به نام خدا

دانشکده مهندسی کامپیوتر

مبانی و کاربردهای هوش مصنوعی ترم پاییز ۱۴۰۲

تمرین سوم

مهلت تحویل ۸ دی ۱۴۰۲ ساعت ۲۳:۵۵

سوال اول (۱۸ نمره)

تصور کنید رباتی داریم که روی یک سطح شیب‌دار کار می‌کند و با استفاده از صفحات خورشیدی شارژ می‌شود. این ربات می‌تواند در سه موقعیت قرار گیرد: پایین، وسط، و بالای سطح شیب‌دار. اگر این ربات چرخ‌هایش را بچرخاند، به یک موقعیت بالاتر می‌رود یا اگر بالای سطح شیب‌دار باشد همانجا می‌ماند، و اگر چرخ‌هایش را نچرخاند به یک موقعیت پایین‌تر می‌رود یا اگر پایین باشد همانجا می‌ماند. هر بار چرخاندن چرخ‌ها یک واحد انرژی صرف می‌کند. اگر ربات پایین سطح شیب‌دار باشد، انرژی دریافت نمی‌کند ولی اگر وسط یا بالای آن باشد، در هر واحد زمان ۳ واحد انرژی دریافت می‌کند. این ربات می‌خواهد تا حد ممکن انرژی جمع کند.

این مسئله را در قالب یک فرایند تصمیم‌گیری مارکوف در نظر بگیرید و گراف مربوط به آن را رسم کنید. سپس با در نظر گرفتن فاکتور تخفیف ۰.۸ مسئله را با روش value iteration تا سه iteration حل کنید. سیاست بهینه را برای این مسئله توصیف کنید.

سوال دوم (۶ نمره)

درستی یا نادرستی عبارات زیر را با ذکر دلیل مشخص کنید:

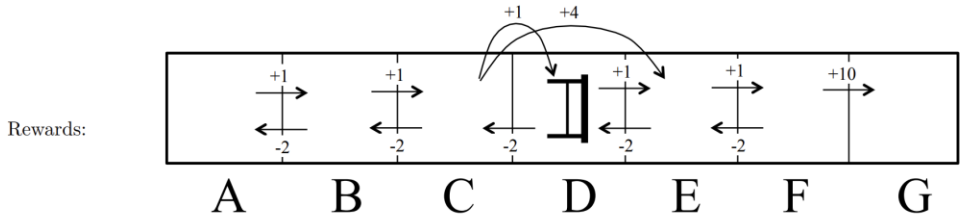
الف) اگر در فرایند q-learning تابع exploration بهینه باشد در یادگیری سیاست بهینه regret نداریم.

ب) ضریب تخفیف کوچک و مقدار بزرگ و منفی living reward رفتار حریصانه را تقویت می‌کند.

ج) در یک مسئله‌ی MDP قطعی، بروزرسانی q-learning با نرخ یادگیری ۱ به درستی مقادیر q بهینه را یاد می‌گیرد.

سوال سوم (۱۸ نمره)

یک MDP را در نظر بگیرید که که یک مسیر دوی از روی موانع را مطابق شکل زیر نشان می‌دهد. یک مانع در مربع D وجود دارد و وضعیت پایانی در مربع G است. عامل می‌تواند به سمت چپ یا راست بدود. اگر عامل در مربع C باشد، نمی‌تواند به سمت راست بدود ولی بجای آن می‌تواند بپرد، که این عمل ممکن است منجر به سقوط به مربع مانع (D) یا پرش موفقیت‌آمیز بر روی مربع E شود. پاداش‌ها در زیر نمایش داده شده‌اند. تخفیف را با مقدار $\gamma = 1$ فرض کنید.



اکشن‌ها:

راست: به طور قطعی به راست حرکت میکند. (در خانه C قابل اتخاذ نیست.)

چپ: به طور قطعی به چپ حرکت میکند.

پرش: به طور تصادفی به راست میپرد و فقط برای خانه C قابل اتخاذ است. احتمال موفقیت‌آمیز بودن پرش 50% است.

الف) برای سیاست π که همیشه حرکت مستقیم را پیشنهاد میدهد (همیشه راست یا پرش)، مقدار $V^{\pi}(C)$ را محاسبه کنید.

ب) دو بار value iteration را انجام دهید و مقادیر زیر را حساب کنید. پیمایش 0 معادل مقداردهی اولیه همه ارزش‌ها به صفر است.

$V_2(B)$	
$Q_2(B, right)$	
$Q_2(B, left)$	

ج) خانه‌های خالی جدول زیر را با Q-value هایی که از اعمال به‌روزرسانی Q-learning برای ۴ انتقال مشخص شده توسط اپیزود زیر بدست میایند پر کنید. میتوانید خانه‌های مربوط به Q-value هایی که از این به‌روزرسانی تغییری نمیکنند خالی بگذارید. از نرخ یادگیری α برابر 0.5 استفاده کنید و فرض کنید همه Q-value ها در ابتدا برابر صفر بودند.

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

	$Q(\mathrm{C}, \mathit{left})$	$Q(\mathrm{C}, \mathit{jump})$	$Q(\mathrm{E}, \mathit{left})$	$Q(\mathrm{E}, \mathit{right})$	$Q(\mathrm{F}, \mathit{left})$	$Q(\mathrm{F}, \mathit{right})$
Initial	0	0	0	0	0	0
Transition 1						
Transition 2						
Transition 3						
Transition 4						

سوال چهارم (۱۸ نمره)

پک‌من که به سهمیه ماراتن گیمپیک (المپیک گیم‌ها!) دست یافته است، برای آنالیز وضعیت خود در مسابقه فردا از شما کمک می‌خواهد.

پک‌من در هر زمان از مسابقه می‌تواند یکی از کارهای زیر را انجام دهد:

- بدود (Run)
- آب بنوشد (Water)
- بنشیند و نفسی چاق کند! (Rest)

همچنین از آنجایی که پک‌من به هیچ رنکی جز اول شدن راضی نیست، وضعیت خود را به دو شکل First Rank و Bad Rank توصیف می‌کند.

S	A	S’	T(S, A, S’)	R(S, A, S’)
First Rank	Run	First Rank	1	0
First Rank	Run	Bad Rank	0	0
First Rank	Water	First Rank	0.9	10
First Rank	Water	Bad Rank	0.1	-10
First Rank	Rest	First Rank	1	1
First Rank	Rest	Bad Rank	0	1
Bad Rank	Run	First Rank	0.3	0
Bad Rank	Run	Bad Rank	0.7	0
Bad Rank	Water	First Rank	0.05	10
Bad Rank	Water	Bad Rank	0.95	-10
Bad Rank	Rest	First Rank	0	1
Bad Rank	Rest	Bad Rank	1.0	1

اگر زمان مسابقه را به چند بازه زمانی که هر یک معادل یک iteration هستند تقسیم کنیم، با در نظر گرفتن اینکه پک‌من مسابقه را با دویدن شروع کند، policy iteration را تا دو iteration اعمال کنید.

مقدار utility اولیه را صفر و نرخ تخفیف را $\gamma = 0.5$ در نظر بگیرید.

آیا در انتها سیاست‌ها همگرا می‌شوند؟

سوال پنجم (۲۰ نمره)

در این سوال قرار است با هم به شهر بازی پارک ارم برویم. ☺ این شهر بازی چندین دستگاه بازی مختلف دارد که هر کدام از آنها هیجان خاص خود را القا می‌کند و ویژگی‌های خاص خودش را دارد. ما در تصمیم‌گیری‌های خود می‌توانیم یکی از این دستگاه‌ها را انتخاب کرده و سوار شویم یا می‌توانیم از شهر بازی بیرون برویم.

در ابتدا خوشحال و خندان وارد شهر بازی می‌شویم و از هر بازی، جایزه‌ی مثبت و خوبی دریافت می‌کنیم که حالمان را بهتر می‌کند. ولی ممکن است بعضی بازی‌ها به قدری ترسناک و دلهره‌آور باشند که ما را مریض و بدحال کنند. اگر با مریضی به بازی کردن ادامه دهیم، احتمال دارد دوباره خوب شویم، ولی به اندازه‌ی قبل از بازی لذت نمی‌بریم و ممکن است جوایز کمتر (یا حتی منفی) به دست آوریم.

ما قبلا به این شهر بازی سوار نشده‌ایم و نمی‌دانیم چقدر می‌شود از هر بازی لذت برد، چه خوب باشیم و چه مریض. هم‌چنین اصلا نمی‌دانیم چقدر احتمال دارد با یک بازی مریض شویم یا با بازی دیگری دوباره حالمان خوب شود. تنها چیزی که در مورد بازی‌ها می‌دانیم، اطلاعاتی کلی مانند جدول زیر است:

عمل / بازی (action)	نوع (type)	زمان (wait)	سرعت (speed)
ترن هوایی قدیمی	ترن هوایی	طولانی	تند
ترن هوایی جدید	ترن هوایی	کوتاه	کند
سقوط آزاد	سقوط آزاد	کوتاه	تند
رینجر	رینجر	کوتاه	کند
ترک شهربازی	ترک	کوتاه	کند

ما این تفریح را به عنوان یک مسئله‌ی MDP با دو حالت خوشحال و مریض فرموله می‌کنیم. در هر حالت با ترک شهر بازی، مسئله تمام می‌شود. بازی کردن می‌تواند ما را در یک حالت نگه دارد یا به حالت دیگر ببرد. همانطور که گفتیم در مورد احتمال هر کدام اطلاعاتی نداریم. ما برای پیدا کردن مقادیر Q-value از یک روش تخمین مبتنی بر ویژگی استفاده می‌کنیم که وزن‌ها و مقدار ویژگی‌های آن به شکل زیر تعریف می‌شود:

F definition	F value	Condition	W
F0(state, action)	1	این ویژگی bias بوده و همواره ۱ است	W0 = 1
F1(state, action)	1	نوع عمل ترن هوایی باشد	W1 = 2
F1(state, action)	0	نوع عمل ترن هوایی نباشد	W1 = 2
F2(state, action)	1	زمان عمل کوتاه باشد	W2 = 1
F2(state, action)	0	زمان عمل کوتاه نباشد	W2 = 1
F3(state, action)	1	سرعت عمل تند باشد	W3 = 0.5
F3(state, action)	0	سرعت عمل تند نباشد	W3 = 0.5

الف) مقدار زیر را محاسبه کنید:

Q(ترن هوایی قدیمی, خوشحال) = ?

ب) فرض کنید در حالت خوشحال هستیم و ترن هوایی قدیمی را سوار می‌شویم و در نتیجه‌ی این کار به حالت مریض می‌رویم و امتیاز -۱۰.۵ را دریافت می‌کنیم. با استفاده از این نمونه‌ی یادگیری، وزن‌های داده شده را با q-learning آپدیت کنید. نرخ یادگیری و ضریب تخفیف را هر دو برابر ۰.۵ در نظر بگیرید.

ج) در تخمین مورد استفاده آیا q-value ها در حالت شروع مریض با q-value های حالت شروع خوشحال برابرند؟ یعنی می‌توان گفت به ازای هر عمل در حالت خوشحال مقدار q با همان عمل در حالت مریض یکسان است؟ دلیل پاسخ خود را بنویسید.

حال می‌خواهیم کمی به exploratioon/exploitation tradeoff در این شهربازی بپردازیم.

د) اگر وزن‌ها را مطابق جدولی که داشتیم در نظر بگیریم، روش e-greedy در حالت خوشحال چه عملی را انتخاب می‌کند؟ اگر ممکن است چند عمل را انتخاب کند، هر عمل را به همراه احتمال انتخاب آن بنویسید.

ه) اگر برای tradeoff از exploration function زیر استفاده کنیم، توضیح دهید مقادیر u,k,n چه هستند؟ در یک جمله توضیح دهید این تابع چگونه در معادلات q-learning استفاده می‌شود؟

$$f(u,n) = u + \frac{k}{n}$$

سوال ششم (۲۰ نمره)

سیستم UC (سیستم دانشگاه‌های کالیفرنیا که شامل ۱۰ دانشگاه از جمله دانشگاه برکلی(UCB)، دانشگاه لوس انجلس (UCLA) و ...) میباشد) در حال آزمایش برای گزینه‌های حمل و نقل دانشجویان میان ده دانشگاه این سیستم است.
هر وضعیت ممکن برای یک دانشجو یک تاپل از مکان (یکی از ۱۰ دانشگاه) و رضایت دانشجو (شاد یا ناراحت) است.
اقدام‌های ممکن: اتوبوس، تاکسی، موتورسیکلت. اقدام‌ها ویژگی‌های زیر را دارند:

Action	Makes Stops	Max. Passengers	Has Wi-Fi
Bus	Yes	50	Yes
Taxi	No	4	No
Motorcycle	No	1	No

ما از یک تخمین خطی براساس ویژگی برای Q-value ها استفاده میکنیم:

Linear value function: $Q_{\mathbf{w}}(s,a) = \sum_{i=0}^{3} f_i(s,a) w_i$

Features	Initial Weights
$f_0(state,\ action) = 1$ (this is a bias feature that is always 1)	$w_0 = 1$
$f_1(state,\ action) = \begin{cases} 1 & \text{if } action \text{ makes stops} \\ 0 & \text{otherwise} \end{cases}$	$w_1 = 2.5$
$f_2(state,\ action) = \begin{cases} 1 & \text{if } action \text{ has max. passengers } > 1 \\ 0 & \text{otherwise} \end{cases}$	$w_2 = 0.5$
$f_3(state,\ action) = \begin{cases} 1 & \text{if } action \text{ has Wi-Fi} \\ 0 & \text{otherwise} \end{cases}$	$w_3 = 1$

به یاد داشته باشید که مقادیر تقریبی Q-value ها تابعی از وزن‌ها هستند، پس از قانون زنجیره‌ای مطابق معادله زیر برای به روزرسانی وزن‌ها استفاده کنید:

$$w_i \leftarrow w_i + \alpha \left[r + \gamma \max_{a'} Q_{\mathbf{w}}(s', a') - Q_{\mathbf{w}}(s, a) \right] \frac{\partial}{\partial w_i} Q_{\mathbf{w}}(s, a)$$

الف) مقادیر اولیه Q-value ها را با وزن‌های اولیه بالا بدست آورید:

Q((UCLA, upset), Bus):
Q((UCLA, upset), Taxi):
Q((UCLA, upset), Motorcycle):

ب) Q-value های اولیه برای وضعیت (UCLA, upset) به طور اتفاقی برابر Q-value های (UCB, happy) شده‌اند. آیا وقتی شما وزن‌ها را آپدیت میکنید نیز این مقادیر با هم برابر خواهند ماند؟ به عبارت دیگر، آیا به ازای هر اقدام a و بردار وزن‌های W ای، $Q_w((UCLA, upset), a) = Q_w((UCB, happy), a)$ صادق خواهد ماند؟

ج) exploration/exploitation: با Q-value ی وضعیت (UCLA, upset) که در بخش قبل حساب کردید، احتمال این که هر کدام از اقدام‌ها انتخاب شود با استفاده از epsilon-greedy exploration چه خواهد بود؟

د) با یک نمونه با وضعیت ابتدایی = (UCB, happy)، اقدام = تاکسی، وضعیت پسین = (UCLA, upset)، و جایزه = 7، با استفاده از نرخ یادگیری = 0.5 و ضریب تخفیف = 0.5 تک‌تک وزن‌ها را آپدیت کنید.

ه) یک مزیت و یک زیان در استفاده از approximate Q-learning را در مقایسه با Q-learning عادی توضیح دهید.

توضیحات تکمیلی

- پاسخ به تمرین ها باید به صورت فردی انجام شود. در صورت مشاهده تقلب، برای همه ی افراد نمره صفر لحاظ
- خواهد شد.
- پاسخ خود را در قالب یک فایل PDF به صورت تایپ شده یا دست نویس (مرتب و خوانا) در سامانه کورسز آپلود کنید.
- فرمت نامگذاری تمرین باید مانند AI_HW3_9931099 باشد.
- در صورت هر گونه سوال یا ابهام از طریق ایمیل AUT.AI.Fall2023@gmail.com با تدریس یاران در ارتباط باشید. همچنین خواهشمند است در متن ایمیل به شماره دانشجویی خود اشاره کنید.
- همچنین می توانید از طریق تلگرام نیز با آیدی های زیر در تماس باشید و سوالاتتان را مطرح کنید:
  - @SarvenazSarvghad
  - @AshkanShakiba
  - @AParsa404
- ددلاین این تمرین ۸ دی ۱۴۰۲ ساعت ۲۳:۵۵ است و امکان ارسال با تاخیر وجود ندارد، بنابراین بهتر است انجام تکلیف را به روز های پایانی موکول نکنید.