Belief state MDP

POMDP 简介

马尔可夫决策过程（Markov decision process，MDP）是一个具备完全信息情形下的决策过程。即智能体在每个时刻可以观测到其真实的状态 s_t ， $s_t \in S$ ，在经历行动 $a_t \in A$ 之后，可以到达下一个可观测的状态， s_{t+1} ，从而可以观测到真实的状态转移 $T(s_t, a_t, s_{t+1})$ ：

$$T(s_t, a_t, s_{t+1}) = P(s_{t+1}|s_t, a_t) : S \times A \times \mathcal{S} \to \mathbb{R}^+$$

也可以在每个时刻获得奖励的具体数值 $r_t(s_t, a_t, s_{t+1})$ ：

$$r_t(s_t, a_t, s_{t+1}) : S \times A \times S \to \mathbb{R}$$

但对于那些由于各种客观原因的限制，无法观测到完全信息的决策场景和系统，我们可以预设真实状态的决策过程依然是一个标准的 MDP 。只是智能体不能直接观察底层真实状态，仅能观察到其中一部分的信息 $o_t \in O$ 。那么，则可以称这种决策过程为一个部分观测马尔可夫决策过程（Partially observable Markov decision process，POMDP）。我们发现，相比于标准的马尔可夫决策过程 ，在部分观测的信息限制之下，真实状态依然可以完整确定一个观测，但部分观测已经无法作为一个完备统计量，完整还原真实状态，即对于任意一个观测序列 $[o_t]$ 和动作序列 $[a_t]$ ，其对应的真实状态 s_t 可能是整个状态空间 S 中的任意一个，无法唯一确定：

$$s_t \ \sim \ P(\cdot|[o_1, a_1, o_2, ..., o_{t-1}, a_{a-1}, o_t])$$

因此，POMDP 可以被视为某种 Infinite-State MDP。此外，相比于马尔可夫决策过程，部分观测马尔可夫决策过程可以额外引入观测信息 o_t 和真实状态 s_t 之间的函数关系，称为观测函数， $\Pi(o_{t+1}, s_t, a_t)$ ：

$$\Pi(o_{t+1}, s_t, a_t) = P(o_{t+1}|s_t, a_t) : S \times A \times \mathcal{O} \to \mathbb{R}^+$$

一般来说，解决 POMDP 问题有几种常用的方法，在课程正文中也已分别从数据建模，网络结构，训练方法等角度展开讲解。本文将介绍另外一种经典解法——使用 Belief State MDP 理论来求解 POMDP。

Belief State MDP 求解 POMDP 问题

使用 Belief State MDP 方法解决 POMDP 问题的适用条件是，当这个 POMDP 的真实状态空间 S 中的任意元素都可以被**完整**表述并**参数化**，这样我们才可以建立一个概率模型同时包含 o_t 和 s_t ，来感知智能体当前所处的状态可能是什么。比如对于常见的诸如扑克、麻将或军棋等非完全信息的博弈益智类卡牌游戏里，一些有限情况下，隐藏卡牌的信息是可以被参数化成某种离散的集合上的分类分布的。但对于很多现实世界中更为复杂的决策场景，比如根据股价历史信息来推断公司的运营状况这样的决

策问题，我们就无法简单地对这些真实存在但却复杂抽象的状态进行参数化，而是仅能对显式的观测信息进行参数化。

在具体操作中，会利用既往的观测和动作信息，以及对于初始状态的先验估计，来建立一种对当前真实状态的某种后验估计，在 Belief State MDP 理论中称之为信念状态（Belief State），记为 $b(s_t)$：

$$b(s_t) = P(s_t) : \mathcal{S} \to \mathbb{R}^+$$

也就是说，对于每个时刻 t ，智能体对当前时刻的真实状态的具体取值 s_t ，有一个参数化的分布 $P(s_t)$ 作为其估计。后续通过更多次与环境的交互，可以从交互带来的额外信息中，使用贝叶斯定理推断和逐步更新对当前真实状态信念的估计，可以把更新后的信念记为 $b'(s_t)$ 。

比如对于离散的场景下，其利用贝叶斯定理的更新形式为：

$$b'(s_{t+1}) = P(s_{t+1}|o_{t+1}, a_t, b(s_t)) = \frac{P(o_{t+1}|s_{t+1}, a_t, b(s_t))P(s_{t+1}|a_t, b(s_t))}{\sum_{s_{t+1}\in S}[P(o_{t+1}|s_{t+1}, a_t, b(s_t))P(s_{t+1}|a_t, b(s_t))]}$$

上式中，由于真实状态是观测信息的完备统计量，故有：

$$P(o_{t+1}|s_{t+1}, a_t, b(s_t)) = P(o_{t+1}|s_{t+1})$$

而对于下一时刻状态对于上一时刻信念和动作的条件先验，可以使用转移概率展开求解：

$$P(s_{t+1}|a_t, b(s_t)) = \sum_{s_t\in S} P(s_{t+1}|s_t, a_t)b(s_t)$$

这样一来，对于每一个时刻，我们都可以对当前的奖励做一个基于信念估计的函数计算：

$$R(a_t, b(s_t)) = \sum_{s_t\in S} r(a_t, s_t)b(s_t)$$

这样我们就让一个 POMDP 问题出现了一些熟悉的元素，比如我们用对于信念状态，来替代 MDP 中的状态。用奖励函数来替代 MDP 中的奖励。然后我们就可以使用一些适用于 MDP 的算法来近似解决 POMDP 问题。

比如我们可以定义一种策略 π ，基于信念状态，而非基于当前观测，来决定当前的动作：

$$a \sim \pi(b(s_t))$$

然后对于特定的策略 π ，我们可以引入这个策略对应的价值函数 V_π ：

$$V_\pi(b(s_t)) = \mathbb{E}(\sum_{i=0}^{\infty} \gamma^i R(a_i, b(s_i)))$$

然后使用诸如 Value Iteration 等方式优化 policy，然后依次循环与环境交互，收集数据，更新信念，更新策略。

示例 Toy example: Crying Baby

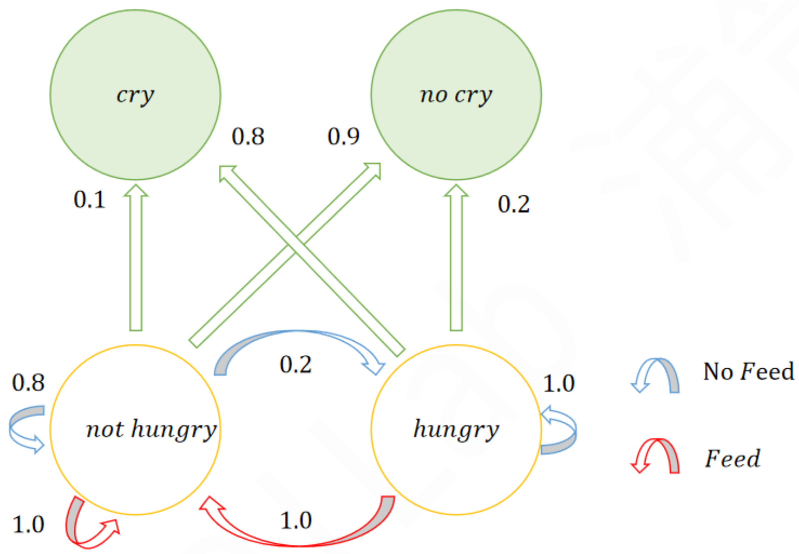
接下来我们用一个简易的例子，Crying Baby Problem，来描述一下 Belief State MDP 的信念更新过程。

Crying Baby Problem 描述的情形是，有一个婴儿宝宝经常大哭，原因主要可能是由于他饿了。已知经过统计，在饿肚子和不饿肚子的情况下，他哭泣的概率分别是：

$$P(o = \text{cry}|s = \text{hungry}) = 0.8$$

$$P(o = \text{cry}|s = \text{not hungry}) = 0.1$$

假如对饿肚子的 Baby 进行投喂 (Feed)，将会喂饱它。假如对不饿肚子的 Baby 不进行投喂，他将有 0.2 的概率转为饥饿状态。因此可以将 Baby 真实状态的马尔可夫过程表述为下图：



假设刚开始，对于 Baby 当前的状态，估计一个均等的信念状态作为先验，即：

$$b(s_0) = [P(\text{hungry}), P(\text{not hungry})] = [0.5, 0.5]$$

假如不对 Baby 进行投喂，发现 Baby 哭泣了，可以以此更新状态：

$$b(s_1) = P(s_1|o_1, a_0, b(s_0)) = \frac{P(o_1|s_1, a_0, b(s_0))P(s_1|a_0, b(s_0))}{\sum_{s_1 \in S}[P(o_1|s_1, a_0, b(s_0))P(s_1|a_0, b(s_0))]}$$

其中，

$$\begin{aligned} P(o_1|s_1, a_0, b(s_0)) &= P(o_1|s_1) \\ &= [[P(\text{cry}|\text{hungry}), P(\text{no cry}|\text{hungry})], [P(\text{cry}|\text{not hungry}), P(\text{no cry}|\text{not hungry}) \\ &= [[0.8, 0.2], [0.1, 0.9]] \end{aligned}$$

$$\begin{aligned} P(s_1|a_0, b(s_0)) = &\sum_{s_0 \in S} P(s_1|a_0, s_0)b(s_0) \\ &= [[P(\text{hungry}|\text{no feed}, \text{hungry})] \times [P(\text{hungry})] + [P(\text{hungry}|\text{no feed}, \text{not hungry})] \times [P(\text{not hungry})], \\ &[P(\text{not hungry}|\text{no feed}, \text{hungry})] \times [P(\text{hungry})] + [P(\text{not hungry}|\text{no feed}, \text{not hungry})] \times [P(\text{not hungry}) \\ &= [[1 \times 0.5] + [0.2 \times 0.5], [0 \times 0.5] + [0.8 \times 0.5]] = [0.6, 0.4] = [P(\text{hungry}), P(\text{not hungry})] \end{aligned}$$

代入可以计算得到：

$$
\begin{aligned}
b(s_1) = P(s_1|o_1, a_0, b(s_0)) &= \frac{P(o_1|s_1, a_0, b(s_0))P(s_1|a_0, b(s_0))}{\sum_{s_1 \in S}[P(o_1|s_1, a_0, b(s_0))P(s_1|a_0, b(s_0))]} \\
&= [\frac{[P(\text{cry}|\text{hungry}) \times P(\text{hungry})]}{[P(\text{cry}|\text{hungry}) \times P(\text{hungry})] + [P(\text{cry}|\text{not hungry}) \times P(\text{not hungry})]}, \\
&\frac{[P(\text{cry}|\text{not hungry}) \times P(\text{not hungry})]}{[P(\text{cry}|\text{hungry}) \times P(\text{hungry})] + [P(\text{cry}|\text{not hungry}) \times P(\text{not hungry})]}] \\
&= [\frac{0.8 \times 0.6}{0.8 \times 0.6 + 0.1 \times 0.4}, \frac{0.1 \times 0.4}{0.8 \times 0.6 + 0.1 \times 0.4}] \\
&= [\frac{12}{13}, \frac{1}{13}] = [P(\text{hungry}), P(\text{not hungry})]
\end{aligned}
$$

从而完成了一次对信念状态的更新。

总结

总的来说，Belief State MDP 方法是一种求解 POMDP 问题的一种较为理想的解法。但由于信念状态是连续的，导致了对于任意观测的无限状态集的出现，这使得与 MDP 相比，POMDP 更难解决。此外，由于信念状态的更新需要涉及到大量的计算，Belief State MDP 方法无法适用于真实状态过于复杂的场景。这是 Belief State MDP 方法虽然作为 POMDP 问题的理想解法，但在实际复杂问题中无法直接使用的原因。