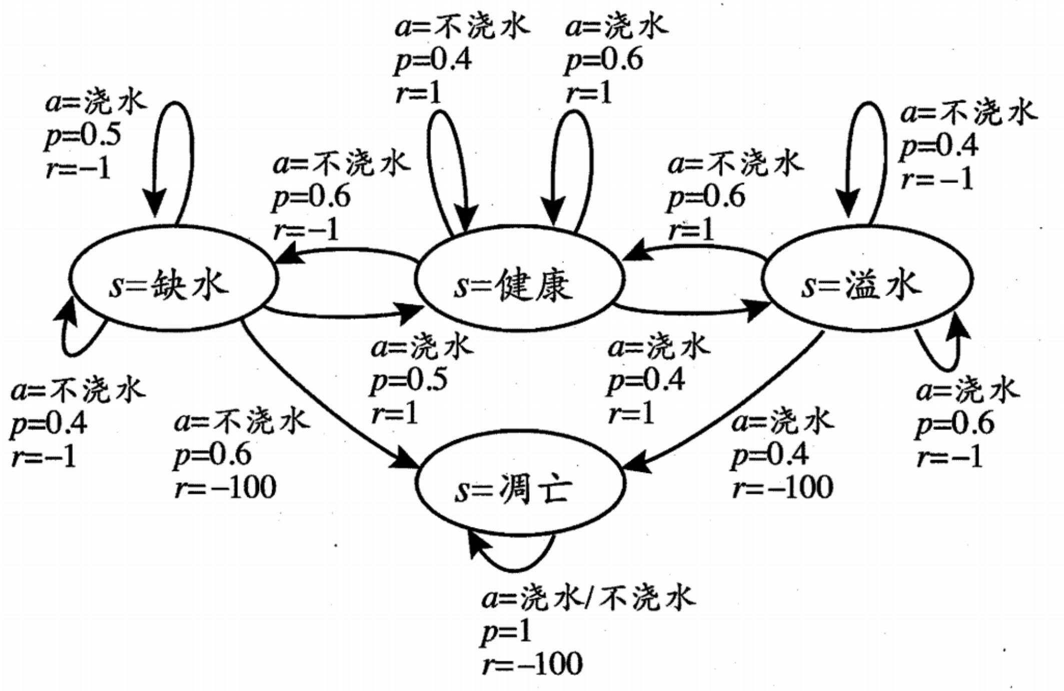
2021 WDS暑期讨论班

目录

- **免模型学习**
- 蒙特卡罗概述
- “同策略”蒙特卡罗强化学习
- “异策略”蒙特卡罗强化学习
- 总结

免模型学习

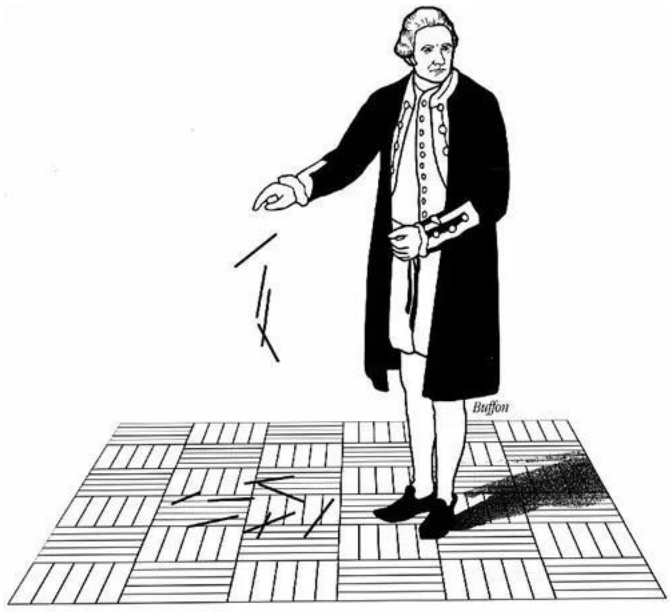
- 环境转移概率、奖赏函数很难得知
- 环境中共有多少状态也难以得知
- 学习算法不依赖于建模



目录

- 免模型学习
- **蒙特卡罗概述**
- “同策略”蒙特卡罗强化学习
- “异策略”蒙特卡罗强化学习
- 总结

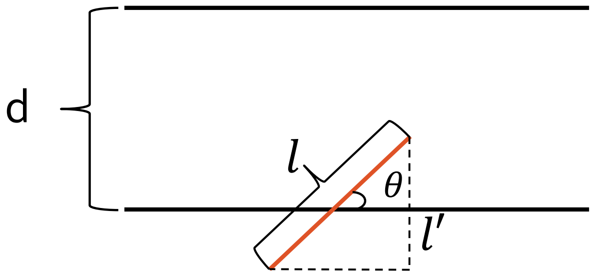
布丰(Buffon)投针



法国博物学家，作家《自然史》
1777年，Buffon投针计算Π

步骤：

1、白纸一张，画有间距为d的平行线
2、取长度为l的针随机抛掷
3、计算针与直线相交的概率P
4、圆周率$\pi=\frac{2l}{Pd}$



$[0,\pi]$按照$\Delta\theta$等分，针与线夹角为θ的概率

$$P_1=\frac{\Delta\theta}{\pi}$$

针与线相交夹角为θ的概率

$$P_2=\frac{l'}{d}=\frac{l\sin\theta}{d}$$

针与线以任意夹角相交的概率

$$P=\sum_{\theta=0}^{\pi}P_1\cdot P_2=\int_0^{\pi}\frac{l\sin\theta}{d\pi}d\theta=\frac{2l}{d\pi}$$

蒙特卡罗概述

- 基于概率统计理论、使用随机模拟的方式来解决问题的数值计算方法
- **按抽样调查法求取统计值来推定未知特性量的计算方法**
- 数学原理：大数定律

$$\frac{1}{\mathrm{n}}\sum_{i=1}^{n} X_i \xrightarrow{p} \frac{1}{\mathrm{n}}\sum_{i=1}^{n} EX_i$$

当样本容量足够大时，均值收敛于期望，频率收敛于概率

蒙特卡罗概述

- 状态x，策略π，动作a, 奖赏r,
- **策略**
  - 确定性策略：将策略表示成函数π：$X \rightarrow A$， $\mathrm{a} = \pi(x)$
  - 随机性策略：用概率表示π：$X \times A \rightarrow R, \pi(x,a)$为状态$x$下选择动作a的概率，有$\sum_a \pi(x,a) = 1$
- **值函数**
  - $V^{\pi}(x)$“状态值函数”：表示从x出发，使用策略π所带来的累计奖赏；
  - $\mathrm{Q}^{\pi}(x,a)$“状态-动作值函数”：表示从状态x出发，执行动作a后再使用策略π带来的累积奖赏。
- 目标：找到能长期使累积奖赏最大的策略

蒙特卡罗概述

- 长期累积奖赏计算方式：
  - “T步累积奖赏”： $\mathrm{E}[\frac{1}{T}\sum_{t=1}^{T} r_t]$
  - “γ折扣累积奖赏”： $E[\sum_{t=0}^{+\infty}\gamma^t r_{t+1}]$
- 蒙特卡罗强化学习：多次“采样”，然后求平均累积奖赏来作为期望累积奖赏的近似
- 由于采样必须为有限次数，因此该方法更适合于使用T步累积奖赏的强化学习任务

蒙特卡罗强化学习步骤

- 从一个起始状态出发(或起始状态集合)开始探索环境;
- 使用某种策略π进行采样，执行T步并获得轨迹

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

- 对轨迹中的每一对状态-动作对，记录其后的累积奖赏，作为该状态-动作对的一次累积奖赏采样值
- 多次采样得到多条轨迹，将每个状态-动作对的累计奖赏进行平均，即得到状态-动作值函数估计

目录

- 免模型学习
- 蒙特卡罗概述
- **“同策略”蒙特卡罗强化学习**
- “异策略”蒙特卡罗强化学习
- 总结

“同策略”蒙特卡罗强化学习

- 欲较好地获得值函数的估计，需要多条不同的轨迹。因此不可以采用确定性策略。
- 原始策略：最大化状态-动作值函数：

$$\pi = argmax_a Q(x,a)$$

- 需要使用ϵ-贪心法，以ϵ的概率从所有动作中随机选择一个，以$1-\epsilon$的概率选择当前最优动作。

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

“同策略”蒙特卡罗强化学习

- ϵ-贪心算法中，当前最优动作被选中的概率为$1-\epsilon+\frac{\epsilon}{|A|}$
- 非最优动作被选中的概率为$\frac{\epsilon}{|\mathrm{A}|}$
- 与策略迭代法类似，蒙特卡罗方法进行策略评估后，仍要进行策略改进
- “被评估”与“被改进”的是同一个策略，因此称为“同策略”蒙特卡罗强化学习算法

“同策略”蒙特卡罗强化学习

对每一个状态-动作对，计算轨迹中的累积奖赏，更新平均奖赏

每条轨迹更新一次策略

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

“同策略”蒙特卡罗强化学习

■缺陷：

- “同策略”蒙特卡罗强化学习算法最终产生的是ϵ-贪心策略
- 而引入ϵ-贪心策略只是为了方便采样，在使用时并不需要ϵ-贪心策略
- 我们需要改进的是原始策略π，该怎么做呢？

目录

“异策略”蒙特卡罗强化学习

- 函数f在概率分布p下的期望可表达为

$$E[f]=\int_{x} p(x)f(x)dx,$$

- 采样概率分布p得到$\{x_1,x_2,\ldots,x_m\}$来评估f的期望

$$\hat{\mathrm{E}}[f]=\frac{1}{m}\sum_{i=1}^{m} f(x_i)$$

“异策略”蒙特卡罗强化学习

- 引入另一个分布q，则函数f在分布p下的期望可以等价于

$$E[f] = \int_x q(x)\frac{p(x)}{q(x)}f(x)dx$$

- 可以看作$\frac{p(x)}{q(x)}f(x)$在分布q下的期望，在q上采样得到$\{x'_1, x'_2, \ldots, x'_m\}$

$$\hat{\mathrm{E}}[f] = \frac{1}{m}\sum_{i=1}^{m}\frac{p(x'_i)}{q(x'_i)}f(x'_i)$$

“异策略”蒙特卡罗强化学习

- 使用策略π的采样轨迹来评估策略π，就是对累积奖赏估计期望

$$Q(x,a)=\frac{1}{m}\sum_{i=1}^{m}R_i$$

R_i表示第i条轨迹上，自状态x至结束的累积奖赏。

- 若改用策略π'的采样轨迹来评价策略π，则需对累积奖赏进行加权

$$Q(x,a)=\frac{1}{m}\sum_{i=1}^{m}\frac{P_i^{\pi}}{P_i^{\pi'}}R_i$$

P_i^{π}和$P_i^{\pi'}$分别表示两个策略产生第i条轨迹的概率。

“异策略”蒙特卡罗强化学习

- 对于一条轨迹$< x_0, a_0, r_1, x_1, a_1, r_2, \dots, x_{T-1}, a_{T-1}, r_T, x_T >$，策略$\pi$产生该轨迹的概率为

$$P^\pi = \prod_{i=0}^{T-1} \pi(x_i, a_i)\, P_{x_i \to x_{i+1}}^{a_i}$$

- 消去共有的$P_{x_i \to x_{i+1}}^{a_i}$得

$$\frac{P_i^\pi}{P_i^{\pi'}} = \prod_{i=0}^{T-1} \frac{\pi(x_i, a_i)}{\pi'(x_i, a_i)}$$

- 若π为确定性策略，π'为ϵ-贪心策略，则

$$\pi(x_i, a_i) = \begin{cases} 1 & \mathrm{a}_i = \pi(x_i) \\ 0 & \mathrm{a}_i \neq \pi(x_i) \end{cases} \qquad \pi'(x_i, a_i) = \begin{cases} \dfrac{\epsilon}{|A|} & \mathrm{a}_i \neq \pi(x_i) \\ 1 - \epsilon + \dfrac{\epsilon}{|A|} & \mathrm{a}_i = \pi(x_i) \end{cases}$$

“异策略”蒙特卡罗强化学习

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：

$Q(x,a)=0,\ \text{count}(x,a)=0,\ \pi(x,a)=\frac{1}{|A(x)|}$;
for $s=1,2,\ldots$ **do**
  在 E 中执行 π 的 ϵ-贪心策略产生轨迹
  $< x_0, a_0, r_1, x_1, a_1, r_2, \ldots, x_{T-1}, a_{T-1}, r_T, x_T >$;
  $p_i = \begin{cases} 1-\epsilon+\epsilon/|A|, & a_i=\pi(x); \\ \epsilon/|A|, & a_i \neq \pi(x), \end{cases}$
  for $t=0,1,\ldots,T-1$ **do**
    $R=\frac{1}{T-t}\sum_{i=t+1}^{T}(r_i \times \prod_{j=i}^{T-1}\frac{1}{p_j})$;
    $Q(x_t,a_t)=\frac{Q(x_t,a_t)\times\text{count}(x_t,a_t)+R}{\text{count}(x_t,a_t)+1}$;
    $\text{count}(x_t,a_t)=\text{count}(x_t,a_t)+1$
  end for
  $\pi(x)=\arg\max_{a'} Q(x,a')$
end for

输出： 策略 π

重要性系数采样，每一轮迭代都可能发生改变

计算修正的累积奖赏，更新平均奖赏

根据值函数得到策略

目录

- **免模型学习**
- 蒙特卡罗概述
- “同策略”蒙特卡罗强化学习
- “异策略”蒙特卡罗强化学习
- **总结**

总结

- 蒙特卡罗强化学习
  - 多次采样产生多条轨迹
  - 对每一条轨迹状态-动作对，计算轨迹中的累积奖赏，增量更新状态-动作值函数
  - 通过最优化状态-动作值函数更新策略并重复迭代
- “同策略”蒙特卡罗：“被评估”与“被改进”的策略是同一个
- “异策略”蒙特卡罗：引进评估策略，改进原始策略

总结

- **蒙特卡罗强化学习优点：**
  - 通过考虑采样轨迹，克服了模型未知给策略造成的困难
- **缺陷：**
  - 与动态规划的策略迭代和值迭代算法相比，效率低下，没有充分利用强化学习任务的马尔可夫决策过程
  - 在求平均时是批处理式进行的，即在完成一个完整的轨迹后在对所有状态-动作对进行更新，实际上更新过程能增量式进行（时序差分学习）。

谢 谢