تمرین سوم مبانی و کاربردهای هوش مصنوعی

اشکان شکیبا (9931030)

سوال اول

- نادرست؛ نتیجه عمل ارتباطی با حالت قبلی ندارد و تنها حالت کنونی در آن موثر است.
- نادرست؛ در direct evaluation ارزش هر حالت پس از یک policy معین به شکل میانگین همه ارزش‌های آن حالت بیان می‌شود.
- نادرست؛ کاهش discount factor می‌تواند منجر به این شود که عامل ما هدف‌های کم ارزش‌تر و نزدیک‌تر را به هدف‌های دورتر و ارزشمندتر ترجیح دهد که این موضوع باعث تغییر policy می‌شود.
- درست؛ با توجه به عدم توجه به محیط، این روش model freeست و به دلیل عدم اهمیت به policy این روش off policyست.

سوال دوم

می‌توان توابعی تعریف کرد که با دریافت هر حالت، فاصله هلی‌کوپتر از درخت‌ها و فاصله هلی‌کوپتر از ساختمان‌ها را محاسبه کرده و به عنوان ارزش بازگردانی کند.

ممکن است دو حالت با وجود فواصل یکسان از درخت‌ها و نیز ساختمان‌ها ارزش متفاوتی داشته باشند؛ برای مثال در یکی از آن‌ها هلی‌کوپتر بین موانعی گیر کند و از ادامه حرکت بازمانَد.

سوال سوم

الف)

نه؛ مقادیر اولیه معمولا مقادیری تقریبی و بعضا تصادفی هستند که با هر بار اجرای policy evaluation یک گام به مقادیر دقیق نزدیک‌تر می‌شوند و تا پیش از همگرایی نمی‌توان تضمین کرد که به مقدار صحیح خود رسیده‌اند و دیگر دستخوش تغییرات نخواهند شد.

ب)

بله؛ بالاخره ارزش‌ها converge خواهند شد اما در مدت زمانی طولانی‌تر. به طور کلی افزایش exploration منجر به افزایش زمان رسیدن به هدف می‌شود.

سوال چهارم

الف)

می‌دانیم:

$$V_{k+1}(s) = \max_a \Sigma_{s'} T(s, a, s') [R(s, a, s') + \gamma V_k(s')]$$

بنابراین:

	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

1) $V_2(B) = 3.5$

2) $Q_2(B, right) = T(B, right, C) [R(B, right, C) + \gamma V_1(C)]$

$= 1 + 2.5 = 3.5$

3) $Q_2(B, left) = T(B, left, A) [R(B, left, A) + \gamma V_1(A)]$

$= -2 + 1 = -1$

ب)

	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	0	2	0	0.5	-1	0
Transition 2	0	3.25	0	0.75	-1.125	0
Transition 3	0	4	0	0.825	-1.1875	0
Transition 4	0	4.4125	0	0.9125	-1.1375	0

سوال پنجم

الف)

$V_1\ \pi0(High) = 1 * (0 + \gamma * 0) = 0$

$V_1\ \pi0(Low) = 0.3 * (0 + \gamma * 0) + 0.7 * (0 + \gamma * 0) = 0$

	High	Low
π_0	study	study
V_0	0	0
V_1	0	0
π_1	exam	Netflix
V_0	8	1
V_1	11.65	1.5
V_2	13.2925	1.75

تعیین policy جدید:

$\pi_1(High) = argmax(0, 8, 1) = exam$

$\pi_1(Low) = argmax(0, -9, 1) = Netflix$

در انتها سیاست ۱ همگرا می‌شود. سیاست ۲ با سه حرکت همگرا نمی‌شود اما با توجه به اینکه مقدار **discount factor** برابر ۰/۵ است، پس از تعداد محدودی **evaluation** همگرا خواهد شد.

سوال ششم

الف)

1) $Q(C, Stop) = 0.5 * 0 + 0.5 * (0 + 0) = 0$

2) $Q(C, Go) = 0.5 * 0 + 0.5 * (2 + 0) = 1$

ب)

اولین بروزرسانی:

$Q(A, Go) = w_1 * 1 + w_2 * 1 = 0$

$r + max(Q(B, a)) = 4$

$diff = 4 - 0 = 4$

$w_1 = w_1 + 0.5 * 4 * 1 = 0 + 2 = 2$

$w_2 = w_2 + 0.5 * 4 * 1 = 0 + 2 = 2$

دومین بروزرسانی:

$Q(B, Stop) = w_1 * 1 + w_2 * (-1) = 0$

$r + max(Q(B, a)) = 0 + 4 = 4$

$diff = 4 - 0 = 4$

$w_1 = w_1 + 0.5 * 4 * 1 = 2 + 2 = 4$

$w_2 = w_2 + 0.5 * 4 * (-1) = 2 - 2 = 2$