强化学习基础算法的公式总结（持续更新） - 工程小猿的博客

前言

在最近学习强化学习过程中遇到很多强化学习算法与公式，随着数量越来越多，我开始觉得思路越发混乱，便上网寻找各个算法的公式总结；但搜索一番后发现，没有找到；于是乎决定自己尝试总结一下；好废话结束，开始正文。

第一章：马尔科夫决策过程公式

1.1 马尔科夫性（马尔科夫过程）

马尔科夫过程只涉及到状态到状态的转移，未涉及到动作、策略与奖励

1.1.1 状态转移概率

状态转移概率为从一个状态转移到其他后继状态的转移概率：

$$P_{ss'} = \mathbb{P}\left[S_{t+1} = s' | S_t = s\right]$$

使用矩阵表示：（每一行加起来值为1）

$$P = \begin{array}{cc} & \text{to} \\ \text{from} & \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & & \\ P_{n1} & \cdots & P_{nn} \end{bmatrix} \end{array}$$

1.2 马尔科夫奖励过程（MRP）

它是由 $\langle S, P, R, \gamma \rangle$ 构成的一个元组，其中：

S 是一个有限状态集

P 是集合中状态转移概率矩阵：$P_{ss'} = \mathbb{P}\left[S_{t+1} = s' | S_t = s\right]$

R 是一个奖励函数:$R_s = \mathbb{E}\left[R_{t+1} | S_t = s\right]$

γ 是一个衰减因子:$\gamma \in [0, 1]$

1.2.1 收获

指的是从某一个状态开始直到终止状态时所有奖励的有衰减的之和。

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

1.2.2 价值

指的是马尔科夫奖励过程中某状态的收获的期望：

$$v(s) = \mathbb{E}\left[G_t | S_t = s\right]$$

但是计算时不可能把某状态经过的所有的序列都找到（特别是状态含有自循环时），所以展开上式：

$$\begin{aligned} v(s) &= \mathbb{E}\left[G_t | S_t = s\right] \\ &= \mathbb{E}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots | S_t = s\right] \\ &= \mathbb{E}\left[R_{t+1} + \gamma\left(R_{t+2} + \gamma R_{t+3} + \ldots\right) | S_t = s\right] \\ &= \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} | S_t = s\right] \\ &= \mathbb{E}\left[R_{t+1} + \gamma v\left(S_{t+1}\right) | S_t = s\right] \end{aligned}$$

$$v(s) = \mathbb{E}\left[R_{t+1} + \gamma v\left(S_{t+1}\right) | S_t = s\right]$$

最终得贝尔曼方程：

$$v(s) = R_s + \gamma \sum_{s' \in S} P_{ss'} v\left(s'\right)$$

其中 v(s’) 表示下一个状态的状态值函数。贝尔曼方程也可以写成矩阵形式进行计算：

$$v = R + \gamma P v$$

它表示：

$$\begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix} = \begin{bmatrix} R_1 \\ \vdots \\ R_n \end{bmatrix} + \gamma \begin{bmatrix} P_{11} & \cdots P_{1n} \\ \vdots & \\ P_{n1} & \cdots P_{nn} \end{bmatrix} \begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix}$$

理论上，该方程可以直接求解：

$$\begin{aligned} v &= R + \gamma P v \\ (1 - \gamma P) v &= R \\ v &= (1 - \gamma P)^{-1} R \end{aligned}$$

1.3 马尔科夫决策过程（MDP）

相比马尔科夫奖励过程加入了动作和策略。

由 $\langle S, A, P, R, \gamma \rangle$ 构成的一个元组，其中：

S 是一个有限状态集

A 是一个有限行为集

P 是集合中基于行为的状态转移概率矩阵：$P_{ss'}^a = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$

R 是基于状态和行为的奖励函数：$R_s^a = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$

γ 是一个衰减因子：$\gamma \in [0, 1]$

1.3.1 策略

策略（policy）与状态转移概率(Pss’)虽然都是概率，但意义不同要区分开，状态转移概率(Pss’)在动作执行之后，而策略(policy)在动作执行之前:

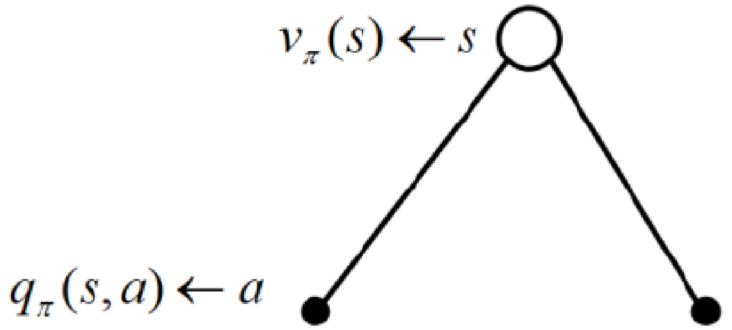
$$\pi(a|s) = \mathbb{P}[A_t = a \mid S_t = s]$$

1.3.2 基于策略的状态值函数

$$v_\pi(s) = \mathbb{E}[G_t | S_t = s]$$

1.3.3 基于策略的状态行为值函数

一个状态行为对下的价值称为状态行为函数，使用 q(s,a) 表示；如下图：



$$q_\pi(s,a) = \mathbb{E}\left[G_t | S_t = s, A_t = a\right]$$

1.3.4 基于策略的霍尔曼方程

$$v_\pi(s) = \mathbb{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s\right]$$

$$q_\pi(s,a) = \mathbb{E}\left[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a\right]$$

转换后得：

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) q_\pi(s,a)$$

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')$$

上式相互带入最终得到计算公式：

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')\right)$$

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \sum_{a' \in A} \pi(a'|s') q_\pi(s',a')$$

参考文献

叶强老师的强化学习笔记：https://zhuanlan.zhihu.com/p/37690204

郭宪老师的《深入浅出强化学习——原理入门》