强化学习读书笔记 - 00 - 术语和数学符号

学习笔记：

Reinforcement Learning: An Introduction, Richard S. Sutton and Andrew G. Barto c 2014, 2015, 2016

基本概念

Agent - 本体。学习者、决策者。
Environment - 环境。本体外部的一切。
s - 状态(state)。一个表示环境的数据。
$S, \mathcal{S}$ - 所有状态集合。环境中所有的可能状态。
a - 行动(action)。本体可以做的动作。
$A, \mathcal{A}$ - 所有行动集合。本体可以做的所有动作。
$A(s), \mathcal{A}(s)$ - 状态 s 的行动集合。本体在状态 s 下，可以做的所有动作。
r - 奖赏(reward)。本体在一个行动后，获得的奖赏。
$\mathcal{R}$ - 所有奖赏集合。本体可以获得的所有奖赏。

S_t - 第t步的状态(state)。t from 0
A_t - 第t步的行动(select action)。t from 0
R_t - 第t步的奖赏(reward)。t from 1
G_t - 第t步的长期回报(return)。t from 0。 **强化学习的目标1：追求最大回报**

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

$where$
k - the sequence number of an action.
γ - discount rate, $0 \leqslant \gamma \leqslant 1$

可以看出，当$\gamma = 0$时，只考虑当前的奖赏。当$\gamma = 1$时，未来的奖赏没有损失。
$G_t^{(n)}$ - 第t步的n步回报(n-step return)。。一个回报的近似算法。

$$G_t^{(n)} \doteq \sum_{k=0}^{n} \gamma^k R_{t+k+1} \tag{2}$$

$where$
k - the sequence number of an action.
γ - discount rate, $0 \leqslant \gamma \leqslant 1$

G_t^{λ} - 第t步的λ回报(λ-return)。一个回报的近似算法。可以说是$G_t^{(n)}$的优化。

Continuing tasks:

$$G_t^\lambda \doteq (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{(n)} \tag{3}$$

Episodic tasks:

$$G_t^\lambda \doteq (1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1}G_t^{(n)} + \lambda^{T-t-1}G_t$$

where

$\lambda \in [0,1]$

$$(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1} = 1$$

$$(1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1} + \lambda^{T-t-1} = 1$$

if $\lambda = 0$, become to 1-step TD algorithm

if $\lambda = 1$, become to Monte Carlo algorithm

策略

π - 策略(policy)。**强化学习的目标2：找到最优策略**。

策略规定了状态s时，应该选择的行动a。

$$\pi = [\pi(s_1), \cdots, \pi(s_n)] \tag{4}$$

$\pi(s)$ - 策略π在状态s下，选择的行动。

π_* - 最优策略(optimal policy)。
$\pi(a|s)$ - **随机策略**π在状态s下，选择的行动a的概率。

$r(s,a)$ - 在状态s下，选择行动a的奖赏。
$r(s,a，s')$ - 在状态s下，选择行动a，变成(状态s‘)的奖赏。
$p(s',r|s,a)$ - (状态s、行动a)的前提下，变成(状态s‘、奖赏r)的概率。
$p(s'|s,a)$ - (状态s、行动a)的前提下，变成(状态s‘)的概率。
$v_\pi(s)$ - 状态价值。使用策略π，(状态s的)长期奖赏G_t。
$q_\pi(s,a)$ - 行动价值。使用策略π，(状态s，行动a的)长期奖赏G_t。
$v_*(s)$ - 最佳状态价值。
$q_*(s,a)$ - 最佳行动价值。
$V(s)$ - $v_\pi(s)$的集合。
$Q(s,a)$ - $q_\pi(s,a)$的集合。

For continuing tasks: (5)

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

For episodic tasks:

$$G_t \doteq \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}$$

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s \right]$$

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t|S_t = s, A_t = a] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a \right]$$

$$v_\pi(s) = \max_{a \in \mathcal{A}} q_\pi(s,a)$$

$$\pi(s) = \underset{a}{argmax}\ v_\pi(s'|s,a)$$

$\pi(s)$ is the action which can get the next state which has the max value.

$$\pi(s) = \underset{a}{argmax}\ q_\pi(s,a)$$

$\pi(s)$ is the action which can get the max action value from the current state.

由上面的公式可以看出：$\pi(s)$可以由$v_\pi(s)$或者$q_\pi(s,a)$决定。

$$
\begin{aligned}
&\text{Reinforcement Learning} \doteq \pi_* \qquad (6)\\
&\quad \updownarrow \\
&\pi_* \doteq \{\pi(s)\},\ s \in \mathcal{S} \\
&\quad \updownarrow \\
&\begin{cases} \pi(s) = \underset{a}{argmax}\ v_\pi(s'|s,a),\ s' \in S(s), \quad \text{or} \\ \pi(s) = \underset{a}{argmax}\ q_\pi(s,a) \end{cases} \\
&\quad \updownarrow \\
&\begin{cases} v_*(s), \quad \text{or} \\ q_*(s,a) \end{cases} \\
&\quad \updownarrow \\
&\text{approximation cases:} \\
&\begin{cases} \hat{v}(s,\theta) \doteq \theta^T \phi(s), \quad \text{state value function} \\ \hat{q}(s,a,\theta) \doteq \theta^T \phi(s,a), \quad \text{action value function} \end{cases} \\
&where \\
&\theta \text{ - value function's weight vector}
\end{aligned}
$$

强化学习的目标3：找到最优价值函数$v_*(s)$或者$q_*(s,a)$。

近似计算

强化学习的目标4：找到最优近似价值函数$\hat{v}(S_t,\theta_t)$或者$\hat{q}(S_t,A_t,\theta_t)$。
强化学习的目标5：找到求解θ。
ρ_t^k - importance sampling ratio for time t to time k - 1。
$\mathcal{J}(s)$ - 状态s被访问的步骤序号。

θ - 近似价值函数的权重向量。
$\phi(s)$ - 近似价值函数的特征函数。是一个将状态s转化成计算向量的方法。这个结果和θ组成近似价值函数。
$\hat{v}(S_t, \theta_t)$ - 近似状态价值函数。

$$\hat{v} \doteq \theta^T \phi(s) \tag{7}$$

$\hat{q}(S_t, A_t, \theta_t)$ - 近似行动价值函数。

$$\hat{q} \doteq \theta^T \phi(s, a) \tag{8}$$

e_t - 第t步资格迹向量(eligibility trace rate)。可以理解为近似价值函数微分的优化值。

$$\begin{aligned} e_0 &\doteq 0 \\ e_t &\doteq \nabla \hat{v}(S_t, \theta_t) + \gamma \lambda e_{t-1} \\ \theta_t &\doteq \theta_t + \alpha \delta_t e_t \end{aligned} \tag{9}$$

α - 学习步长。$\alpha \in (0, 1)$
γ - 未来回报的折扣率(discount rate)。$\gamma \in [0, 1]$
λ - λ-return中的比例参数。$\lambda \in [0, 1]$
h（horizon）- 水平线h表示on-line当时可以模拟的数据步骤。$t < h \leq T$

老O虎O机问题

$q_*(a)$ - 行动 a 的真实奖赏(true value)。这个是（实际中）不可知的。期望计算的结果收敛(converge)与它。
$N_t(a)$ - 在第t步之前，行动a被选择的次数。
$Q_t(a)$ - 行动 a 在第t步前（不包括第t步）的实际平均奖赏。

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \times 1_{A_i=a}}{N_t(a)} \tag{10}$$

$H_t(a)$ - 对于行动a的学习到的倾向(reference)。
ϵ - 在ε-贪婪策略中，采用随机行动的概率 $[0, 1)$。

通用数学符号

$\doteq$ - 定义上的等价关系。
$\mathbb{E}[X]$ - X的期望值。
$Pr\{X = x\}$ - 变量X值为x的概率。
$v \mapsto g$ - v渐近g。
$v \approx g$ - v约等于g。
$\mathbb{R}$ - 实数集合。
$\mathbb{R}^n$ - n个元素的实数向量。
$\underset{a \in \mathcal{A}}{max}\ F(a)$ - 在所有的行动中，求最大值$F(a)$。
$\underset{c}{argmax}\ F(c)$ - 求当F(c)为最大值时，参数c的值。

术语

episodic tasks - 情节性任务。指（强化学习的问题）会在有限步骤下结束。
continuing tasks - 连续性任务。指（强化学习的问题）有无限步骤。
episode - 情节。指从起始状态（或者当前状态）到结束的所有步骤。
tabular method - 列表方法。指使用了数组或者表格存储每个状态（或者状态-行动）的信息（比如：其价值）。

planning method - 计划性方法。需要一个模型，在模型里，可以获得状态价值。比如： 动态规划。
learning method - 学习性方法。不需要模型，通过模拟（或者体验），来计算状态价值。比如：蒙特卡洛方法，时序差分方法。

on-policy method - on-policy方法。评估的策略和优化的策略是同一个。
off-policy method - off-policy方法。评估的策略和优化的策略不是同一个。意味着优化策略使用来自外部的样本数据。
target policy - 目标策略。off-policy方法中需要优化的策略。
behavior policy - 行为策略μ。off-policy方法中提供样本数据的策略。
importance sampling - 行为策略μ的样本数据。
importance sampling rate - 由于目标策略π和行为策略μ不同，导致样本数据在使用上的加权值。
ordinary importance sampling - 无偏见的计算策略价值的方法。
weighted importance sampling - 有偏见的计算策略价值的方法。
MSE(mean square error) - 平均平方误差。
MDP(markov decision process) - 马尔科夫决策过程
The forward view - We decide how to update each state by looking forward to future rewards and states.
例如：

$$G_t^{(n)} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n \hat{v}(S_{t+n}, \theta_{t+n-1}),\ 0 \leq t \leq T - n \tag{11}$$

The backward or mechanistic view - Each update depends on the current TD error combined with eligibility traces of past events.
例如：

$$\begin{aligned} e_0 &\doteq 0 \\ e_t &\doteq \nabla \hat{v}(S_t, \theta_t) + \gamma \lambda e_{t-1} \end{aligned} \tag{12}$$

参照

- Reinforcement Learning: An Introduction, Richard S. Sutton and Andrew G. Barto c 2014, 2015, 2016
- 强化学习读书笔记 - 01 - 强化学习的问题
- 强化学习读书笔记 - 02 - 多臂老O虎O机问题

- 强化学习读书笔记 - 03 - 有限马尔科夫决策过程
- 强化学习读书笔记 - 04 - 动态规划
- 强化学习读书笔记 - 05 - 蒙特卡洛方法(Monte Carlo Methods)