智能系统设计与应用

Homework 4-6 习题解答

刘旭辉

2021 年 6 月 18 日

Problem 1

什么是马尔科夫假设? 一个马尔科夫决策过程由哪些部分构成?什么是稳态MDP? 画出一个稳态MDP的决策网络表示.

Solution

1. 马尔科夫假设指当且状态s_t只依赖于前一时刻的状态s_{t-1}和执行的行动a_{t-1}.
2. 一个马尔科夫决策过程包括：状态集合$\mathcal{S}$, 行动集合$\mathcal{A}$, 转移函数$T(s'|s,a)$以及奖励函数$R(s,a)$.
3. 稳态MDP指转移函数和奖励函数不随时间变化. 稳态MDP可以表示为:

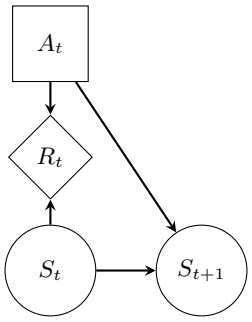

图 1: 稳态MDP的决策网络表示

Problem 2

考虑一个无限步数的MDP(如下图所示).该MDP仅需在顶部状态做决策,有left和right两个行动可供选择.每次行动后会得到确定性的奖赏.有两个确定性策略：π_{left}和π_{right} .请问,当γ分别为0,0.5和0.9时,哪个策略最优?

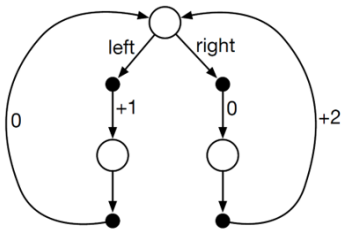

图 2: 无限步数MDP状态转移图

Solution

- $\gamma = 0$时, π_{left}为最优策略;
- $\gamma = 0.5$时, π_{left}和π_{right}均为最优策略;
- $\gamma = 0.9$时, π_{right}为最优策略;

Problem 3

用h和l表示状态high和low,用s,w和re表示search,wait和recharge.写出吸尘机器人MDP的最优状态值函

数的Bellman最优方程.

Solution

两个状态的Bellman最优值函数可分别表示为:

$$
\begin{aligned}
v_*(\mathrm{h}) &= \max\left\{ \begin{array}{l} p(\mathrm{h}|\mathrm{h},\mathrm{s})\left[r(\mathrm{h},\mathrm{s},\mathrm{h})+\gamma v_*(\mathrm{h})\right]+p(1|\mathrm{h},\mathrm{s})\left[r(\mathrm{h},\mathrm{s},1)+\gamma v_*(1)\right] \\ p(\mathrm{h}|\mathrm{h},\ \mathrm{w}\ )\left[r(\mathrm{h},\mathrm{w},\mathrm{h})+\gamma v_*(\mathrm{h})\right]+p(1|\mathrm{h},\ \mathrm{w}\ )\left[r(\mathrm{h},\ \mathrm{w}\ ,1)+\gamma v_*(1)\right] \end{array} \right\} \\
&= \max\left\{ \begin{array}{l} \alpha\left[r_{\mathrm{s}}+\gamma v_*(\mathrm{h})\right]+(1-\alpha)\left[r_{\mathrm{s}}+\gamma v_*(1)\right] \\ 1\left[r_{\mathrm{w}}+\gamma v_*(\mathrm{h})\right]+0\left[r_{\mathrm{w}}+\gamma v_*(1)\right] \end{array} \right\} \\
&= \max\left\{ \begin{array}{l} r_{\mathrm{s}}+\gamma\left[\alpha v_*(\mathrm{h})+(1-\alpha)v_*(1)\right] \\ r_{\mathrm{w}}+\gamma v_*(\mathrm{h}) \end{array} \right\} \\
v_*(l) &= \max\left\{ \begin{array}{l} \beta r_s-3(1-\beta)+\gamma\left[(1-\beta)v_*(\mathrm{h})+\beta v_*(1)\right] \\ r_w+\gamma v_*(l) \\ \gamma v_*(\mathrm{h}) \end{array} \right\}
\end{aligned}
$$

Bellman最优方程由具体的$r_s, r_w, \alpha, \beta, \gamma$的值确定.

Problem 4

考虑图3(a)中的3×3世界，每个格子中的数值表示的是R(s)，即状态s的立即奖赏，右上角含有+10的格子是终止状态(进入终止状态得到+10的奖赏后，采取任意行动都会导致情节结束). 转移模型如图3(b)所示, 它表示的含义是，以0.8的概率向选择的方向移动，各以0.1的概率向与它垂直的两个方向移动. 假设Agent的可选行动为上(U), 下(D), 左(L), 右(R),使用折扣因子为0.99的折扣奖赏定义效用(即回报). 对于下面的每种情况,计算最优策略.

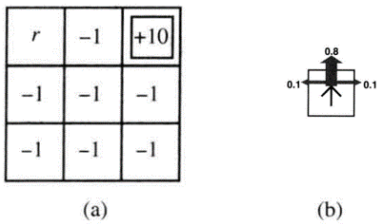

图 3: 3x3网格世界及其转移模型

1. $r = 100$
2. $r = -3$
3. $r = 0$
4. $r = 3$

Solution

u/l	l	.
u	l	d
u	l	l

$r = 100$时的最优策略

r	r	.
r	r	u
r	r	u

$r = -3$时的最优策略

r	r	.
u	u	u
u	u	u

$r = 0$时的最优策略

u/l	l	.
u	l	d
u	l	l

$r = 3$时的最优策略

Problem 5

已知Bellman最优方程：

$$U^*(s) = \max_a \left(R(s,a) + \gamma \sum_{s'} T\left(s'|s,a\right) U^*\left(s'\right) \right)$$

和值迭代的更新公式：

$$U_{k+1}(s) \leftarrow \max_a \left[R(s,a) + \gamma \sum_{s'} T\left(s'|s,a\right) U_k\left(s'\right) \right]$$

证明: 当$||U_k - U_{k-1}||_\infty < \delta, \delta = \frac{\epsilon(1-\gamma)}{\gamma}$时, $||U^* - U_k||_\infty < \epsilon$

Solution

$$\begin{aligned}
U_{k+1}(s) - U_k(s) =& \max_a \left\{ R(s,a) + \gamma \sum_{s'} T\left(s'|s,a\right) U_k(s) \right\} - \max_a \left\{ R(s,a) + \gamma \sum_{s'} T\left(s'|s,a\right) U_{k-1}(s) \right\} \\
\leq& \max_a \left\{ \gamma \sum_{s'} T\left(s'|s,a\right) \left[U_k(s) - U_{k-1}(s)\right] \right\} \\
=& \gamma \left[U_k(s) - U_{k-1}(s)\right] \max_a \left\{ \sum_{s'} T\left(s'|s,a\right) \right\} \\
=& \gamma \left[U_k(s) - U_{k-1}(s)\right]
\end{aligned}$$

所以 $||U_{k+1} - U_k||_\infty < \gamma ||U_k - U_{k-1}||_\infty < \gamma\delta$

$\|U^* - U_k\|_\infty \leq \sum_{i=0}^{\infty} \|U_{k+i-1} - U_{k+i}\|_\infty = \sum_{i=1}^{\infty} \gamma^i \|U_k - U_{k-1}\|_\infty < \sum_{i=1}^{\infty} \gamma^i \delta = \frac{\gamma\delta}{1-\gamma} = \epsilon$

Problem 6

考虑一个连续的MDP。状态由位置x和速度v构成，即$s = \begin{bmatrix} x \\ v \end{bmatrix}$。行动由加速度$a$构成，其每个时间步$\Delta t = 1$执行一次。奖赏函数如下二次奖赏：

$$R(s,a) = -x^2 - v^2 - 0.5a^2$$

即$R_s = -I$，$R_a = -[0.5]$。转移函数为：

$$\begin{bmatrix} x' \\ v' \end{bmatrix} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} + \begin{bmatrix} 0.5\Delta t^2 \\ \Delta t \end{bmatrix} [a] + \mathbf{w}$$

其中，w服从均值为0，协方差为$0.1I$的多元高斯分布。

该系统的控制目标是达到零平衡状态$s = 0$。试求出一个从$s_0 = \begin{bmatrix} -10 \\ 0 \end{bmatrix}$开始的最优5步策略。

Solution

最优5步策略为

$$\begin{aligned}
\pi_1(\mathbf{s}) &= \begin{bmatrix} 0 & 0 \end{bmatrix} \mathbf{s} \\
\pi_2(\mathbf{s}) &= \begin{bmatrix} -0.286 & -0.857 \end{bmatrix} \mathrm{s} \\
\pi_3(\mathbf{s}) &= \begin{bmatrix} -0.462 & -1.077 \end{bmatrix} \mathrm{s} \\
\pi_4(\mathbf{s}) &= \begin{bmatrix} -0.499 & -1.118 \end{bmatrix} \mathrm{s} \\
\pi_5(\mathbf{s}) &= \begin{bmatrix} -0.504 & -1.124 \end{bmatrix} \mathbf{s}
\end{aligned}$$

Problem 7

给定如下3个状态：

$$\mathbf{s}_1 = \begin{bmatrix} 4 \\ 5 \end{bmatrix}, \quad \mathbf{s}_2 = \begin{bmatrix} 2 \\ 6 \end{bmatrix}, \quad \mathbf{s}_3 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

以及它们对应的值：

$$\mathcal{U}(\mathbf{s}_1) = 2, \quad \mathcal{U}(\mathbf{s}_2) = 10, \quad \mathcal{U}(\mathbf{s}_3) = 30$$

分别使用L_1, L_2, L_∞距离度量计算状态$s = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$的2-最近邻局部近似值。

Solution

- L_1
  距离分别为6，5，5，局部近似值20；
- L_2
  距离分别为$\sqrt{18}$，$\sqrt{17}$，$\sqrt{13}$，局部近似值20；
- L_∞
  距离分别为3，4，3，局部近似值16；

Problem 8

考虑如下4个状态：

$$\mathbf{s}_1 = \begin{bmatrix} 0 \\ 5 \end{bmatrix}, \quad \mathbf{s}_2 = \begin{bmatrix} 0 \\ 25 \end{bmatrix}, \quad \mathbf{s}_3 = \begin{bmatrix} 1 \\ 5 \end{bmatrix}, \quad \mathbf{s}_4 = \begin{bmatrix} 1 \\ 25 \end{bmatrix}$$

以及采样状态$s = \begin{bmatrix} 0.7 \\ 10 \end{bmatrix}$，写出$\mathcal{U}(s)$的双线性插值方程。

Solution

$$\begin{aligned}
U_\theta(s) &= \tfrac{(x_2-x)(y_2-y)}{(x_2-x_1)(y_2-y_1)}\theta_1 + \tfrac{(x_2-x)(y-y_1)}{(x_2-x_1)(y_2-y_1)}\theta_2 + \tfrac{(x-x_1)(y_2-y)}{(x_2-x_1)(y_2-y_1)}\theta_3 + \tfrac{(x-x_1)(y-y_1)}{(x_2-x_1)(y_2-y_1)}\theta_4 \\
U_\theta(s) &= \tfrac{(1-0.7)(25-10)}{(1-0)(25-5)}\theta_1 + \tfrac{(1-0.7)(10-5)}{(1-0)(25-5)}\theta_2 + \tfrac{(0.7-0)(25-10)}{(1-0)(25-5)}\theta_3 + \tfrac{(0.7-0)(10-5)}{(1-0)(25-5)}\theta_4 \\
U_\theta(s) &= \tfrac{9}{40}\theta_1 + \tfrac{3}{40}\theta_2 + \tfrac{21}{40}\theta_3 + \tfrac{7}{40}\theta_4
\end{aligned}$$

Problem 9

试比较动态规划, 近似动态规划和在线规划, 说明每类规划方法在何种情况下更有优势.

Solution

动态规划：适合状态空间和行动空间比较小的问题, 可以保证收敛到最优解.

近似动态规划：适合状态空间和行动空间比较大(包括连续空间)的问题, 这种情况下动态规划一般是不可解的, 近似动态规划找到的近似最优策略一般可以满足要求.

在线规划：适合状态空间和行动空间非常大(包括连续空间)的问题. 这种情况下在全空间下找到近似最优策略一般是不可解的, 在线规划对当前状态找到一个近似最优策略,可以大幅减少计算量, 但每次见到新状态时需要重新计算.

Problem 10

在稀疏采样方法中，如果令$n = |\mathcal{S}|$，那么它与前向搜索方法是等价的吗？为什么？

Solution 不等价。尽管计算复杂度相同$O(|\mathcal{S}|^d|\mathcal{A}|^d)$，前向搜索在状态空间的所有状态上进行分支，

而稀疏采样在$|\mathcal{S}|$随机采样状态上进行分支。

Problem 11

给定一个MDP问题，其中$|\mathcal{S}| = 10$，$|\mathcal{A}| = 3$, $T(s'|s,a) = \frac{1}{|\mathcal{S}|}$是对所有$s, a$都成立的均匀转移分布。问：用样本数$n = |\mathcal{S}|$和深度$d = 1$的稀疏采样方法产生与深度$d = 1$的前向搜索方法完全相同的搜索树的概率是多少？

Solution 对于一个动作下的子树，相同的概率为$\frac{10!}{10^{10}}$

三个动作下全相同的概率$(\frac{10!}{10^{10}})^3$

Problem 12

1. 为什么探索与利用的概念在强化学习中如此重要?

2. 什么是n-摇臂赌博机问题?

3. 假想你有一个2-摇臂赌博机,已经知道拉其中一个摇臂会有0.9的概率输出\$1, 但你还没有拉过另外一个摇臂, 不能确定拉它是否有回报. 试说明这个问题中的探索与利用.

Solution

1. 探索指寻找新的信息的行为, 利用指基于当前信息最大化收益的行为. 如果只进行利用, 可能会错过收益更大的解, 如果只进行探索, 可能不会获得任何收益, 所以必须平衡探索与利用.

2. n-摇臂赌博机指一个具有n个摇臂的机器,摇每个摇臂都能会从一个固定的分布(一般为未知分布)中获得收益,需要学习一个最大化期望收益的摇摇臂策略.

3. 探索：摇另一个摇臂, 利用：继续摇当前摇臂. 如果未知摇臂以0.95的概率输出\$1, 那么不经过探索就不会知道这个更优解, 但它也可能以更低的概率输出1, 那么继续摇当前摇臂可以获得最大收益.

Problem 13

假设我们有一个2-摇臂赌博机. 我们估计第1个摇臂的回报为0.7, 第2个摇臂的回报为0.6, 即$\rho_1 = 0.7, \rho_2 = 0.6$. θ_1和θ_2的95%置信区间为$(0.6, 0.8), (0.3, 0.9)$. 问：θ_i和ρ_i之间的不同是什么? 假设你使用了一个ε贪心策略, 其中$\varepsilon = 0.5$. 你如何决定拉哪个摇臂? 假设你使用了一个95%置信区间的区间探索策略, 你会拉哪个摇臂?

Solution

1. ρ是对θ的估计.

2. 以50%概率随机选一个摇臂,以50%概率选摇臂1.

3. 选摇臂2.

Problem 14

试证明：如果学习率是常数α,有：

$$\hat{x}_n = (1-\alpha)^n \hat{x}_0 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} x_i$$

Solution

$$\begin{aligned}
\hat{x}_n &= \hat{x}_{n-1} + \alpha\left(x_{n-1} - \hat{x}_{n-1}\right) \\
&= \alpha x_{n-1} + (1-\alpha)\hat{x}_{n-1} \\
&= \alpha x_{n-1} + (1-\alpha)\left[\alpha x_{n-2} + (1-\alpha)\hat{x}_{n-2}\right] \\
&= \alpha x_{n-1} + (1-\alpha)\alpha x_{n-2} + (1-\alpha)^2 \hat{x}_{n-2} \\
&= \alpha x_{n-1} + (1-\alpha)\alpha x_{n-2} + (1-\alpha)^2 \alpha x_{n-3} + \\
&\quad \cdots + (1-\alpha)^{n-1}\alpha x_0 + (1-\alpha)^n \hat{x}_0 \\
&= (1-\alpha)^n \hat{x}_0 + \sum_{i=0}^{n} \alpha(1-\alpha)^{n-i} x_i
\end{aligned}$$

Problem 15

写出增量估计方程,指出其中的学习率.假想你有某一随机变量x的一个估计,想象这个估计是$\hat{x} = 3$.如果学习率为0.1, 在观察到一个新样本$x = 7$后, 你的估计会发生什么变化? 如果学习率为0.5呢? 试说明学习率在增量估计中所起的效果.

Solution

1. $\hat{x} = \hat{x} + \alpha(x - \hat{x})$.

2. 如果学习率为0.1, 新的估计值为0.34.

3. 如果学习率为0.5, 新的估计值为0.5.

4.学习率决定了新样本对当前估计的影响程度.

Problem 16

使用Q值,贝尔曼方程和增量估计方程来推导Q学习和Sarsa的更新方程.

Solution

$\hat{Q}(s,a) = \hat{Q}(s,a) + \alpha\left(Q(s,a) - \hat{Q}(s,a)\right)$. 其中$Q(s,a) = r + \gamma \max_{a'} Q(s',a')$, 所以:

$$\hat{Q}(s,a) = \hat{Q}(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a') - \hat{Q}(s,a)\right)$$

对于Q-learning:

$$Q(s_t,a_t) = Q(s_t,a_t) + \alpha\left(r_t + \gamma \max_{a} Q(s_{t+1},a) - Q(s_t,a_t)\right)$$

Problem 16 continued on next page.

对于Sarsa:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha \left(r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right)$$

Problem 17

写出线性近似Sarsa算法的伪代码.

Solution

Algorithm 1 LinearApproximationSarsa

$t \leftarrow 0$
$s_0, a_t \leftarrow$ initial state and action
初始化 θ
loop
    观测新状态s_{t+1}和奖励r_t
    基于$\theta^{\mathrm{T}}\beta(s_{t+1}, a)$和某种探索策略选择行动$a_{t+1}$
    $\theta \leftarrow \theta + \alpha \left(r_t + \gamma \theta^{\top} \beta \left(s_{t+1}, a_{t+1} \right) - \theta^{\top} \beta \left(s_t, a_t \right) \right) \beta \left(s_t, a_t \right)$
    $t \leftarrow t + 1$
end loop

Problem 18

假设我们有下表给定的参数化策略$\pi_\theta, \pi_{\theta'}$:

	a_1	a_2	a_3	a_4
$\pi_\theta \left(a \mid s_1 \right)$	0.1	0.2	0.3	0.4
$\pi_{\theta'} \left(a \mid s_1 \right)$	0.4	0.3	0.2	0.1
$\pi_\theta \left(a \mid s_2 \right)$	0.1	0.1	0.6	0.2
$\pi_{\theta'} \left(a \mid s_2 \right)$	0.1	0.1	0.5	0.3

有5个采样出的状态，分别是s_1, s_2, s_1, s_1, s_2。试用KL散度的定义估算$\mathbb{E}_s[D_{KL}(\pi_\theta(\cdot|s)||\pi_{\theta'}(\cdot|s)]$.

Solution

根据频率估计概率

$$P(s_1) = \frac{3}{5}, \quad P(s_2) = \frac{2}{5}$$

根据KL散度定义

$$D_{KL}(\pi_\theta(s_1)||\pi_{theta'}(s_1)) \approx 0.456$$

$$D_{KL}(\pi_\theta(s_2)||\pi_{theta'}(s_2)) \approx 0.028$$

估算得到

$$\mathbb{E}_s[D_{KL}(\pi_\theta(\cdot|s)||\pi_{\theta'}(\cdot|s)] \approx 0.2848$$

Problem 19

假设我们有一个与转移和奖赏模型未知的环境交互的Agent，下表是交互产生的数据。试用极大似然估计方法从这组数据中估计状态转移函数$T(s'|s,a)$和奖赏函数$R(s,a)$。

s	a	r	s'
s_2	a_1	2	s_1
s_1	a_2	1	s_2
s_2	a_2	1	s_1
s_1	a_2	1	s_2
s_2	a_2	1	s_3
s_3	a_2	2	s_2
s_2	a_2	1	s_3
s_3	a_2	2	s_3
s_3	a_1	2	s_2
s_2	a_1	2	s_3

Solution

奖赏函数

s	a	$N(s,a)$	$\rho(s,a)$	$\hat{R}(s,a)=\frac{\rho(s,a)}{N(s,a)}$
s_1	a_1	0	0	0
s_1	a_2	2	2	1
s_2	a_1	2	4	2
s_2	a_2	3	3	1
s_3	a_1	1	2	2
s_3	a_2	2	4	2

转移函数

s	a	s'	$N(s,a,s')$	$\hat{T}(s' \mid s,a)=\frac{N(s,a,s')}{N(s,a)}$
s_1	a_1	s_1	0	1/3
s_1	a_1	s_2	0	1/3
s_1	a_1	s_3	0	1/3
s_1	a_2	s_1	0	0
s_1	a_2	s_2	2	1
s_1	a_2	s_3	0	0
s_2	a_1	s_1	1	1/2
s_2	a_1	s_2	0	0
s_2	a_1	s_3	1	1/2
s_2	a_2	s_1	1	1/3
s_2	a_2	s_2	0	0
s_2	a_2	s_3	2	2/3
s_3	a_1	s_1	0	0
s_3	a_1	s_2	1	1
s_3	a_1	s_3	0	0
s_3	a_2	s_1	0	0
s_3	a_2	s_2	1	1/2
s_3	a_2	s_3	1	1/2

Problem 19 continued on next page.

Problem 20

POMDP是什么的缩写?它与MDP有何不同?画出POMDP的决策网络结构,它与MDP的决策网络结构有何不同?

Solution

POMDP是Partially Observable Markov Decision Process的缩写. POMDP中无法直接观测状态S,而是只能观测到一个可以部分反映状态的性质的观测O. POMDP的决策网络比MDP要多一个依赖于状态的观测结点,如下图(注：观测结点也可同时依赖于状态和行动):

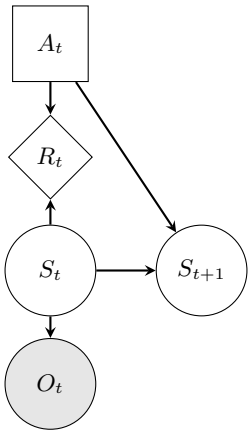

图 4: POMDP的决策网络表示

Problem 21

有如下两个栅格世界.在左侧的栅格世界中,已知Agent的位置(表示为红色方块),在右侧的栅格世界中,只有可能状态的一个概率分布.对每种情况,如何表示状态?请用这个例子来解释为什么称有些POMDP为信念状态MDP?为什么求解信念状态MDP是困难的?

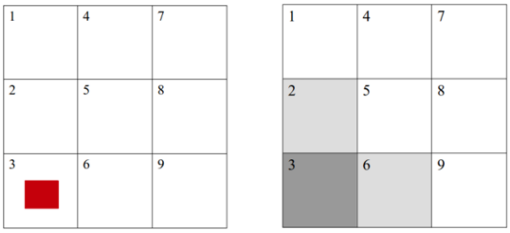

图 5: 两个栅格世界的示意图

Solution

左侧栅格世界的状态可表示直接用栅格编号表示为为$s = 3$,

右侧可用在所有各栅格上的分布概率表示为$s = (0, 0.25, 0.5, 0, 0, 0.25, 0, 0, 0)$, 这是一个关于状态的信念, 所以称有些POMDP为信念状态MDP.

右侧的栅格世界虽然也只有9个栅格, 但是信念却有无穷多种, 直接在连续的信念空间上进行值迭代是不可解的, 即使将每维信念离散化成10个值, 总的状态空间也有10^9, 比左侧栅格世界的状态空间要大得多.

Problem 22

可以用哪些方法来更新POMDP中的信念状态? 使用时怎样在这些方法中选择?

Solution

更新方法主要有离散状态滤波, 线性高斯滤波, 粒子滤波等.
离散状态滤波适合小的离散的状态空间.
线性高斯滤波适合连续的观测空间和转移模型, 但需要转移模型本身是线性高斯的.
粒子滤波适合上述两种方法不适用的情形, 利用采样的方法更新信念状态.

Problem 23

对一个离散状态的滤波器, 从信念状态更新的定义

$$b'(s') = P(s'|o, a, b)$$

出发, 推导出如下方程:

$$b'(s') = O(o|s', a) \sum_s T(s'|s, a)b(s)$$

.

Solution

$$\begin{aligned}
b'(s') \propto & P(o|s', a, b)\, P(s'|a, b) && \#\text{贝叶斯公式} \\
= & P(o|s', a)\, P(s'|a, b) && \#o\text{与}b\text{独立} \\
= & P(o|s', a) \sum_s P(s', s|a, b) \\
\propto & O(o|s', a) \sum_s P(s'|s, a, b)\, P(s|a, b) && \#\text{全概率公式} \\
= & O(o|s', a) \sum_s T(s'|s, a)\, P(s|a, b) && \#s'\text{与}b\text{独立, 此时该概率为转移概率} \\
= & O(o|s', a) \sum_s T(s'|s, a)\, b(s) && \#s\text{与}a\text{独立, 此时该概率为信念状态}
\end{aligned}$$

Problem 24

假想你已经解出了一个3个状态的POMDP问题的策略, 可以表示为如下阿尔法向量:

$$\begin{pmatrix} 300 \\ 100 \\ 0 \end{pmatrix}, \begin{pmatrix} 167 \\ 10 \\ 100 \end{pmatrix}, \begin{pmatrix} 27 \\ 50 \\ 50 \end{pmatrix}$$

Problem 24 continued on next page.

第1, 3个阿尔法向量对应的行动为1, 第2个阿尔法向量对应的行动为2. 这是一个有效的策略吗? 能每个行动有多个阿尔法向量吗? 如果该策略有效, 请确定在信念状态$\vec{b} = \begin{pmatrix} 0 \\ 0.7 \\ 0.3 \end{pmatrix}$应采取何种行动?

Solution

是个有效的策略, 每个行动可以对应多个阿尔法向量.
$\vec{\alpha}_1 \cdot \vec{b} = 70, \vec{\alpha}_2 \cdot \vec{b} = 37, \vec{\alpha}_3 \cdot \vec{b} = 50$, 因此应该选择第1个阿尔法向量对应的行动1.

Problem 25

离线求解一个POMDP与在线求解一个POMDP有何不同? 各有什么优缺点? QMDP, FIB, 基于点的值迭代有何不同?各有什么优缺点?

Solution

1.
离线求解一个POMDP意味着在执行前完成计算. 这样可以预先算好一个策略, 在执行时计算量非常小, 只需要将每个阿尔法向量与当前的信念状态相乘即能获得当前应该执行的行动.
在线求解一个POMDP意味着在执行时完成计算. 在执行时计算量比较大, 但是可以处理大规模的问题, 因为每次求解时可以利用当前状态(或信念状态)的信息,减小搜索空间.

2.
QMDP假设所有状态在执行完当前行动后会变成完全可观测的. 它的计算最简便,但不适合处理信息收集的问题, 因为一旦假设了状态会变成完全可观测的, 就不会采取任何收集信息的行动.
FIB考虑了部分可观测性, 但计算复杂.
基于点的值迭代将信念状态空间离散成一组信念点的集合, 在此基础上进行值迭代.

Problem 26

假设行动空间$\mathcal{A} = \{a^1, a^2\}$，信念状态$b = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$，奖赏总为1，观察函数为$O(o^1|a^1) = 0.8, O(o^1|a^2) = 0.4$，折扣因子$\gamma = 0.9$。我们有一个阿尔法向量$\alpha = \begin{bmatrix} -3 \\ 4 \end{bmatrix}$，并用它表示近似值函数。请使用深度为1的前向搜索方法计算$U(b)$ 。计算中可能用到的数据如下：

a	o	UPDATEBELIEF (b, a, o)
a^1	0^1	$[0.3, 0.7]^\top$
a^2	0^1	$[0.2, 0.8]^\top$
a^1	0^2	$[0.5, 0.5]^\top$
a^2	0^2	$[0.8, 0.2]^\top$

Solution

$$Q_d(\mathbf{b}, a) = R(\mathbf{b}, a) + \gamma \sum_o P(o \mid b, a) U_{d-1}(\text{Update}(\mathbf{b}, a, o))$$

Problem 26 continued on next page.

$$
\begin{aligned}
U_0\left(\text{Update}\left(\mathbf{b}, a^1, o^1\right)\right) &= \boldsymbol{\alpha}^\top \mathbf{b}' = 0.3 \times -3 + 0.7 \times 4 = 1.9 \\
U_0\left(\text{ Update }\left(\mathbf{b}, a^2, 0^1\right)\right) &= 0.2 \times -3 + 0.8 \times 4 = 2.6 \\
U_0\left(\text{ Update }\left(\mathbf{b}, a^1, o^2\right)\right) &= 0.5 \times -3 + 0.5 \times 4 = 0.5 \\
U_0\left(\text{ Update }\left(\mathbf{b}, a^2, o^2\right)\right) &= 0.8 \times -3 + 0.2 \times 4 = -1.6
\end{aligned}
$$

$$
\begin{aligned}
Q_1\left(\mathbf{b}, a^1\right) &= 1 + 0.9\left(\left(P\left(o^1 \mid \mathbf{b}, a^1\right) U_0\left(\text{ Update }\left(\mathbf{b}, a^1, o^1\right)\right) + \left(P\left(o^2 \mid \mathbf{b}, a^1\right) U_0\left(\text{Update}\left(\mathbf{b}, a^1, o^2\right)\right)\right)\right. \\
&= 1 + 0.9(0.8 \times 1.9 + 0.2 \times 0.5) = 2.458 \\
Q_1\left(\mathbf{b}, a^2\right) &= 1 + 0.9\left(\left(P\left(o^1 \mid \mathbf{b}, a^2\right) U_0\left(\text{Update}\left(\mathbf{b}, a^2, o^1\right)\right) + \left(P\left(o^2 \mid \mathbf{b}, a^2\right) U_0\left(\text{Update}\left(\mathbf{b}, a^2, o^2\right)\right)\right)\right. \\
&= 1 + 0.9(0.4 \times 2.6 + 0.6 \times -1.6) = 1.072
\end{aligned}
$$

最后我们有$U_1(\mathbf{b}) = \max_a Q_1(\mathbf{b}, a) = 2.458$。