

دانشکده مهندسي کامپيوتر هوش مصنوعي و سيستمهاي خبره

تمرین تشریحی پنجم۱

خانوادگي - شماره دانشجويي غزل زماني نژاد 97522166	نام و نام .
محمدطاهر پيلهور - سيد صالح اعتمادي	مدرس
، و تدوین مرسده ایراني - مهسا قادران	طراحي
تشار ۴۲ آبان ۹۹۳۱	تاريخ ان
عویل گروه ۱ آذر ۹۹۳۱ عویل گروه ۱	تاريخ تد
ویل گروه ۲	تاريخ تح

۱ در طراحي اين تمرين از منابع كورس CS188 دانشگاه بركلي استفاده شده است.

١

1.1



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	В	2
1	В	Down	В	-4
2	В	Up	В	0
3	В	Up	A	3
4	A	Up	A	-1

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

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?

$$Q(A, Down) = \underline{1}, \qquad Q(B, Up) = \underline{1.75}$$

Your Solution:

$$Q(A, down) \leftarrow (1-0.5) Q(A, down) + 0.5(2 + 0.5 \times 0) = 1 \quad \text{up} \qquad 1.75$$

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

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

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

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.

$\hat{T}(A, Up, A) = \underline{\hspace{1cm}},$	$\hat{T}(A, Up, B) = \underline{\hspace{1cm}},$	$\hat{T}(B, Up, A) = \underline{\hspace{1cm}},$	$\hat{T}(B, Up, B) = \underline{\hspace{1cm}}$
$\hat{R}(A, Up, A) = \underline{\hspace{1cm}},$	$\hat{R}(A, Up, B) = \underline{\hspace{1cm}},$	$\hat{R}(B, U_{P}, A) = \underline{\hspace{1cm}},$	$\hat{R}(B, Up, B) =$

Your Solution:
$$T(A, \cup p, A) = 1 \qquad T(A, \cup p, B) = 0 \qquad T(B, \cup p, A) = \frac{1}{2} \qquad T(B, \cup p, B) = \frac{1}{2}$$

$$R(A, \cup p, A) = -1 \qquad R(A, \cup p, B) = n/\alpha \qquad R(B, \cup p, A) = 3 \qquad R(B, \cup p, B) = 0$$

$$R(A, \cup p, A) = -1 \qquad R(B, \cup p, B) = 0$$



٣.١

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
\mathbf{A}	Down	A	0.5	2
A	Down	В	0.5	2
В	Up	A	1	-5
В	Down	В	1	8

1.7.1

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 $1 + x + x^2 + x^3 + x^4 + ... = 1/(1 - x)$

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

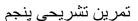
Your Solution:

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

A $\xrightarrow{up} I[10 + 0.5 v_{k}] = 10$
 $cog_{max} ((z, down), (10, up)] = up$
 $down = 0.5[2 + 0.5 v^{\pi}_{k}(A)] + 0.5[2 + 0.5 v^{\pi}_{k}(B)] = 2$
 $\hat{v}^{\pi}_{A} = \sum_{i=0}^{\infty} (0.5)^{i} 10 = \frac{10}{1 - 0.5} = \frac{20}{20}$

B $\xrightarrow{up} I[-5 + 0.5 v^{\pi}_{k}(A)] = -5$
 $down = 1[8 + 0.5 \times v^{\pi}_{k}(B)] = 8$
 $\hat{v}^{\#}_{B} = \sum_{i=0}^{\infty} (0.5)^{i} 8 = \frac{8}{1 - 0.5} = \frac{16}{10}$







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 2 lines:

جواب گزیند یا * فی ، م کن Q-learning مادر دورد الدورد مادر مواد و مادر کرد و دارد الدخت را مادر کرد و دارد الدخت را عام مادد . اما به مادر الدخت را موادد الدورد و دارد الدخت را مادر کرد در دارد الدخت را عام مادد .



Policy Evaluation ۲

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 $\pi: S \to A$, 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 :

$$Q^{\pi}(s, a) = \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma Q^{\pi}(s', \pi(s'))]$$

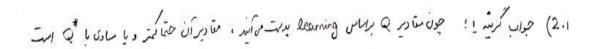
The policy π will not change while running any of the algorithms below.

1.7

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

- 1) $Q^{\pi}(s, a) \leq Q^*(s, a)$ 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 2 lines:



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).

1.7.7

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.

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha$$

Explain your answer in less than 2 lines:

7.7.7

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 $w_1,...,w_d$ 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 wi?

- 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:

in up fi us difference in in approximate Q is in the in use (2.2.2 نَا بَوَاشِ عِلْمَا يُكُم نَا مِ حَالَ مُنْ هُذُهُ اللَّهُ مُرْتِ بَوْمِ بِدِمْتَ بِيادِيمٍ



تمرين تشريحي پنجم هوش مصنوعي و سيستمهاي خبره

٣.٢.٢

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:

model_free (3.2.2 مستد؛ دران کوال ها بر اماس الوست کرد به مادیر Q و لا دختری بداردیم اما مدل از امتیاب ما داده منه