6 ccs3ain,教程 04 (版本 1.0)

1.我在繁忙的日程中有半个小时的空闲时间，我可以在安静地在办公室工作和出去喝杯咖啡之间做出选择。

如果我呆在我的办公室里,会发生三件事:我可以完成一些工作(效用= 8),或者我可以分心看美国中期选举预测(效用= 1),或一个同事可能会停止谈论一些工作我们做修改课程(实用= 5)。

如果我出去喝咖啡，我很可能会享受到一杯平滑的咖啡化(效用= 10)，但也有可能我会把咖啡洒在自己身上(效用= - 20)。

如果我选择呆在办公室，完成工作的概率是 0.5，而分心的概率，以及一个同事路过的概率分别是 0.3 和 0.2。

如果我出去喝咖啡，我享受饮料的机会是 0.95，而打翻饮料的机会是 0.05。(a)计算呆在办公室和出去喝杯咖啡的预期效用。

(b)根据最大期望效用原则，我应该选择哪种行为?

(c)如果我使用 maximin 或 maximax 决策标准，这个决策会改变吗?

2.考虑我们在讲座中学习的简单世界(图 1)。

(a)将其写成马尔可夫决策过程的正式描述(如幻灯片所示)。

(b)假设行为是确定性的(所以 agent 以 1 的概率向它试图移动的方向移动)，写下在这种情况下适用的贝尔曼方程。

提示:把幻灯片中的 Bellman 方程简化一下，对于每个 a, P(s 0|s, a)对于一对 s = 1，对于其他所有的 s 都是 0 和 0。

(c)使用这个 Bellman 的确定性版本在世界上运行值迭代，并获得每个状态的效用值。

您将需要运行值迭代直到它稳定下来。假设 $\gamma = 1$，初始 U (s) = 0。

(d)写下给定值迭代确定性版本的解决方案的最优策略。

3.现在考虑相同的世界，但假设行为是非确定性的。对于任何操作，操作成功的概率为 0.9，操作完全失败的概率为 0.1(因此代理不移动)。

例如，如果动作是向上的，代理以 0.9 的概率向上移动，并以 0.1 的概率停留在相同的位置。

(a)使用 Bellman 的非确定性版本对世界进行值迭代，获取每个状态的效用值。

请注意，这个问题要求的是所有状态的值收敛之后的值。

提示:电子表格是简化计算的好方法。

(b)写下最优政策。这与您在 Q2 中计算的非确定性模型的最佳策略有何不同?这说明了什么?

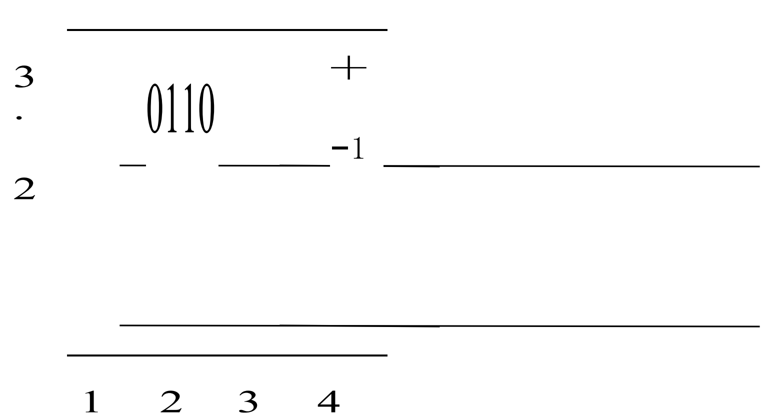


图 1:我们在讲座中学习的简单世界。

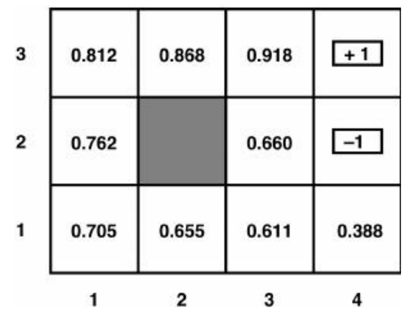


图 2:我们在课堂上学习的简单世界的效用值，其中问题有回报和课堂上的转换模型。

4.现在考虑一下这个问题的版本——回想一下，在这个版本中，过渡模型是这样的:当主体试图朝一个给定的方向移动时，有 10%的时间它会向左移动[1]， 10%的时候它会向右移动。图 2 显示了在此问题版本的最佳策略下获得的实用程序。

(a)使用图 2 中的效用值，计算状态(3,1)的期望效用。关于这个值，你注意到什么?

(b)现在计算最大化期望效用的动作。

(c)给定图 2 中每个状态的效用值，如果我们选择使用以下操作，会有什么策略:

我极大极小。

2 极大极大

将这些策略与 MEU 选择的策略进行比较(幻灯片中有)。

5.本教程没有额外的计算部分。部分原因是问题 2 和问题 3 中的电子表格是一种计算形式;这部分是因为你必须为课程作业实现一个 MDP 求解器，所以你将有机会接触到 Bellman 的计算方面。

1 在这里，“左”和“右”是相对于 agent 试图移动的方向。如果你觉得这很难想象，你可以试着把“左”想象成“逆时针 90 度”，把“右”想象成“顺时针 90 度”。

第 04 课回答 (版本 1.0)

1.我们要在留在办公室和出去之间做一个决定。这两种选择都将导致一个或多个状态。

待在办公室意味着我要么工作，要么分心，要么和同事聊天。这些州有以下实用程序:

$$U(work) = 8$$

U(分心)= 1u(同
事)= 5

而这些发生的概率，假设我在办公室里，是:

$$P(work|office) = 0.5$$

P(分心|办公室)= 0.3

P(同事|办公室)= 0.2 因
此，待在办公室的预期效用是:

$$EU(office) = 0.5 \cdot 8 + 0.3 \cdot 1 + 0.2 \cdot 5$$
$$= 5.3$$

外出可能会导致咖啡和泄漏，与公用事业:

$$U(coffee) = 10$$
$$U(spill) = -20$$

相关概率为:

P(coffee|out) = 0.95

因此 EU (out) = 8.5。

(b)最大期望效用原则说的是选择期望效用最大的期权，在这个例子中就是向外。

(c)极大值原则是指根据最坏结果的效用来选择选项，并根据最坏结果选择效用最大的选项。

正式我们有:

$$\begin{aligned} a &= \arg \max_{s \in \{office, out\}} \{min(U(s))\} \\ &= \arg \max_{s \in \{office, out\}} \{min^{office}(8,1,5), min^{out}(10,-20)) \\ &= \arg \max_{s \in \{office, out\}} \{1^{office}, -20^{out}\} \\ &= office \end{aligned}$$

所以最大化原则会说我应该留在办公室。

maximax 原则指的是根据每个行动的最佳结果选择一个选项，并选择一个具有最大效用的选项来获得最佳结果，换句话说:

$$a = \arg \max_{s \in \{office, out\}} \{max(U(s))\}$$

这样就能挑出来
了。2.网格世界是:

3

.

2

0110 +

-1

1 2 3 4

(a)作为 MDP 的正式描述如下:

- state:(1,1)，(1,2)，(1,3)，⋯(2,1)，(2,3)，⋯(4,3)
  注意，有障碍物的位置(2,2)并不对应于状态。
- 初始状态:(1,1)。
- 动作:所有状态上、下、左、右。
- 奖赏:

$$R(s) = \begin{cases} 1 & \text{for } s = (4,3) \\ -1 & \text{for } s = (4,2) \\ -0.04 & \text{otherwise} \end{cases}$$

- 转移模型，P(s [0]|年代,):

$$P((1,2)|(1,1), Up) = 0.8$$
$$P((1,1)|(1,1), Up) = 0.1$$
$$P((2,1)|(1,1), Up) = 0.1$$
$$P((1,1)|(1,1), Down) = 0.9$$
$$\vdots$$

(b)对于确定性模型，我们不再需要担心行为的预期效用，因为我们知道它是成功的，并且我们有:

$$U(s) = R(s) + \gamma max_{a \in A(s)} U(s')$$

其中 s 0 是动作 a 导致的状态。

(c)假设 $\gamma = 1$, U (s)初始值为 0。然后，利用上式进行第一轮值迭代后，得到:

3

.

2

-0. 04-0. 04

- 0. 96

0

1

-0. 04 1 0-0. 04-1

-0. 04-0. 04-0. 04

1 2 3. 4

第二轮之后，我们得到:

3

.

2

1 2 3 4

所以在 3 轮之后，我们会得到:

3 20.840.88 01100.920.960.92
+ 1-1
1 0.8 0.84 0.88 0.84

1 2 3 4

当值不再改变时。

有关计算的细节，请参阅电子表格(在 KEATS 上)。

(d)在每个状态下，最优策略是选择期望效用最高的行动。因为在这种情况下，动作是确定性的，所以这等同于决定使用最高的实用程序移动到状态(确保您理解为什么会这样)。因此，我们有:

3. — — —
0110
2
1

— — —

1 2 3 4

注意，我们假设代理在到达(4,3)或(4,2)时不会移动。3.(a)我的电子表格告诉我，当数值收敛时，我们有(四舍五入到小数点后两位):

3 0.87 0.910.96
20.820110 - 0.91-1
1 0.78 0.82 0.87
0.82

1 2 3 4

对于 γ =
1。

(3, 3)的第一次迭代计算为:

$$
\begin{array}{lll}
U(s) & = -0.04 + \gamma max(& (1 \cdot 0) \quad\quad\quad\quad\quad\quad \textit{Up} \\
 & & (0.9 \cdot 0 + 0.1 \cdot 0) \quad \textit{Down} \\
 & & (0.9 \cdot 0 + 0.1 \cdot 0) \quad \textit{Left} \\
 & & (0.9 \cdot 1 + 0.1 \cdot 0) \quad \textit{Right}
\end{array}
$$

其他州也是如此。

参阅电子表格(在 KEATS 上)了解其余的中间计算。

(b)每个州最佳行动的计算与上述计算并行。对于(3,3)，我们想确定 4 个可能动作的最大期望值:

(1 • 0.96)上(0.9 • 0.91 + 0.1 • 0.96)
下(0.9 • 0.91 + 0.1 • 0.96)左(0.9 • 1
+ 0.1 • 0.96)右

所以正确将是最好的行动。

重复这所有的州,我们很快发现,最优政策是一样的在确定性情况下(尽管政策的期望效用会降低自代理将花更多的时间在效用较低的州,并将采取更多的措施来达到我们的目标)。

这表明，对于像我们这里的这个运动模型，数值迭代的确定性版本是一个非常好的非确定性版本的近似值(尽管从图中可以明显看出实用程序的不同)。

4.(一)

$$U((3,1)) \quad = -0.04 + \gamma max(\begin{array}{l} 0.8 \cdot 0.660 + 0.1 \cdot 0.655 + 0.1 \cdot 0.388 \text{ 上涨} \\ 0.8 \cdot 0.611 + 0.1 \cdot 0.655 + 0.1 \cdot 0.388 \text{ 下降} \\ \text{左:}0.8 \cdot 0.655 + 0.1 \cdot 0.660 + 0.1 \cdot 0.611 \\ 0.8 \cdot 0.388 + 0.1 \cdot 0.660 + 0.1 \cdot 0.611 \text{ 正确} \end{array}$$

这样可以减少:

$$U((3,1)) \quad = -0.04 + \gamma max(\begin{array}{ll} 0.6323 & Up \\ 0.5931 & Down \\ 0.6511 & Left \\ 0.4375 & Right \end{array}$$

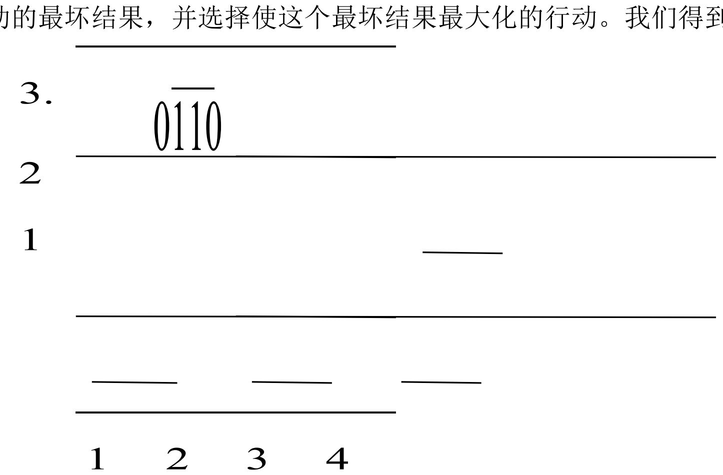
等于 0.611。

这和国家的效用是一样的。这正是我们从 Bellman 方程中所期望的，一旦我们得到了处于最优策略下的状态实用程序(这就是值迭代终止时所发生的情况)。

(b)上述方程中的数字是行动的预期效用。所以期望效用最大的动作是 Left。

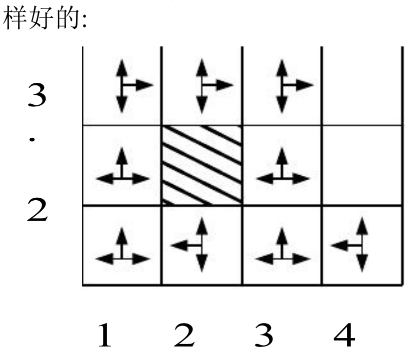
请注意，这与使用最大效用的状态是不同的。

对于最大化政策，我们着眼于每个行动的最坏结果，并选择使这个最坏结果最大化的行动。我们得到:



该策略确保在每个状态下，代理都没有机会进入(4,2)，同时确保它最终会到达(4,3)。

2 对于极大化策略，我们着眼于每个行动的最佳结果，并选择使其最大化的行动。因为每个行动都有三种可能的结果，所以总有三种行动可以带来最大的效用。因为 maximax 忽略了结果的概率，所以它认为所有这些行为都是同样好的:



maximax 没有选择的唯一操作是指向具有最大效用的状态的操作。

当然，一个代理不能选择三个操作，所以任何想要在这个环境中使用 maximax 的代理都必须选择这三个操作中的一个。

5.没有第 5 题。