دانشکده مهندسي کامپيوتر

هوش مصنوعي و سيستم‌هاي خبره

تمرين تشريحي پنجم[1]

نام و نام خانوادگي - شماره دانشجويي . . . . . غزل زمانی نژاد 97522166. . .

مدرس محمدطاهر پيله‌ور - سيد صالح اعتمادي

طراحي و تدوين . . . . مرسده ايراني - مهسا قادران

تاريخ انتشار . . . . ۶۲ آبان ۹۹۳۱

تاريخ تحويل گروه ۱ . . . . آذر ۹۹۳۱

تاريخ تحويل گروه ۲ . . . . آذر ۹۹۳۱

[1] در طراحي اين تمرين از منابع کورس CS188 دانشگاه برکلي استفاده شده است.

١ 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 Y *= 0.5 and a learningrate* α *= 0.5*

١.١

Recall the update function of Q-learning is:

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

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

$Q(\text{A}, \text{Down}) =$ 1 , $\quad Q(\text{B}, \text{Up}) =$ 1.75

Your Solution:

$$Q(A, down) \leftarrow (1-0.5)\overbrace{Q(A, down)}^{0} + 0.5(2 + 0.5 \times 0) = 1$$

	A	B	C
up		1.75	
down	1	-2	

$$Q(B, down) \leftarrow 0.5\underbrace{Q(B, down)}_{0} + 0.5(-4 + \underbrace{0.5 \times 0}_{0}) = -2$$

$$Q(B, up) \leftarrow 0.5Q(B, up) + 0.5(0 + 0.5 \times 0) = 0$$

$$Q(B, up) \leftarrow 0.5\underbrace{Q(B, up)}_{0} + 0.5(3 + 0.5 \times \max[\underbrace{(A, up)}_{0}, \underbrace{(A, down)}_{1}]) = \frac{3.5}{2} = 1.75$$

۲.۱

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.

$\hat{T}(\mathrm{A}, \mathrm{Up}, \mathrm{A}) =$ ________, $\hat{T}(\mathrm{A}, \mathrm{Up}, \mathrm{B}) =$ ________, $\hat{T}(\mathrm{B}, \mathrm{Up}, \mathrm{A}) =$ ________, $\hat{T}(\mathrm{B}, \mathrm{Up}, \mathrm{B}) =$ ________

$\hat{R}(\mathrm{A}, \mathrm{Up}, \mathrm{A}) =$ ________, $\hat{R}(\mathrm{A}, \mathrm{Up}, \mathrm{B}) =$ ________, $\hat{R}(\mathrm{B}, \mathrm{Up}, \mathrm{A}) =$ ________, $\hat{R}(\mathrm{B}, \mathrm{Up}, \mathrm{B}) =$ ________

Your Solution:

$T(A, Up, A) = 1$ $\quad$ $T(A, up, B) = 0$ $\quad$ $T(B, up, A) = \frac{1}{2}$ $\quad$ $T(B, up, B) = \frac{1}{2}$

$R(A, Up, A) = -1$ $\quad$ $R(A, up, B) = n/a$ $\quad$ $R(B, up, A) = 3$ $\quad$ $R(B, up, B) = 0$

طبق جدول بالا در 50٪ موارد با اکشن up به A رفته‌ایم

۳.۱

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

۱.۳.۱

Give the optimal policy $\hat{\pi}^*$ (s) and $\hat{V}^*$ (s) for the MDP with transition function $\hat{T}$ and reward function $\hat{R}$.
Hint: for any $x \in R$, $|x| < 1$, we have $1 + x + x^2 + x^3 + x^4 + \ldots = 1/(1 - x)$

$\hat{\pi}^*(A) =$ up, $\hat{\pi}^*(B) =$ down $\hat{V}^*(A) =$ 20, $\hat{V}^*(B) =$ 16.

Your Solution:

1.3) $v^{\pi}_{k+1} \leftarrow \sum_{s'} T(a,\pi(s),s')[R(s,\pi(s),s') + \gamma v^{\pi}_{k}(s')]$

$A \xrightarrow{up} 1[10 + 0.5\, \hat{v}_k] = 10$

$A \xrightarrow{down} 0.5[2 + 0.5\, v^{\pi}_k(A)] + 0.5[2 + 0.5\, v^{\pi}_k(B)] = 2$

argmax[(2, down), (10, up)] = $\boxed{\text{up}}$

$\hat{v}^*_A = \sum_{i=0}^{\infty} (0.5)^i\, 10 = \frac{10}{1-0.5} = \boxed{20}$

$B \xrightarrow{up} 1[-5 + 0.5\, v^{\pi}_k(A)] = -5$

$B \xrightarrow{down} 1[8 + 0.5 \times v^{\pi}_k(B)] = 8$

~ argmax[(-5, up), (8, down)] = $\boxed{\text{down}}$

$\hat{v}^*_B = \sum_{i=0}^{\infty} (0.5)^i\, 8 = \frac{8}{1-0.5} = \boxed{16}$

۲.۳.۱

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.

1) the values found above, $\hat{V}^*$

2) the optimal values, V^*

3) neither $\hat{V}^*$ nor V^*

4) not enough information to determine

Explain your answer in less than 2lines:

جواب گزینه ۱ $\hat{V}^*$. در کل Q-learning ، مقادیر لزوما به optimal value ، converge نمی‌شود

1.3.2)

چون همه چیز براساس تجربه و موارد آموخته شده محاسبه می‌شود . اما به مقادیر آموخته شده converge می‌کند

۲ Policy Evaluation

In this question, you will be working in an MDP with states S, actions A, discount factor Y, transition function T and reward function R.
We have some fixed policy π : S → A, which returns an action a = π(s) for each state $s \in S$.We want to learn the Q function Q^{π}(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.

۱.۲

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

1) $Q^{\pi}(s,a) \le Q^*(s,a)$ for all s,a
2) $Q^{\pi}(s,a) = Q^*(s,a)$ for all s,a
3) $Q^{\pi}(s,a) \ge Q^*(s,a)$ for all s,a
4) None of the above are guaranteed

Explain your answer in less than 2lines:

2.1) جواب گزینه ۱؛ چون مقادیر Q براساس learning بدست می‌آید، مقادیر آن حتما کمتر و یا مساوی با Q^* است

۲.۲

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), ..., (s_T,a_T,r_T) from acting according to this policy (where $a_t = \pi(s_t)$, for all t).

۱.۲.۲

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}, Y , α, π in your equation, as well as Σ and max with any index variables (i.e. you could write max_a, or Σ_a and then use a somewhere else), but no other terms.

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\,[\underline{\hspace{6cm}}]$$

Explain your answer in less than 2 lines:

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

مشابه با V است.

۲.۲.۲

Now, we will approximate Q^π using a linear function: $Q(s,a) = \Sigma^d_{i=1} w_i f_i(s,a)$ for weights $w_1,..,w_d$ and feature functions $f_1(s,a),...,f_d(s,a)$.
To decouple this part from the previous part, use Q_{samp} for the value in the blank in part (2.2.1) (i.e. $Q(s_t,a_t) \leftarrow (1-\alpha)\, Q(s_t,a_t) + \alpha Q_{samp}$).
Which of the following is the correct sample-based update for each w_i?

1) $w_i \leftarrow w_i + \alpha[Q(s_t, a_t) - Q_{samp}]$
2) $w_i \leftarrow w_i - \alpha[Q(s_t, a_t) - Q_{samp}]$
3) $w_i \leftarrow w_i + \alpha[Q(s_t, a_t) - Q_{samp}]f_i(s_t, a_t)$
4) $w_i \leftarrow w_i - \alpha[Q(s_t, a_t) - Q_{samp}]f_i(s_t, a_t)$
5) $w_i \leftarrow w_i + \alpha[Q(s_t, a_t) - Q_{samp}]w_i$
6) $w_i \leftarrow w_i - \alpha[Q(s_t, a_t) - Q_{samp}]w_i$

Explain your answer in less than 2 lines:

2.2.2) جواب گزینه 4: که بدست آوردن Q approximate، باید مقدار difference را در f_i ضرب کنیم
تا بتوانیم جاهایی که تا به حال مشاهده نشده اند نیز تجربه بدست بیاوریم

۳.۲.۲

The algorithms in the previous parts (part 2.2.1 and 2.2.2) are:
1)model-based   2)model-free

Explain your answer in less than 2 lines:

3.2.2) model-free هستند؛ دانش سوال‌ها بر اساس learn و کسب تجربه به مقادیر Q و V دسترسی پیدا کردیم اما
در model-based، مدل از ابتدا به ما داده شده