计算机科学188：人工智能导论
2024年春季 笔记17

作者（所有其他笔记）：尼基尔·夏尔马
作者（贝叶斯网络笔记）：乔希·胡格和杰基·梁，由王瑞吉编辑
作者（逻辑笔记）：亨利·朱，由考佩林编辑
致谢（机器学习与逻辑笔记）：部分章节改编自教材《人工智能：一种现代方法》。

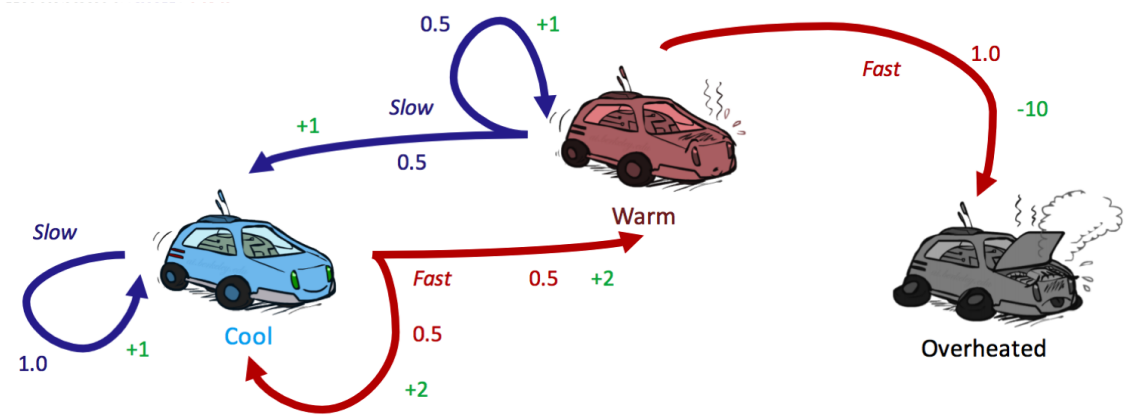
最后更新时间：2024年3月8日

马尔可夫决策过程

马尔可夫决策过程由几个属性定义：

- 一组状态 S。马尔可夫决策过程中的状态与传统搜索问题中的状态表示方式相同。
- 一组动作 A。马尔可夫决策过程中的动作与传统搜索问题中的表示方式也相同。
- 一个起始状态。
- 可能有一个或多个终止状态。
- 可能有一个折扣因子 γ。我们稍后会介绍折扣因子。
- 一个转移函数 $T\left(s, a, s'\right)$。由于我们引入了非确定性动作的可能性，我们需要一种方法来描述从任何给定状态采取任何给定动作后可能结果的可能性。MDP的转移函数正是为此而设计的——它是一个概率函数，表示智能体从状态 $s \in S$ 采取动作 $a \in A$ 后最终处于状态 $s' \in S$ 的概率。
- 一个奖励函数 $R\left(s, a, s'\right)$。通常，MDP在每一步都用小的“生存”奖励来建模，以奖励智能体的生存，同时对到达终止状态给予大的奖励。奖励可能是正的或负的，这取决于它们是否对相关智能体有益，而智能体的目标自然是在到达某个终止状态之前获得尽可能多的奖励。

为一种情况构建马尔可夫决策过程（MDP）与为搜索问题构建状态空间图非常相似，但有一些额外需要注意的地方。考虑赛车这个具有启发性的例子：



存在三种可能的状态， $S = \{$ 凉爽、温暖、过热 $\}$ ，以及两种可能的动作 $A = \{$ 慢速、快速 $\}$ 。就像在状态空间图中一样，三种状态中的每一种都由一个节点表示，边表示动作。过热是一个终端状态，因为一旦赛车智能体到达这个状态，它就不能再执行任何动作以获得进一步的奖励（它是马尔可夫决策过程中的一个汇点状态，没有出边）。值得注意的是，对于非确定性动作，从同一状态出发表示相同动作的边有多个，其后续状态不同。每条边不仅标注了它所代表的动作，还标注了转移概率和相应的奖励。这些总结如下：

- **Transition Function**: $T(s,a,s')$
  - $T(cool,slow,cool) = 1$
  - $T(warm,slow,cool) = 0.5$
  - $T(warm,slow,warm) = 0.5$
  - $T(cool,fast,cool) = 0.5$
  - $T(cool,fast,warm) = 0.5$
  - $T(warm,fast,overheated) = 1$
- **Reward Function**: $R(s,a,s')$
  - $R(cool,slow,cool) = 1$
  - $R(warm,slow,cool) = 1$
  - $R(warm,slow,warm) = 1$
  - $R(cool,fast,cool) = 2$
  - $R(cool,fast,warm) = 2$
  - $R(warm,fast,overheated) = -10$

我们用离散时间步长来表示智能体随时间在不同马尔可夫决策过程（MDP）状态间的移动，分别将 $s_t \in S$ 和 $a_t \in A$ 定义为智能体在时间步 t 所处的状态和采取的动作。智能体在时间步0时处于状态 s_0 ，并在每个时间步采取一个动作。因此，智能体在MDP中的移动可以建模如下：

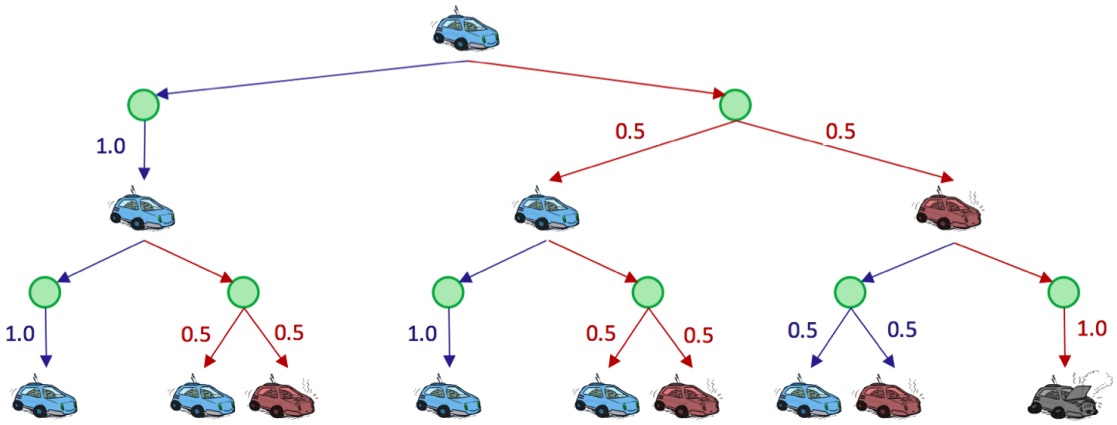
$$s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} s_3 \xrightarrow{a_3} \dots$$

此外，考虑到智能体的目标是在所有时间步上最大化其奖励，我们可以相应地将其数学表达为以下效用函数的最大化：

$$U([s_0,a_0,s_1,a_1,s_2,\dots]) = R(s_0,a_0,s_1) + R(s_1,a_1,s_2) + R(s_2,a_2,s_3) + \dots$$

马尔可夫决策过程与状态空间图一样，可以展开为搜索树。在这些搜索树中，不确定性是用Q状态来建模的，Q状态也称为动作状态，本质上与期望最大化机会节点相同。这是一个合适的选择，因为Q状态使用概率来建模环境将智能体置于给定状态的不确定性，就像期望最大化机会节点使用概率来建模对抗智能体通过它们选择的动作将我们的智能体置于给定状态的不确定性一样。从状态 s 采取动作 a 所表示的Q状态被记为元组(s,a)。

观察我们的赛车展开后的搜索树，截断到深度为2：



绿色节点表示Q状态，即已从某个状态采取了行动，但尚未解析为后继状态。重要的是要理解，智能体在Q状态下花费的时间步长为零，它们只是为了便于MDP算法的表示和开发而创建的一种结构。

有限视野与折扣

我们的赛车MDP存在一个固有问题——我们没有对赛车可以采取行动并收集奖励的时间步数设置任何时间限制。按照我们目前的公式，它可能会在每个时间步长都永远选择 $a =$ 慢速行驶，安全有效地获得无限奖励，而不会有任何过热风险。通过引入有限视野和/或折扣因子可以防止这种情况。一个实施有限视野的MDP很简单——它本质上为智能体定义了一个"生命周期"，这给了它们一些固定数量的时间步长 n ，以便在自动终止之前尽可能多地积累奖励。我们很快会回到这个概念。

折扣因子稍微复杂一些，引入它是为了对奖励价值随时间的指数衰减进行建模。具体来说，对于折扣因子 γ ，在时间步 <b3></b3 从状态 s_t 采取动作 a_t 并最终处于状态 s_{t+1} 会获得奖励 $\gamma^t R\left(s_t, a_t, s_{t+1}\right)$ ，而不是仅仅 $R\left(s_t, a_t, s_{t+1}\right)$ 。现在，不是最大化加法效用

$$U([s_0, a_0, s_1, a_1, s_2, \ldots]) = R(s_0, a_0, s_1) + R(s_1, a_1, s_2) + R(s_2, a_2, s_3) + \ldots$$

我们尝试最大化折扣效用

$$U([s_0, a_0, s_1, a_1, s_2, \ldots]) = R(s_0, a_0, s_1) + \gamma R(s_1, a_1, s_2) + \gamma^2 R(s_2, a_2, s_3) + \ldots$$

注意到上述折扣效用函数的定义看起来类似于公比为 γ 的几何级数，我们可以通过以下逻辑证明，只要满足约束 $|\gamma| < 1$ （其中 $|n|$ 表示绝对值运算符），它就保证是有限值：

$$\begin{aligned} U([s_0, s_1, s_2, \ldots]) &= R(s_0, a_0, s_1) + \gamma R(s_1, a_1, s_2) + \gamma^2 R(s_2, a_2, s_3) + \ldots \\ &= \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) \leq \sum_{t=0}^{\infty} \gamma^t R_{max} = \boxed{\frac{R_{max}}{1-\gamma}} \end{aligned}$$

其中 $R_{\max}$ 是在马尔可夫决策过程（MDP）中任何给定时间步可获得的最大可能奖励。通常，γ 严格从 $0 < \gamma < 1$ 范围内选择，因为在 $-1 < \gamma \leq 0$ 范围内的值在大多数实际情况中根本没有意义——γ 为负值意味着状态 s 的奖励在交替的时间步会在正值和负值之间翻转。

马尔可夫性

马尔可夫决策过程是"马尔可夫的"，因为它们满足马尔可夫性质，即无记忆性，该性质表明在给定当前状态的情况下，未来和过去是条件独立的。直观地说，这意味着，如果我们知道当前状态，那么知道过去并不能给我们提供更多关于未来的信息。为了用数学方式表达这一点，考虑一个在某个马尔可夫决策过程中采取行动 $a_0, a_1, \ldots, a_{t-1}$ 后访问了状态 $s_0, s_1, \ldots, s_t$ ，并且刚刚采取行动 a_t 的智能体。给定其先前访问的状态和采取的行动的历史记录，该智能体随后到达状态 s_{t+1} 的概率可以写成如下形式：

$$P(S_{t+1} = s_{t+1} | S_t = s_t, A_t = a_t, S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, ..., S_0 = s_0)$$

其中每个 S_t 表示代表我们智能体状态的随机变量， A_t 表示代表我们智能体在时间 t 采取的动作的随机变量。马尔可夫性质表明上述概率可以简化如下：

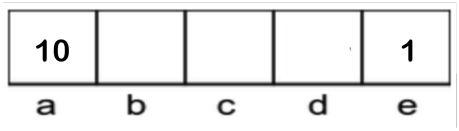
$$P(S_{t+1} = s_{t+1} | S_t = s_t, A_t = a_t, S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, ..., S_0 = s_0) = P(S_{t+1} = s_{t+1} | S_t = s_t, A_t = a_t)$$

它是"无记忆的"，即智能体在时间 $t+1$ 到达状态 s' 的概率仅取决于在时间 t 的状态 s 和动作 a ，而不取决于任何更早的状态或动作。实际上，正是这些无记忆概率由转移函数 $T(s, a, s') = P(s' \mid s, a)$ 进行编码。

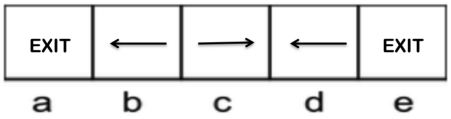
求解马尔可夫决策过程

回想一下，在确定性的、非对抗性搜索中，解决搜索问题意味着找到一个到达目标状态的最优计划。另一方面，解决马尔可夫决策过程意味着找到一个最优策略 $\pi^* : S \to A$ ，即一个将每个状态 $s \in S$ 映射到一个动作 $a \in A$ 的函数。一个显式策略 π 定义了一个反射智能体——给定一个状态 s ，在 s 处实施 π 的智能体将选择 $a = \pi(s)$ 作为合适的动作，而不考虑其动作的未来后果。最优策略是指如果实施智能体遵循它，将产生最大期望总奖励或效用的策略。

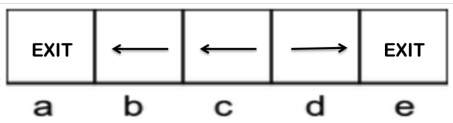
考虑以下马尔可夫决策过程，其状态有东、西、出口（出口仅在状态 a 和 e 时为有效动作，分别产生10和1的奖励），折扣因子 $\gamma = 0.1$ ，以及确定性转移：



此马尔可夫决策过程（MDP）的两种潜在策略如下：



(a) Policy 1



(b) Policy 2

经过一些研究，不难确定策略2是最优的。遵循该策略直到采取动作 $a =$ 退出，对于每个起始状态会产生以下奖励：

Start State	Reward
a	10
b	1
c	0.1
d	0.1
e	1

我们现在将学习如何使用马尔可夫决策过程的贝尔曼方程，通过算法解决此类马尔可夫决策过程（以及更复杂的情况！）。

贝尔曼方程

为了讨论马尔可夫决策过程的贝尔曼方程，我们必须首先引入两个新的数学量：

- 一个状态 $s, U^*(s)$ 的最优值 - s 的最优值是从 s 开始的最优行为智能体在其剩余生命周期内将获得的效用的期望值。请注意，在文献中，这个量经常用 $V^*(s)$ 表示。
- Q状态的最优值 $(s,a), Q^*(s,a)$ - (s,a)的最优值是智能体从 s 开始、采取 a 行动并从此最优行动后所获得效用的期望值。

使用这两个新量以及前面讨论的其他马尔可夫决策过程量，贝尔曼方程定义如下：

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

在开始解释这意味着什么之前，让我们也定义Q状态的最优值（更常见的称为最优Q值）的方程：

$$Q^*(s,a) = \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma U^*(s')]$$

请注意，这第二个定义使我们能够将贝尔曼方程重新表示为

$$U^*(s) = \max_a Q^*(s,a)$$

这是一个极其简单的量。贝尔曼方程是动态规划方程的一个例子，该方程通过固有的递归结构将一个问题分解为更小的子问题。我们可以在状态的Q值方程中的项 $\left[R\left(s,a,s'\right)+\gamma U^*\left(s'\right)\right]$ 中看到这种固有的递归。该项表示智能体通过首先从 s 采取 a 行动并到达 s'，然后从那时起采取最优行动所获得的总效用。所采取的行动 a 的即时奖励 $R\left(s,a,s'\right)$ 被加到从 $s',U^*\left(s'\right)$ 可获得的最优折扣奖励总和中，该总和通过 γ 进行折扣，以考虑在采取行动 a 时经过的一个时间步长。尽管在大多数情况下，从 s' 到某个终端状态存在大量可能的状态和行动序列，但所有这些细节都被抽象掉并封装在一个单一的递归值 $U^*\left(s'\right)$ 中。

我们现在可以再向外迈出一步，考虑Q值的完整方程。已知 $\left[R\left(s,a,s'\right)+\gamma U^*\left(s'\right)\right.$ 表示从Q状态（s，a）到达状态 s' 后通过最优行动获得的效用，很明显，这个量

$$\sum_{s'} T(s,a,s')[R(s,a,s')+\gamma U^*(s')]$$

仅仅是效用的加权和，每个效用都由其发生概率加权。根据定义，这就是从Q状态(s,a)开始最优行动的期望效用！这完成了我们的分析，并为我们提供了足够的洞察力来解释完整的贝尔曼方程——一个状态的最优值，$U^*\left(s\right)$，仅仅是从 s 的所有可能行动中的最大期望效用。计算状态 s 的最大期望效用本质上与运行期望最大化算法相同——我们首先从每个Q状态(s,a)计算期望效用（等同于计算机会节点的值），然后在这些节点上计算最大值以计算最大期望效用（等同于计算最大化节点的值）。

关于贝尔曼方程的最后一点说明——它的用途是作为最优性的一个条件。换句话说，如果我们能以某种方式为每个状态 $s\in S$ 确定一个值 $U\left(s\right)$，使得贝尔曼方程对这些状态中的每一个都成立，那么我们就可以得出这些值是其各自状态的最优值。实际上，满足这个条件意味着 $\forall s\in S,U\left(s\right)=U^*\left(s\right)$。