1 Markov Decision Process

1.1 马尔科夫决策过程理论

1.1.1 马尔科夫性

马尔科夫性：状态 s_t 是马尔科夫的，当且仅当 $P[s_{t+1}|s_t]=P[s_{t+1}|s_1,\ldots,s_t]$ 。

马尔科夫性是指系统的下一个状态 s_{t+1} 仅与当前状态 s_t 有关，而与以前的状态无关。

随机过程：指随机变量序列。

马尔科夫随机过程：随机变量序列中的每个状态都是马尔科夫的。

1.1.2 马尔科夫过程

马尔科夫过程：马尔科夫过程是一个二元组 (S,P) ，且满足： S 是有限状态集合， P 是状态转移概率。状态转移概率矩阵为 $\begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \vdots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix}$ 。

状态序列称为马尔科夫链，当给定状态转移概率时，从某个状态出发存在多条马尔科夫链。

1.1.3 马尔科夫决策过程

马尔科夫决策过程：将动作（策略）和回报考虑在内的马尔科夫过程。

马尔科夫决策过程由元组 (S,A,P,R,γ) 描述，其中：

S 为有限的状态集

A 为有限的动作集

P 为状态转移概率

R 为回报函数

γ 为折扣因子，用来计算累积回报

强化学习的目标是给定一个马尔科夫决策过程，寻找最优策略。所谓策略是指状态到动作的映射，策略通常用符号 π 表示，它是指给定状态 s 时，动作集上的一个分布，即

$$\pi(a|s)=p[A_t=a|S_t=s] \tag{1}$$

即策略的定义是用条件概率分布给出的。公式（1）的含义是：策略 π 在每个状态 s 指定一个动作概率。如果给出的策略 π 是确定性的，那么策略 π 在每个状态 s 指定一个确定的动作。

当给定一个策略 π 时，就可以计算累积回报。首先定义累积回报：

$$G_t=R_{t+1}+\gamma R_{t+2}+\cdots=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \tag{2}$$

由于策略 π 是随机的，因此累积回报也是随机的。为了评价状态 s_1 的价值，我们需要定义一个确定量来描述状态 s_1 的价值，很自然的想法是利用累积回报来衡量状态 s_1 的价值。然而，累积回报 G_1 是一个随机变量，不是一个确定值，因此无法描述，但其期望是个确定值，可以作为状态值函数的定义。

（1）状态值函数。

当智能体采用策略 π 时，累积回报服从一个分布，累积回报在状态 s 处的期望值定义为状态-值函数：

$$\nu_\pi(s) = E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s] \tag{3}$$

注意：状态值函数是与策略 π 相对应的，这是因为策略 π 决定了累积回报 G 的状态分布。

相应地，状态-行为值函数为

$$q_\pi(s,a) = E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a] \tag{4}$$

（2）状态值函数与状态-行为值函数的贝尔曼方程。

由状态值函数的定义式（ 3 ）可以得到：

$$\begin{gathered} \nu(s) = E[G_t|S_t = s] \\ = E[R_{t+1} + \gamma R_{t+2} + \cdots|S_t = s] \\ = E[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots)|S_t = s] \\ = E[R_{t+1} + \gamma G_{t+1}|S_t = s] \\ = E[R_{t+1} + \gamma\nu(S_{t+1})|S_t = s] \end{gathered} \tag{5}$$

同样我们可以得到状态-动作值函数的贝尔曼方程：

$$q_\pi(s,a) = E_\pi[R_{t+1} + \gamma q(S_{t+1}, A_{t+1})|S_t = s, A_t = a] \tag{6}$$

计算状态值函数的目的是为了构建学习算法从数据中得到最优策略。每个策略对应着一个状态值函数，最优策略自然对应着最优状态值函数。

定义：最优状态值函数 $\nu^*(s)$ 为在所有策略中值最大的值函数，即 $\nu^*(s) = \max \nu_\pi(s)$ ，最优状态-行为值函数 $q^*(s,a)$ 为在所有策略中最大的状态-行为值函数，即

$$q^*(s,a) = \max q_\pi(s,a) \tag{7}$$

最优状态值函数和最优状态-行为值函数的贝尔曼最优方程为：

$$\nu^*(s) = \max R_s^a + \gamma \sum_{s' \in S} P_{SS'}^a \nu(s') \tag{8}$$

$$q^*(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{SS'}^a \max_{a'} q^*(s',a') \tag{9}$$

若已知最优状态-行为值函数，最优策略可通过直接最大化 $q^*(s,a)$ 来决定。

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max\limits_{a \in A} q_*(s,a) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

综上所述，我们定义了一个离散时间有限范围的折扣马尔科夫决策过程 $M = (S, A, P, r, \rho_0, \gamma, T)$ ，其中 S 为状态集， A 为动作集， $P: S \times A \times S \to R$ 是转移概率， $r: S \times A \to [-R_{max}, R_{max}]$ 为立即回报函数， $\rho_0: S \to R$ 是初始状态分布， $\gamma \in [0,1]$ 为折扣因子， T 为水平范围（步数）。 τ 为一个轨迹序列，即 $\tau = (s_0, a_0, s_1, a_1, \cdots)$ ，累积回报为 $R = \sum_{t=0}^{T} \gamma^t r_t$ ，强化学习的目标是找到最优策略 π ，使得该策略下的累积回报期望最大，即 $\max\limits_{\pi} \int R(\tau) p_\pi(\tau) d\tau$ 。

1.2 MDP中的概率学基础

在强化学习算法中，随机策略得到广泛应用，因为随机策略耦合了探索。随机策略常用符号 π 表示，它是指给定状态 s 时动作集上的一个分布。

1.2.1 随机变量及其分布

（1）随机变量。

随机变量是指可以随机地取不同值的变量，常用小写字母表示。在MDP中随机变量是指当前的动作，用字母 a 表示。

（2）概率分布。

概率分布用来描述随机变量在每个可能取到的值处的可能性大小。离散性随机变量的概率分布常用概率质量函数来描述，即随机变量在离散点处的概率。连续性随机变量的概率分布则用概率密度函数来描述。

（3）条件概率。

策略 $\pi(a|s)$ 是条件概率。条件概率是指在其他事件发生时，我们所关心的事件所发生的概率。$\pi(a|s)$ 是指在当前状态 s 处，采取某个动作 a 的概率。当给定随机变量后，状态 s 处的累积回报 $G(s)$ 也是随机变量，而且其分布由随机策略 π 决定。状态值函数定义为该累积回报的期望。

（4）期望。

函数 $f(x)$ 关于某分布 $P(x)$ 的期望是指，当 x 由分布 $P(x)$ 产生、 f 作用于 x 时， $f(x)$ 的平均值。

对于离散型随机变量，期望公式为

$$E_{x\sim P}[f(x)] = \sum_x P(x)f(x) \tag{11}$$

对于连续型随机变量，期望通过积分求得

$$E_{x\sim P}[f(x)] = \int p(x)f(x)dx \tag{12}$$

（5）方差。

方差是衡量利用当前概率分布采样时，采样值的差异大小，可用如下公式得到：

$$Var(f(x)) = E[(f(x) - E[f(x)])^2] \tag{13}$$

方差越小，采样值离均值很近，不确定性越小。方差的平方根被称为标准差。

1.2.2 随机策略

（1）贪婪策略。

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max\limits_{a\in A} q_*(s,a) \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

贪婪策略是一个确定性策略，即只有在使得动作值函数 $q^*(s,a)$ 最大的动作处取概率1,选其他动作的概率为0。

（2） ϵ - greedy 策略。

$$\pi(a|s) \leftarrow \begin{cases} 1-\epsilon+\frac{\epsilon}{|A(s)|} & \text{if } a = \text{argmax}_a Q(s,a) \\ \frac{\epsilon}{|A(s)|} & \text{if } a \neq \text{argmax}_a Q(s,a) \end{cases} \tag{15}$$

ϵ - greedy策略是强化学习最基本最常用的随机策略。其含义是选取使得动作值函数最大的动作的概率为 $1-\epsilon+\frac{\epsilon}{|A(s)|}$ ，而其他动作的概率为等概率，都为 $\frac{\epsilon}{|A(s)|}$ 。 ϵ - greedy平衡了利用（exploitation）和探索（exploration），其中选取动作值函数最大的部分为利用，其他非最优动作仍有概率为探索部分。

（3）高斯策略。

一般高斯策略可以写成 $\pi_\theta=\mu_\theta+\epsilon$， $\epsilon\sim N(0,\sigma^2)$ 。其中 μ_θ 为确定性部分， ϵ 为零均值的高斯噪声。高斯策略也平衡了利用和探索，其中利用由确定性部分完成，探索由 ϵ 完成。高斯策略在连续系统的强化学习中应用广泛。

（4）波尔兹曼分布。

对于动作空间是离散的或者动作空间并不太大的情况，可采用波尔兹曼分布作为随机策略，即

$$\pi(a|s,\theta)=\frac{\exp(Q(s,a,\theta))}{\sum_b\exp(h(s,b,\theta))} \tag{16}$$

其中 $Q(s,a,\theta)$ 为动作值函数。该函数的含义是，动作值函数大的动作被选中的概率大，动作值函数小的动作被选中的概率小。