

Homework 4 Written

June 17th, 2020 at 11:59pm

1 Reinforcement Learning

Imagine an unknown game which has only two states {A, B} and in each state the agent has two actions to choose from: {Up, Down}. Suppose a game agent chooses actions according to some policy π and generates the following sequence of actions and rewards in the unknown game:

t	s_t	a_t	s_{t+1}	r_t
0	A	Down	B	-2
1	B	Down	B	-4
2	B	Up	B	0
3	B	Up	A	3
4	A	Up	A	1

Unless specified otherwise, assume a discount factor $\gamma = 0.5$ and a learning rate $\alpha = 0.5$.

1. Recall the update function of Q-learning is:

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha \left(r_t + \gamma \max_{a'} Q(s_{t+1}, a') \right)$$

Assume that all Q-values initialized as 0. What are the following Q-values learned by running Q-learning with the above experience sequence?

$$Q(A, \text{Down}) = \underline{\text{---}}, \quad Q(B, \text{Up}) = \underline{\text{---}}$$

2. In model-based reinforcement learning, we first estimate the transition function $T(s, a, s')$ and the reward function $R(s, a, s')$. Fill in the following estimates of T and R, estimated from the experience above. Write "n/a" if not applicable or undefined.

$$\hat{T}(A, \text{Up}, A) = \underline{\text{---}}^{\text{1}} \quad \hat{T}(A, \text{Up}, B) = \underline{\text{---}}^{\text{0}} \quad \hat{T}(B, \text{Up}, A) = \underline{\text{---}}^{\text{0.5}} \quad \hat{T}(B, \text{Up}, B) = \underline{\text{---}}^{\text{0.5}}$$

$$\hat{R}(A, \text{Up}, A) = \underline{\text{---}}^{\text{1}} \quad \hat{R}(A, \text{Up}, B) = \underline{\text{---}}^{\text{n/a}} \quad \hat{R}(B, \text{Up}, A) = \underline{\text{---}}^{\text{3}} \quad \hat{R}(B, \text{Up}, B) = \underline{\text{---}}^{\text{0}}$$

3. To decouple this question from the previous one, assume we had a different experience and ended up with the following estimates of the transition and reward functions:

s	a	s'	$\hat{T}(s,a,s')$	$\hat{R}(s,a,s')$
A	Up	A	1	10
A	Down	A	0.5	2
A	Down	B	0.5	2
B	Up	A	1	-5
B	Down	B	1	8

- (a) Give the optimal policy $\hat{\pi}^*(s)$ and \hat{V}^*s for the MDP with the transition function \hat{T} and the reward function \hat{R} .

Hint: for any $x \in \mathbb{R}$, $|x| < 1$, we have $1 + x + x^2 + x^3 + x^4 + \dots = 1/(1 - x)$

$$\hat{\pi}^*(A) = \underline{\text{Up}} \quad \hat{\pi}^*(B) = \underline{\text{Down}} \quad \hat{V}^*(A) = \underline{20} \quad \hat{V}^*(B) = \underline{16}$$

- (b) If we repeatedly feed this new experience sequence through our Q-learning algorithm, what values will it converge to? Assume the learning rate α_t is properly chosen so that convergence is guaranteed.
- the value found above, \hat{V}^*
 - the optimal values, V^*
 - neither \hat{V}^* nor V^*
 - not enough information to determine

2 Policy Evaluation

In this question, you will be working in an MDP with states S , actions A , discount factor γ , transition function T , and reward function R .

We have some fixed policy $\pi : S \rightarrow A$, which returns an action $a = \pi(s)$ for each state $s \in S$. We want to learn the Q function $Q^\pi(s, a)$ for this policy: the expected discounted reward from taking action a in state s and then continuing to act according to π : $Q^\pi(s, a) = \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma Q^\pi(s', \pi(s'))]$. The policy π will not change while running any of the algorithms below.

1. Can we guarantee anything about how the values Q^π compare to the values Q^* for an optimal policy π^* ?

- (a) $Q^\pi(s, a) \leq Q^*(s, a)$ for all s, a
- (b) $Q^\pi(s, a) = Q^*(s, a)$ for all s, a
- (c) $Q^\pi(s, a) \geq Q^*(s, a)$ for all s, a
- (d) None of the above guaranteed

2. Suppose T and R are unknown. You will develop sample-based methods to estimate Q^π . You obtain a series of samples $(s_1, a_1, r_1), (s_2, a_2, r_2), \dots, (s_T, a_T, r_T)$ from acting according to this policy (where $a_t = \pi(s_t)$, for all t)

- (a) Recall the update equation for the Temporal Difference algorithm, performed on each sample in sequence:

$$V(s_t) \leftarrow (1 - \alpha)V(s_t) + \alpha(r_t + \gamma V(s_{t+1}))$$

which approximates the expected discounted reward $V^\pi(s)$ for following policy π from each state s , for a learning rate α . Fill in the blank below to create a similar update equation which will approximate Q^π using the samples. You can use any of the terms $Q, s_t, s_{t+1}, a_t, a_{t+1}, r_t, r_{t+1}, \gamma, \alpha, \pi$ in your equation, as well as \sum and \max with any index variables (i.e. you could write \max_a or \sum_a and then use a somewhere else), but no other terms.

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha[r_t + \gamma Q(s_{t+1}, a_{t+1})]$$

- (b) Now, we will approximate Q^π using a linear function: $Q(s, a) = \sum_{i=1}^d w_i f_i(s, a)$ for weights w_1, \dots, w_d and feature functions $f_1(s, a), \dots, f_d(s, a)$.

To decouple this part from the previous part, use Q_{samp} for the value in the blank in part (a) (i.e. $Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha Q_{\text{samp}}$). Which of the following is the correct sample-based update for each w_i ?

- i. $w_i \leftarrow w_i + \alpha [Q(s_t, a_t) - Q_{\text{samp}}]$
- ii. $w_i \leftarrow w_i - \alpha [Q(s_t, a_t) - Q_{\text{samp}}]$
- iii. $w_i \leftarrow w_i + \alpha [Q(s_t, a_t) - Q_{\text{samp}}] f_i(s_t, a_t)$
- iv. $w_i \leftarrow w_i - \alpha [Q(s_t, a_t) - Q_{\text{samp}}] f_i(s_t, a_t)$
- v. $w_i \leftarrow w_i + \alpha [Q(s_t, a_t) - Q_{\text{samp}}] w_i$
- vi. $w_i \leftarrow w_i + \alpha [Q(s_t, a_t) - Q_{\text{samp}}] w_i$

- (c) The algorithms in the previous parts (part a and b) are:

- i. model-based
- ii. model-free