به نام خدا

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

مبانی و کاربردهای هوش مصنوعی ترم پاییز ۱۴۰۱

تمرین سوم

مهلت تحویل ۱۶ دی ساعت ۲۳:۵۹

سوال ۱ (۲۰ نمره)

صحیح یا غلط بودن موارد زیر را با ذکر دلیل بیان کنید.

۱- در فرایند تصمیم گیری مارکوف نتیجه هر عمل به حالت کنونی و قبلی وابسته است.

۲- در direct evaluation بعد از انجام هر انتقال، ارزش آن حالت را دوباره محاسبه می‌کنیم.

۳- اگر تنها تفاوت بین دو MDP، مقدار discount factor باشد، این دو قطعا سیاست بهینه (optimal policy) یکسان دارند.

۴- Q-learning یک شیوه یادگیری تقویتی model-free و off-policy است.

سوال ۲) (۱۰نمره)

فرض کنید میخواهیم از روش یادگیری Q تخمینی در یک هلیکوپتر کوچک برای خودداری از برخورد با درختان و ساختمان ها استفاده کنیم. ابتدا مشخص کنید از کدام یک از ویژگی های محیط را برای تابع ارزش خطی استفاده می کنید و سپس دو حالتی که بر اساس ویژگی های گفته شده مشابه هستند اما ارزش بسیار متفاوتی دارند را بیان کنید.

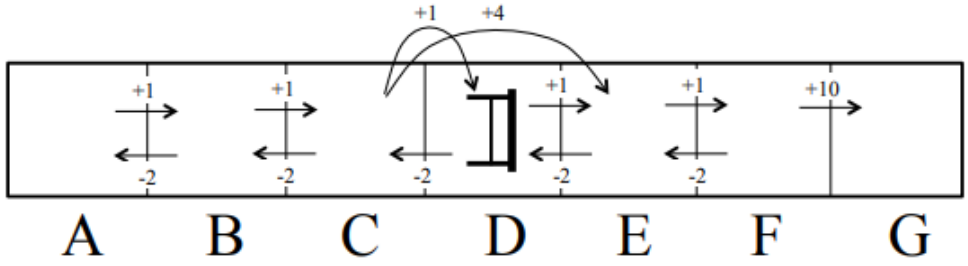
سوال۳) (۱۰ نمره)

الف) اگر در طول اجرای policy iteration، فقط یک iteration از policy evaluation را به جای اجرای آن تا زمان همگرایی انجام دهیم، آیا همچنان به سیاست بهینه می‌رسیم؟ توضیح دهید.

ب) فرض کنید در Q-learning، مقدار ε=1 باشد. در این صورت آیا تضمینی وجود دارد که به سیاست بهینه همگرا شود؟ توضیح دهید.

سوال۴) (۲۰ نمره)

مدل MDP شکل زیر را در نظر بگیرید. در خانه D یک مانع وجود دارد و عامل ما می‌تواند به صورت deterministic به چپ (left) یا راست (right) حرکت کند. خانه G نیز حالت ترمینال است. اگر عامل در خانه C باشد، نمی‌تواند به راست حرکت کند و باید بپرد (jump). پرش ممکن است موفقیت‌آمیز باشد و به خانه E برسد و T(C, jump, E) = 0.5. همچنین ممکن است شکست خورده و به مانع خانه D برخورد کند و T(C, jump, D) = 0.5. به سوالات زیر با فرض γ = 1 پاسخ دهید.



الف) دو مرحله value iteration را انجام دهید و مقادیر زیر را محاسبه کرده و جدول را کامل کنید. Iteration صفر برای تخصیص مقدار اولیه صفر برای همه مقادیر است.

1) $V_2(B)$

2) $Q_2(B, right)$

3) $Q_2(B, left)$

	A	B	C	D	E	F	G
V_1							
V_2							

ب) با توجه به اپیزود زیر و چهار انتقال انجام شده، به روز رسانی Q-learning را اعمال کنید و Q-valueهای به دست آمده را در خانه‌های جدول پر کنید. خانه‌هایی که تحت تاثیر قرار نمی‌گیرند را خالی بگذارید.

Episode

s	a	r	s	a	r	s	a	r	s	a	r	s
C	*jump*	+4	E	*right*	+1	F	*left*	-2	E	*right*	+1	F

	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						
Transition 2						
Transition 3						
Transition 4						

سوال۵) (۲۰ نمره)

فرض کنید یک عامل هوشمند داریم که میتواند درس بخواند! به جهت سادگی، این عامل فقط دارای حالت وضعیت درسی {High, Low} میباشد. در هر کدام از این حالت ها، عامل میتواند یکی از عمل های زیر را انجام دهد:

- درس بخواند
- امتحان بدهد
- به تماشای Netflix بنشیند.

این عامل را می توانیم با MDP زیر مدل کنیم:

S	A	S'	T(S, A, S')	R (S, A, S')
High	study	High	1.0	0
High	study	Low	0.0	0
High	exam	High	0.9	10
High	exam	Low	0.1	-10
High	Netflix	High	1.0	1
High	Netflix	Low	0.0	1
Low	study	High	0.3	0
Low	study	Low	0.7	0
Low	exam	High	0.05	10
Low	exam	Low	0.95	-10
Low	Netflix	High	0.0	1
Low	Netflix	Low	1.0	1

Policy iteration را برای این MDP تا دو iteration اعمال کنید. سیاست اولیه را عمل study برای هر حالت در نظر بگیرید. Utility اولیه هر دو حالت را برابر با صفر در نظر بگیرید. مقدار $\gamma = 0.5$ است. راه حل خود را شرح دهید. آیا در انتها سیاست ها همگرا می‌شوند؟ توضیح دهید.

سوال ۶) (۲۰ نمره)

یک مسئله MDP را در نظر بگیرید که در آن سه حالت [A,B,C] و دو اکشن حرکت (Go) و توقف (Stop) وجود دارد. نمونه‌های جدول زیر از انجام اکشن‌های مختلف در این MDP تولید شده‌اند. همچنین فرض کنید $\gamma = 1$ و $\alpha = 0.5$. به سوالات زیر پاسخ دهید.

الف) Q-learning را روی نمونه‌های زیر اجرا کنید. با فرض مقدار اولیه Q-value صفر، مقادیر Q-value زیر که از Q-learning به دست آمده‌اند را محاسبه کنید.

s	a	s'	r
A	Go	B	2
C	Stop	A	0
B	Stop	A	-2
B	Go	C	-6
C	Go	A	2
A	Go	A	-2

1) Q (C, Stop)

2) Q (C, Go)

ب) در این بخش از دو ویژگی داده شده زیر استفاده کنید.

- $f_1(s,a) = 1$
- $f_2(s,a) = \begin{cases} 1 & a = \text{Go} \\ -1 & a = \text{Stop} \end{cases}$

وزن‌های w_1 و w_2 را با توجه به نمونه‌های مشاهده شده پس از اولین به روزرسانی (با استفاده از نمونه اول) و دومین بروزرسانی (با استفاده از نمونه دوم) بنویسید. وزن‌های اولیه صفر هستند.

s	a	s'	r
A	Go	B	4
B	Stop	A	0

توضیحات تکمیلی(توجه کنید این تمرین سوال امتیازی ندارد)

- پاسخ به تمرین ها باید به صورت فردی انجام شود. در صورت مشاهده تقلب، برای همه‌ی افراد نمره صفر لحاظ خواهد شد.
- پاسخ خود را در قالب یک فایل PDF بصورت تایپ شده یا دست نویس(مرتب و خوانا) در سامانه کورسز آپلود کنید.
- فرمت نامگذاری تمرین باید مانند AI_HW3_9931099.pdf باشد.
- در صورت هرگونه سوال یا ابهام از طریق ایمیل ai.aut.fall2022@gmail.com با تدریسیاران در تماس باشید، همچنین خواهشمند است در متن ایمیل به شماره دانشجویی خود اشاره کنید.
- همچنین می‌توانید از طریق تلگرام نیز با آیدی‌های زیر در تماس باشید و سوالاتتان را مطرح کنید:
    - @theycallmenami
    - @Elahehere
    - @ali_nrb
- ددلاین این تمرین **۱۶ دی ۱۴۰۱ ساعت ۲۳:۵۹** است و امکان ارسال با تاخیر وجود ندارد،بنابراین بهتر است انجام تکلیف را به روز های پایانی موکول نکنید.