دانشگاه صنعتی امیرکبیر

دانشکده مهندسی کامپیوتر

«مبانی و کاربردهای هوش مصنوعی ترم بهار ۱۴۰۵»

تمرین چهارم

فرایند تصمیم‌گیری مارکوف و یادگیری تقویتی

در انجام تمرین‌ها به نکات زیر توجه فرمائید:

۱- مطابق قوانین دانشگاه هر نوع کپی‌برداری و اشتراک کار دانشجویان غیرمجاز است و پاسخ به تمرین‌ها باید به صورت انفرادی انجام شود.

۲- پاسخ خود را باید در قالب یک فایل PDF به صورت تایپ‌شده و یا دستنویس (مرتب و خوانا) در سامانه کورسز آپلود نمائید؛ توجه کنید که مشکل ناخوانا بودن پاسخ‌ها در هنگام تصحیح، بر عهده دانشجو خواهد بود.

۳- در صورت هر گونه سوال و یا ابهام از طریق ایمیل زیر می‌توانید در ارتباط باشید:

h.masroor@aut.ac.ir

۴- توجه نمائید پاسخ تمرین‌ها تنها در صورت آپلود در سامانه کورسز پذیرفته خواهد شد و ارسال پاسخ از طریق ایمیل و یا روش‌های دیگر بررسی نخواهد شد.

۵- فایل پاسخ‌نامه خود را با فرمت StudentID_HW04.pdf تا ساعت ۲۳:۵۹ روز ۱۴۰۵/۰۳/۲۳ فقط در بخش مربوطه در سایت درس آپلود نمائید.

سوال اول

خرگوشی در یک جدول مطابق با شکل زیر قرار دارد. در خانه‌های شماره ۱ تا ۴ او ۲ حرکت مجاز دارد. می‌تواند به راست برود(right) یا از بازی خارج شود(exit). اگر حرکت راست را انتخاب کند به احتمال p به راست می‌رود و هویج خانه بعد را می‌خورد(یعنی پاداش می‌گیرد) و به احتمال 1-p از جدول خارج می‌شود و امتیازی نمی‌گیرد. در خانه پنجم تنها حرکت مجاز خارج شدن از جدول است که با انتخاب آن، خرگوش قطعا از جدول خارج می‌شود. با خوردن هر هویج، خرگوش ۴ امتیاز بدست می‌آورد. در صورت انتخاب exit به اختیار خود خرگوش، خرگوش ۸ امتیاز کسب می‌کند. خرگوش حرکت خود را از خانه یک آغاز می‌کند. ضریب تخفیف برابر با γ است.

سیاست‌های زیر را درنظر بگیرید:

$$\pi_0(s) = exit \quad for\ all\ s$$

$$\pi_1(s) = right \quad if\ s \leq 2, exit \quad if\ s > 2$$

$$\pi_2(s) = right \quad if\ s \leq 3, exit \quad if\ s > 3$$

الف) گراف مسئله MDP را رسم کنید.

ب) فرض کنید γ برابر با 1 است. p را به گونه‌ای تعیین کنید که دو مقدار زیر برابر شوند. محاسبات خود را نشان دهید.

$$V^{\pi_1}(2), V^{\pi_2}(2)$$

از مقدار p در قسمت قبل استفاده کنید. در صورت اینکه دو مقدار بدست آوردید، مقدار بیشتر را انتخاب کنید.

ج) بازه‌ای برای γ بیابید که $V^{\pi_0}(1) > V^{\pi_1}(1)$ محاسبات خود را بنویسید.

د) بازه‌ای برای $V^{\pi_1}(1) > V^{\pi_2}(1)$ محاسبات خود را بنویسید.

ه) فرض کنید تعداد خانه‌ها از سمت راست نامحدود است و در هر خانه هویج قرار دارد. اگر $p = 1$ باشد، ضریب تخفیف چقدر باشد تا خرگوش ترجیح دهد در خانه اول از جدول خارج شود؟ اگر مقداری پیدا کردید آن را بنویسید در غیر این صورت توضیح دهید چه جریمه‌ای برای خرگوش در نظر گرفته شود تا در هر حالت، خروج ارجحیت داشته باشد.

سوال دوم

الف) policy iteration چه مزیتی نسبت به value iteration دارد؟

ب) در یک MDP محدود اگر هر پاداش در عددی مثبت ضرب شود، سیاست بهینه تغییر می‌کند؟ توضیح دهید.

ج) تابع اکتشاف زیر را در نظر بگیرید. (در اسلاید درس موجود است.)

$$f(u,n) = u + k/n$$

این تابع چگونه بر جست‌وجوی محیط توسط عامل تاثیر می‌گذارد؟ نقش هر پارامتر در این تابع را توضیح دهید.

د) اگر عاملی از سیاست اکتشاف تصادفی استفاده کند، در صورتی که به سیاست بهینه برسد، آیا عملکرد بهینه دارد ؟ پاسخ خود را با توجه به مفهوم regret توضیح دهید.

ه) دو خودروی خودران تصور کنید. یکی از q-learning سنتی استفاده می‌کند(فرض کنید تعداد حالت ها به گونه ای است که این قابلیت را دارد) و یکی از q-learning تقریبی. بار اول هر دو در یک خیابان آسفالت شده شهری آموزش می بینند. بار دوم در یک جاده خاکی صعب العبور. در صورتی که آن دو برای تست به خیابان آسفالت شده شهری برگردند، چه تفاوتی میان عملکرد آن دو است؟

سوال سوم

یک عامل در یک MDP ناشناخته قرار دارد. این MDP شامل سه وضعیت روبه‌رو است: A, B, C

عامل در هر وضعیت می‌تواند یکی از دو اقدام روبه‌رو را انتخاب کند: $\leftarrow, \rightarrow$

عامل هنگام تعامل با محیط، چند تجربه جمع‌آوری کرده است. هر تجربه به‌صورت چهارتایی زیر نمایش داده می‌شود:

$$(s, a, s', r)$$

که در آن: s: وضعیت فعلی، a: اقدام انتخاب‌شده، s': وضعیت بعدی و r: پاداش دریافت‌شده

داده‌های مشاهده‌شده به‌صورت زیر هستند:

وضعیت فعلی(s)	اقدام(a)	وضعیت بعدی(s')	پاداش(r)
A	(→)	B	2
C	(←)	B	2
B	(→)	C	-2
A	(→)	B	4

فرض کنید: $\gamma = 1$

و همه‌ی مقدارهای اولیه‌ی Q برابر صفر هستند: $Q(s, a) = 0$

همچنین نرخ یادگیری برابر است با: $\alpha = \frac{1}{2}$

رابطه‌ی به‌روزرسانی Q-learning به شکل زیر است:

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

الف) تجربه‌ها را به همان ترتیبی که در جدول آمده‌اند پردازش کنید و مقدارهای زیر را پس از پایان همه‌ی به‌روزرسانی‌ها به‌دست آورید:

$Q(A, \rightarrow)$ $Q(B, \rightarrow)$ $Q(C, \leftarrow)$ $Q(B, \leftarrow)$

در محاسبات خود نشان دهید که هر مقدار Q چگونه تغییر می‌کند.

ب) پس از انجام یادگیری Q، سیاست greedy زیر ساخته می‌شود: $\pi_Q(s) = \arg\max_a Q(s, a)$

برای وضعیت‌های A، B، و C، اقدام انتخاب‌شده توسط این سیاست را مشخص کنید.

اگر در یک وضعیت بین دو اقدام تساوی وجود داشت، آن را صریحاً مشخص کنید و بنویسید سیاست در آن وضعیت یکتا نیست.

پ) در یک MDP گسسته‌ی دلخواه، فرض کنید عامل به‌اندازه کافی محیط را اکتشاف کرده باشد. کدام‌یک از روش‌های زیر در حالت همگرایی، اطلاعات کافی برای استخراج یک سیاست بهینه فراهم می‌کنند؟ برای هر مورد، پاسخ دهید «کافی است» یا «کافی نیست» و دلیل کوتاه بیاورید.

۱. یادگیری مدل‌مبنا برای تخمین T(s, a, s') و R(s, a, s')
۲. ارزیابی مستقیم برای تخمین V(s)
۳. یادگیری TD برای تخمین V(s)
۴. Q-learning برای تخمین Q(s, a)

ت) در حد بی‌نهایت گام زمانی، تحت کدام سیاست‌های اکتشافی زیر، Q-learning تضمین دارد که به مقدارهای بهینه‌ی Q همگرا شود؟

فرض کنید:

- نرخ یادگیری مناسب انتخاب شده است.
- در MDP از هر وضعیت، با احتمال غیرصفر می‌توان به وضعیت‌های دیگر رسید.

گزینه‌ها:

۱. سیاست ثابت که در هر وضعیت، اقدام‌ها را یکنواخت تصادفی انتخاب می‌کند.
۲. سیاست greedy.
۳. سیاست greedy-ε.
۴. سیاست ثابت بهینه.

برای هر مورد دلیل آورده و توضیح دهید چرا شرط بازدید بی‌نهایت از همه‌ی زوج‌های وضعیت–اقدام مهم‌تر است.

سوال چهارم

در این سؤال، می‌خواهیم برای Pacman از Q-learning استفاده کنیم. اما فضای وضعیت‌ها بسیار بزرگ است و نگهداری جدول کامل Q برای همه‌ی وضعیت‌ها و اقدام‌ها عملی نیست. بنابراین از نمایش مبتنی بر ویژگی استفاده می‌کنیم.

در هر وضعیت s، Pacman می‌تواند یکی از اقدام‌های روبه‌رو را انجام دهد: $left, right, up, stay$

برای هر زوج (s, a)، دو ویژگی تعریف می‌کنیم:

$$F_p(s, a) = D(s) + 2E(s, a) \qquad F_g(s, a) = A(s) + B(s, a) + C(s, a)$$

که در آن:

- $A(s)$: تعداد روح‌هایی که در فاصله‌ی حداکثر یک حرکت از وضعیت فعلی Pacman هستند .
- $B(s, a)$: تعداد روح‌هایی که Pacman پس از انجام اقدام a مستقیماً با آن‌ها برخورد می‌کند .
- $C(s, a)$: تعداد روح‌هایی که پس از انجام اقدام a، در فاصله‌ی حداکثر یک حرکت از وضعیت جدید Pacman قرار دارند .
- $D(s)$: تعداد غذاهایی که در فاصله‌ی حداکثر یک حرکت از وضعیت فعلی Pacman هستند .
- $E(s, a)$: تعداد غذاهایی که Pacman پس از انجام اقدام a می‌خورد .

در وضعیت فعلی، مقدارهای زیر داده شده‌اند:

$$D(s) = 1 \qquad A(s) = 2$$

و مقدارهای وابسته به اقدام‌ها در جدول زیر آمده‌اند:

اقدام a	$B(s, a)$	$C(s, a)$	$E(s, a)$
up	0	0	1
left	1	1	0
right	1	1	0
stay	0	2	0

الف) برای هر اقدام، مقدار ویژگی‌های زیر را محاسبه کنید:

$$F_g(s, a)$$
$$F_p(s, a)$$

یعنی مقدارهای زیر را به‌دست آورید:

$$F_g(s, up), F_p(s, up)$$
$$F_g(s, left), F_p(s, left)$$
$$F_g(s, right), F_p(s, right)$$
$$F_g(s, stay), F_p(s, stay)$$

ب) فرض کنید پس از چند اپیزود یادگیری، وزن ویژگی‌ها به‌صورت زیر شده‌اند:

$$w_g = -10$$
$$w_p = 100$$

مقدار Q برای هر اقدام با رابطه‌ی روبه‌رو محاسبه می‌شود: $Q(s, a) = w_g F_g(s, a) + w_p F_p(s, a)$

مقدارهای زیر را محاسبه کنید:

$$Q(s, up)$$
$$Q(s, left)$$
$$Q(s, right)$$
$$Q(s, stay)$$

سپس مشخص کنید اگر Pacman سیاست greedy داشته باشد، کدام اقدام را انتخاب می‌کند.

پ) فرض کنید Pacman در وضعیت s، اقدام up را انجام می‌دهد، به وضعیت جدید s' می‌رسد و پاداش روبه‌رو را دریافت می‌کند: $R(s, up, s') = 250$

در وضعیت جدید s'، فقط دو اقدام ممکن وجود دارد: $down, stay$

ویژگی‌های این دو اقدام در وضعیت s' به‌صورت زیر هستند:

اقدام در s'	$F_g(s', a')$	$F_p(s', a')$
down	2	0
stay	0	0

با استفاده از وزن‌های قبلی: $w_g = -10,\ w_p = 100$

ابتدا مقدارهای زیر را محاسبه کنید:

$$Q(s', down)$$

$$Q(s', stay)$$

سپس با فرض ضریب تخفیف $\gamma = 0.5$ هدف Q-learning برای تجربه‌ی مشاهده‌شده را محاسبه کنید:

$$Q_{\text{target}}(s, up) = R(s, up, s') + \gamma \max_{a'} Q(s', a')$$

ت) **(امتیازی)** اکنون با استفاده از هدف محاسبه‌شده در قسمت قبل، وزن‌ها را به‌روزرسانی کنید.

فرض کنید نرخ یادگیری برابر است با: $\alpha = 0.5$

و به‌روزرسانی وزن‌ها به شکل روبه‌رو انجام می‌شود: $w_i \leftarrow w_i + \alpha\left[Q_{\text{target}}(s, a) - Q(s, a)\right]F_i(s, a)$

وزن‌های جدید زیر را محاسبه کنید:

$$w_g^{new}$$

$$w_p^{new}$$

توجه کنید که در این قسمت، اقدام انجام‌شده up بوده است؛ بنابراین باید از ویژگی‌های $F_g(s, up)$ و $F_p(s, up)$ استفاده کنید.

سوال پنجم

در این مسئله یک ماشین در دره‌ای بین دو تپه قرار دارد. هدف ماشین این است که با انتخاب اقدام‌های مناسب، به بالای تپه‌ی سمت راست برسد.

وضعیت ماشین در هر لحظه با دو مقدار مشخص می‌شود: $s = [position, velocity]$

اقدام‌های ممکن عبارت‌اند از:

اقدام	معنا
0	شتاب به چپ
1	بدون شتاب
2	شتاب به راست

در هر گام زمانی، اگر ماشین به هدف نرسیده باشد، پاداش روبه‌رو را دریافت می‌کند: $r = -1$

اگر ماشین به هدف برسد، اپیزود پایان می‌یابد.

الف) فرض کنید ماشین در وضعیت s قرار دارد و مقدارهای فعلی Q برای این وضعیت به‌صورت زیر هستند:

اقدام	$Q(s, a)$
0	-8
1	-6
2	-7

عامل اقدام $a = 1$ را انتخاب می‌کند و پس از انجام این اقدام، پاداش $r = -1$ را دریافت می‌کند و به وضعیت جدید s' منتقل می‌شود. مقدارهای Q در وضعیت جدید به‌صورت زیر هستند:

اقدام	$Q(s', a')$
0	-5
1	-4
2	-6

فرض کنید: $\alpha = 0.4$ و $\gamma = 0.9$

با استفاده از رابطه‌ی Q-learning، مقدار جدید $Q(s, 1)$ را محاسبه کنید:

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

ب) با توجه به جدول اولیه‌ی Q در وضعیت s:

اقدام	$Q(s, a)$
0	-8
1	-6
2	-7

اگر عامل از سیاست greedy استفاده کند، کدام اقدام را انتخاب می‌کند؟ دلیل پاسخ خود را کوتاه توضیح دهید.

پ) فرض کنید عامل در همان وضعیت s از سیاست $greedy - \epsilon$ استفاده می‌کند و مقدار ϵ برابر است با:

$$\epsilon = 0.2$$

در این سیاست:

- با احتمال $1 - \epsilon$، عامل بهترین اقدام فعلی را انتخاب می‌کند .
- با احتمال ϵ، عامل یکی از اقدام‌های ممکن را به‌صورت تصادفی و یکنواخت انتخاب می‌کند .

احتمال انتخاب هر یک از اقدام‌های زیر را محاسبه کنید:

$$P(a = 0)$$

$$P(a = 1)$$

$$P(a = 2)$$

توجه کنید که اقدام greedy نیز در بخش انتخاب تصادفی می‌تواند دوباره انتخاب شود.

ت) فرض کنید اقدام $a = 2$ در ابتدا مقدار Q کمتری نسبت به اقدام $a = 1$ دارد، اما در واقع اگر چندین بار امتحان شود، در بلندمدت می‌تواند ماشین را سریع‌تر به هدف برساند.

توضیح دهید چرا یک سیاست کاملاً greedy ممکن است هیچ‌وقت این موضوع را کشف نکند.

همچنین توضیح دهید چرا سیاست $greedy - \epsilon$ می‌تواند در این وضعیت عملکرد بهتری داشته باشد.

ث) در محیط‌های واقعی، بعضی اقدام‌ها ممکن است خطرناک باشند. در این مسئله فرض کنید اگر سرعت ماشین در گام بعدی از یک حد مشخص بیشتر شود، آن اقدام ناایمن محسوب می‌شود. سرعت گام بعدی با رابطه‌ی زیر محاسبه می‌شود:

$$velocity_{t+1} = velocity_t + (action - 1) \times force - \cos(3 \times position_t) \times gravity$$

فرض کنید:

$$position_t = 0.4$$
$$velocity_t = 0.052$$
$$force = 0.001$$
$$gravity = 0.0025$$

و آستانه‌ی سرعت ایمن برابر است با: $velocity_{threshold} = 0.055$

برای هر اقدام $a \in \{0, 1, 2\}$، مقدار $velocity_{t+1}$ را محاسبه کنید و مشخص کنید آن اقدام ایمن است یا ناایمن.

یک اقدام زمانی ایمن است که: $velocity_{t+1} < velocity_{threshold}$

سوال ششم

یک MDP با n حالت فرض کنید. این N حالت از شماره یک تا n نام‌گذاری شده اند. در حالت‌های یک تا n-1 دو حرکت مجاز است. اولی right است که عامل را از حالت n به حالت n+1 منتقل می‌کند و پاداش یک را برای عامل به همرا دارد. دومی reset است که باعث می‌شود عامل بدون هیچ پاداشی به حالت اول برگردد. در حالت N تنها یک حرکت مجاز است که آن stay است و پاداش 10 را به همراه دارد. توجه کنید در حالت آخر عامل از محیط خارج نشده و همچنان به حرکت ادامه می‌دهد. فرض کنید ضریب تخفیف 0.2 است.

الف) سیاست بهینه در هر حالت چیست ؟

ب) $V^*(N)$ را محاسبه کنید.

ج) برای حالت دلخواه n به غیر از حالت N ام، مقدار $V^*(n)$ را محاسبه کنید.

د) **(امتیازی)** فرض کنید از value iteration استفاده می‌کنید. مقادیر اولیه صفر هستند. مقادیر همه حالت‌ها را بعد از تکرار اول، تکرار دوم و تکرار سوم بدست بیاورید.

ه) اگر پاداش reset منفی باشد، سیاست بهینه که در قسمت الف بدست آوردید تغییر می‌کند؟ توضیح دهید.

و) اگر پاداش reset مثبت باشد، امکان دارد که سیاست بهینه تغییر کند ؟ اگر جوابتان مثبت است بیشترین مقدار پاداش reset که سیاست بهینه شما در قسمت الف را تغییر ندهد چیست ؟ محاسبات خود را بنویسید.