به نام خدا

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

مبانی و کاربردهای هوش مصنوعی ترم پاییز ۱۴۰۱

پاسخنامه تمرین سوم

سوال ۱)

۱- نادرست است. مطابق با اسلایدها، فقط به حالت کنونی وابسته است:

- "Markov" generally means that given the present state, the future and the past are independent

۲- نادرست است. مطابق با اسلایدها، در direct evaluation پس از انجام چند sample با میانگین گرفتن ارزش را آپدیت می کنیم.

۳- نادرست است. یک مثال نقض در نظر می‌گیریم. یک MDP با دو حالت ترمینال A و B را در نظر بگیرید. انتقال به حالت A پاداش 1 و انتقال به حالت B پاداش 10 دارد. همه انتقال های دیگر پاداش صفر دارند. فرض کنید A یک قدم به سمت شمال از حالت شروع باشد. همچنین B دو قدم به سمت جنوب از حالت شروع باشد. (اقدامات همیشه موفق هستند). حالا اگر $\gamma < 0.1$ باشد سیاست بهینه عامل را یک قدم به سمت شمال از حالت شروع به A می‌برد، اما اگر $\gamma > 0.1$ سیاست بهینه عامل را دو گام به سمت جنوب از حالت شروع به B می‌رساند.

۴- درست است.Q-learning نیازی به مدلی از محیط ندارد و model-free است. همچنین اگر غیر بهینه (suboptimal) عمل شود، باز هم Q-learning به سیاست بهینه همگرا می‌شود و off-policy است.

سوال ۲) سرعت هلیکوپتر، موقعیت و زاویه با زمین هلیکوپتر، فاصله تا نزدیکترین مانع، سرعت باد و وضعیت آب و هوا و ...

حتی اگر از ویژگی های گفته شده استفاده کنیم و همه ضرایب را پیدا کنیم هم باز ممکن است به مشکل بر بخوریم. مثلا فرض کنید که هلیکوپتر به شکل برعکس در حال پرواز است. ممکن است با موانع فاصله داشته باشیم و همچنین همه چیز خوب باشد اما قطعا پرواز به شکل برعکس مطلوب نیست.

* در این سوال پاسخ های قابل قبول دیگر نیز درست هستند.

سوال ۳)

الف) بله، همچنان یک سیاست بهینه دریافت می‌شود. این روش عملاً مشابه value iteration است، زیرا value iteration شامل یک مرحله ارزیابی(evaluation) نیز می شود. در این مورد، ما مقادیر را بر اساس بهترین سیاست فعلی خود به روز می کنیم و این کار را تا زمان همگرایی یک سیاست ادامه می دهیم.

فرمول value iteration به صورت زیر است:

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s')\right]$$

و فرمول policy evaluation به صورت زیر است:

$$V_{k+1}^{\pi_i}(s) \leftarrow \sum_{s'} T(s,\pi_i(s),s') \left[R(s,\pi_i(s),s') + \gamma V_k^{\pi_i}(s')\right]$$

با داشتن یک سیاست ثابت π که در مرحله policy improvement به روز رسانی می‌کنیم:

$$\pi_{i+1}(s) = \arg\max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^{\pi_i}(s')\right]$$

که بسیار شبیه به policy extraction ای است که در value iteration انجام می‌شود.

ب) بله. با تکرارهای کافی، عامل می‌تواند همچنان Q-value ها را از اکشن‌های خود محاسبه کند. با این حال، یادگیری به طور قابل توجهی بیشتر از حالت $\varepsilon < 1$ طول می‌کشد، زیرا عامل نمی‌تواند مقادیر Qای که یاد گرفته را exploit کند، بنابراین اشتباهات مشابه را مکررا انجام می دهد تا زمانی که راه حل صحیح به طور تصادفی پیدا شود.

سوال ۴) طبق روابط زیر محاسبات را انجام می‌دهیم.

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s')\right]$$

$$Q_{k+1}(s,a) \leftarrow \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} Q_k(s',a')\right]$$

$V_1(A) = \max \{1\ [1 + 0]\} = 1$

$V_1(B) = \max \{1\ [1 + 0], 1[-2 + 0]\} = 1$

$V_1(C) = \max \{(0.5)[1 + 0] + (0.5)[4 + 0], 1[-2 + 0]\} = 2.5$

$V_1(D) = \max \{1[1 + 0], 1[-2 + 0]\} = 1$

$V_1(E) = \max \{1[1 + 0], 1[-2 + 0]\} = 1$

$V_1(F) = \max \{1[10 + 0], 1[-2 + 0]\} = 10$

$V_1(G) = 0$

$V_2(A) = \max \{1\ [1 + 1]\} = 2$

$V_2(B) = \max \{1\ [1 + 2.5], 1[-2 + 1]\} = 3.5$

$V_2(C) = \max \{(0.5)[1 + 1] + (0.5)[4 + 1], 1[-2 + 2.5]\} = 3.5$

$V_2(D) = \max \{1\ [1 + 1], 1[-2 + 2.5]\} = 2$

$V_2(E) = \max \{1[1 + 10], 1[-2 + 1]\} = 11$

$V_2(F) = \max \{1[10 + 0], 1[-2 + 1]\} = 10$

$V_2(G) = 0$

پاسخ‌های نهایی به صورت زیر است:

$V_2(B) = 3.5$

$Q_2(B, right) = 1\ [1 + 2.5] = 3.5$

$Q_2(B, left) = 1[-2 + 1] = -1$

	A	B	C	D	E	F	G
V_1	1	1	2.5	1	1	10	0
V_2	2	3.5	3.5	2	11	10	0

ب) طبق به روز رسانی در Q-learning داریم:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + (\alpha)\left[r + \gamma \max_{a'} Q(s',a')\right]$$

$Q(C, jump) = (0.5)(0) + (0.5)(4 + 0) = 2$

$Q(E, right) = (0.5)(0) + (0.5)(1 + 0) = 0.5$

$Q(F, left) = (0.5)(0) + (0.5)(-2 + \max\{Q(E, right), Q(E,left)\}) = 0.5(-2 + \max\{0.5, 0\}) = -0.75$

$Q(E, right) = (0.5)(0.5) + (0.5)(1 + \max\{Q(F, right), Q(F,left)\}) = 0.5(1 + \max\{0,-0.75\}) = 0.75$

	Q(C, left)	Q(C, jump)	Q(E, left)	Q(E, right)	Q(F, left)	Q(F, right)
Initial	0	0	0	0	0	0
Transition 1		2				
Transition 2				0.5		
Transition 3					-0.75	
Transition 4				0.75		

یک state-action تنها زمانی به روز می‌شود که یک انتقال از آن انجام گیرد. Q(C,left), Q(E,left), Q(F, right) هرگز تجربه نشده‌اند و بنابراین این مقادیر به روز نمی‌شوند.

سوال ۵)

اولین مرحله از policy iteration, انجام policy evaluation است که مطابق با فرمول زیر انجام می‌شود:

$$V_{k+1}^{\pi_i}(s) \leftarrow \sum_{s'} T(s, \pi_i(s), s') \left[R(s, \pi_i(s), s') + \gamma V_k^{\pi_i}(s') \right]$$

در ابتدا از k=0 و i=0 شروع می‌کنیم که مقدار $V_0^{\pi 0}$ برای هر وضعیت در صورت سوال داده شده است. ابتدا مقدار V1 را برای سیاست ابتدایی محاسبه میکنیم:

Iteration 1:
Policy evaluation (calculate utilities):

$V^{\pi 0}{}_1(\text{High}) = T(\text{High, study, High}) \cdot (R(\text{High, study, High}) + \gamma V_0(\text{High}))$
$T(\text{High, study, Low}) \cdot (R(\text{High, study, Low}) + \gamma V_0(\text{Low}))$
$= 1.0 \times (0 + \gamma 0) + 0 = 0$

$V^{\pi 0}{}_1(\text{Low}) = T(\text{Low, study, High}) \cdot (R(\text{Low, study, High}) + \gamma V_0(\text{High}))$
$T(\text{Low, study, Low}) \cdot (R(\text{Low, study, Low}) + \gamma V_0(\text{Low}))$
$= 0.3 \times (0 + \gamma 0) + 0.7 \times (0 + \gamma 0) = 0$

یک بار دیگر ارزیابی را انجام می دهیم تا ببینیم که آیا value همگرا می شود یا خیر:

$V^{\pi 0}{}_2(\text{High}) = T(\text{High, study, High}) \cdot (R(\text{High, study, High}) + \gamma V_1(\text{High}))$
$T(\text{High, study, Low}) \cdot (R(\text{High, study, Low}) + \gamma V_1(\text{Low}))$
$= 1.0 \times (0 + \gamma 0) + 0 = 0$

$V^{\pi 0}{}_2(\text{Low}) = T(\text{Low, study, High}) \cdot (R(\text{Low, study, High}) + \gamma V_1(\text{High}))$
$T(\text{Low, study, Low}) \cdot (R(\text{Low, study, Low}) + \gamma V_1(\text{Low}))$
$= 0.3 \times (0 + \gamma 0) + 0.7 \times (0 + \gamma 0) = 0$

می‌بینیم که مقدار V تغییر نکرد، پس همگرا شده. در مرحله بعد، policy improvement را مطابق زیر انجام می‌دهیم:

$$\pi_{i+1}(s) = \arg\max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^{\pi_i}(s') \right]$$

توجه داشته باشید که V در فرمول بالا و محاسبات زیر، برابر با V ای است که در قسمت policy evaluation بدست آوردیم. مطابق با روند زیر، سیاست را بهبود می بخشیم.

Policy improvement:
For state "high":

- Utility of choosing "study": $1 \cdot (0 + 0.5 \cdot 0) + 0 \cdot (0 + 0.5 \cdot 0) = 0$
- Utility of choosing "exam": $0.9 \cdot (10 + 0.5 \cdot 0) + 0.1 \cdot (-10 + 0.5 \cdot 0) = 8$
- Utility of choosing "Netflix": $1.0 \cdot (1 + 0.5 \cdot 0) = 1$

So set $\boxed{\hat{\pi}_1(high) \leftarrow exam}$.
For state "low":

- Utility of choosing "study": $0.3 \cdot (0 + 0.5 \cdot 0) + 0.7 \cdot (0 + 0.5 \cdot 0) = 0$
- Utility of choosing "exam": $0.05 \cdot (10 + 0.5 \cdot 0) + 0.95 \cdot (-10 + 0.5 \cdot 0) = -9$
- Utility of choosing "Netflix": $1.0 \cdot (1 + 0.5 \cdot 0) = 1$

So set $\boxed{\hat{\pi}_1(low) \leftarrow Netflix}$.

در iteration دوم، باز هم مانند قبل عمل می‌کنیم:

مرحله اول: policy evaluation

ابتدا توجه کنید چون بر روی value همگرا شدیم، در ابتدای این پیمایش خواهیم داشت:

$$V^{\pi 1}{}_0(\text{High}) = 0$$
$$V^{\pi 1}{}_0(\text{Low}) = 0$$

حال تا سه مرحله همگرایی را بررسی می کنیم. توجه کنید که سیاست استفاده شده در این مراحل از سیاست بدست آمده در policy improvement است.

Iteration 2:
Policy evaluation:

$V^{\pi 1}{}_1$(High) = T(High, exam, High) . (R(High, exam, High) + γ V_0(High))
T(High, exam, Low) . (R(High, exam, Low) + γ V_0(Low))
= 0.9 x (10 + γ 0) + 0.1 x (-10 + γ 0) = 8

$V^{\pi 1}{}_1$(Low) = T(Low, Netflix, High) . (R(Low, Netflix, High) + γ V_0(High))
T(Low, Netflix, Low) . (R(Low, Netflix, Low) + γ V_0(Low))
= 0.0 x (1 + γ 0) + 1.0 x (1+ γ 0) = 1

$V^{\pi 1}{}_2$(High) = T(High, exam, High) . (R(High, exam, High) + γ V_1(High))
T(High, exam, Low) . (R(High, exam, Low) + γ V_1(Low))
= 0.9 x (10 + γ 8) + 0.1 x (-10 + γ 1) = 12.6 – 0.95 = 11.65

$V^{\pi 1}{}_2$(Low) = T(Low, Netflix, High) . (R(Low, Netflix, High) + γ V_1(High))
T(Low, Netflix, Low) . (R(Low, Netflix, Low) + γ V_1(Low))
= 0.0 x (1 + γ 8) + 1.0 x (1+ γ 1) = 1.5

$V^{\pi 1}{}_3$(High) = T(High, exam, High) . (R(High, exam, High) + γ V_2(High))
T(High, exam, Low) . (R(High, exam, Low) + γ V_2(Low))
= 0.9 x (10 + γ 11.65) + 0.1 x (-10 + γ 1.5) = 13.3175

$V^{\pi 1}{}_3$(Low) = T(Low, Netflix, High) . (R(Low, Netflix, High) + γ V_2(High))
T(Low, Netflix, Low) . (R(Low, Netflix, Low) + γ V_2(Low))
= 0.0 x (1 + γ 11.65) + 1.0 x (1+ γ 1.5) = 1.75

Policy improvement:
For state "high":

- Utility of choosing "study": 1 . (0 + 0.5 . 13.3175) + 0 . (0 + 0.5 . 1.75) = 6.66
- Utility of choosing "exam": 0.9 . (10 + 0.5 . 13.3175) + 0.1 . (-10 + 0.5 . 1.75) = 14.99 – 0.91 = 14.08
- Utility of choosing "Netflix": 1 . (1 + 0.5 . 13.3175) + 0 . (0 + 0.5 . 1.75) = 7.66

So set $\hat{\pi}_2(high) \leftarrow exam$
For state "low":

- Utility of choosing "study": 0.3 . (0 + 0.5 . 13.3175) + 0.7 . (0 + 0.5 . 1.75) = 1.99 + 0.6125 = 2.6
- Utility of choosing "exam": 0.05 . (10 + 0.5 . 13.3175) + 0.95 . (-10 + 0.5 . 1.75) = 0.83 -8.66 = -7.83
- Utility of choosing "Netflix" 1 . (1 + 0.5 . 1.75) = 1.875

So set $\hat{\pi}_2(low) \leftarrow study$

می‌بینیم که سیاست تغییر کرد فلذا سیاست‌ها هنوز همگرا نشده‌اند.

سوال ۶)

الف) با استفاده از Q-learning و محاسبه سه نمونه زیر داریم:

$Q(A, Go) \leftarrow (1 - \alpha)Q(A, Go) + \alpha(r + \gamma \max Q(B, a')) = 0.5(0) + 0.5(2) = 1$

$Q(C, Stop) \leftarrow (1 - \alpha)Q(C, Stop) + \alpha(r + \gamma \max Q(A, a')) = 0.5(0) + 0.5(0 + 1) = 0.5$

$Q(C, Go) \leftarrow (1 - \alpha)Q(C, Go) + \alpha(r + \gamma \max Q(A, a')) = 0.5(0) + 0.5(2 + 1) = 1.5$

ب) محاسبات را طبق روابط زیر انجام می‌دهیم.

$$Q(s,a) = w_1 f_1(s,a) + w_2 f_2(s,a) + \ldots + w_n f_n(s,a)$$

$$\text{difference} = \left[r + \gamma \max_{a'} Q(s',a')\right] - Q(s,a)$$

$$w_i \leftarrow w_i + \alpha\,[\text{difference}]\, f_i(s,a)$$

پس از اولین به روزرسانی:

$Q(A, Go) = w_1 f_1(A, Go) + w_2 f_2(A, Go) = 0$

$\text{difference} = [r + \max Q(B, a')] - Q(A, Go) = [4 + 0] - 0 = 4$

$w_1 = w_1 + \alpha(\text{difference}) f_1 = 0 + (0.5)(4)(1) = 2$

$$w_2 = w_2 + \alpha(\text{difference})f_2 = 0 + (0.5)(4)(1) = 2$$

پس از دومین به روز رسانی:

$$Q(B, \text{Stop}) = w_1 f_1(B, \text{Stop}) + w_2 f_2(B, \text{Stop}) = 2(1) + 2(-1) = 0$$

$$Q(A, \text{Go}) = w_1 f_1(A, \text{Go}) + w_2 f_2(A, \text{Go}) = 2(1) + 2(1) = 4$$

$$\text{difference} = [r + \max Q(A, a)] - Q(B, \text{Stop}) = [0 + 4] - 0 = 4$$

$$w_1 = w_1 + \alpha(\text{difference})f_1 = 2 + (0.5)(4)(1) = 4$$

$$w_2 = w_2 + \alpha(\text{difference})f_2 = 2 + (0.5)(4)(-1) = 0$$