به نام خدا

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

مبانی و کاربردهای هوش مصنوعی ترم بهار ۱۴۰۱

تمرین سوم

مهلت تحویل ۲۴ اردیبهشت ۱۴۰۱

سوال ۱ (۱۰ نمره)

الف)فرایند مارکوفی را توضیح دهید و بگویید آیا میتوانیم در فرایندهای مارکوفی از روش planning استفاده کنیم؟ اگر پاسختان منفیست شرح دهید چه راه حل جایگزینی برای آن وجود دارد؟

ب) آیا بازی سودوکو یک فرایند تصمیم گیری مارکوف است؟ بیست سوالی چطور؟

پ) تفاوت های کلیدی بین الگوریتم های value iteration و policy iteration چیست و چه زمانی ممکن است یکی را به دیگری ترجیح دهید؟ پاسختان را با دلیل شرح دهید.

ت) آیا می توان همه MDP ها را با استفاده از جستجوی expectimax حل کرد؟ پاسخ خود را توجیه کنید.

سوال ۲ (۱۰ نمره)

الف) یادگیری تقویتی active و passive را با هم مقایسه کنید.

ب) آیا مشکلی در استفاده از روش Epsilon-Greedy برای یافتن سیاست بهینه وجود دارد؟ توضیح دهید.

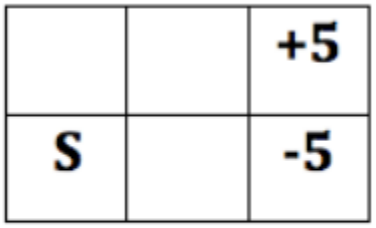
پ) مفاهیم اکتشاف (exploration) و استخراج (exploitation) را توضیح دهید. با استفاده از دو مثال از کاربردهای یادگیری تقویتی، تفاوت و trade off بین این دو را توضیح دهید.

ت) هنگام استفاده از ویژگی‌ها برای نمایش تابع Q، آیا تضمین می‌شود که یادگیری Q مبتنی بر ویژگی همان Q^* بهینه را پیدا کند که هنگام استفاده از نمایش جدولی برای تابع Q یافت می‌شود؟

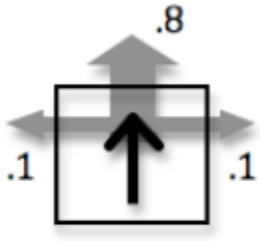
سوال ۳ (۲۰ نمره)

محیط شکل زیر(a) شامل ۶ حالت است که هر یک از حالات با شماره سطر و شماره ستون آن خانه نمایش داده می‌شوند. سطرها از پایین به بالا و ستون‌ها از چپ به راست شماره‌گذاری شده‌اند. ابتدا عامل در خانه‌ی (۱،۱) که با S نشان داده شده است قرار دارد. دو حالت هدف نهایی در میان حالات وجود دارد: (۱،۳) با پاداش -۵ و (۲،۳) با پاداش +۵ . در حالت‌های غیر نهایی، پاداش صفر است. (وقتی عامل وارد یک حالت شود، پاداش آن را دریافت می‌کند.)

اگر عامل در حالت‌های نهایی نباشد، می‌تواند ۴ جهت بالا، پایین، چپ و راست را برای حرکت خود انتخاب کند. تابع انتقال(b) بدین صورت است که به احتمال ۰.۸ حرکت انتخابی عامل انجام می‌شود و ۰.۱ احتمال دارد که عامل در جهتی عمود بر جهت انتخابی خود حرکت کند. اگر برخورد با دیوار رخ دهد، عامل در جای خود ثابت می‌ماند.



(a)



(b)

الف) سیاست بهینه را برای این محیط مشخص کنید.

ب) فرض کنید عامل، مقدار احتمالات تابع انتقال را می‌داند. همچنین فرض کنید ارزش اولیه‌ی تمام حالات صفر است. با ۲ دور اجرای value iteration با ضریب تخفیف ۰.۹ ، ارزش‌های جدید همه‌ی حالات را به دست آورید.

ج) این بار فرض کنید عامل، احتمالات تابع انتقال را نمی‌داند. در این صورت برای به دست آوردن سیاست بهینه، عامل باید چه کاری را بتواند انجام دهد؟ توضیح دهید.

د) فرض کنید همه‌ی حالات ارزش اولیه‌ی صفر دارند. به روش Temporal Difference Learning، پس از ۲ تجربه‌ی اول قسمت قبل، عامل چه بروزرسانی‌هایی در اطلاعات خود خواهد داشت ؟ (نرخ یادگیری را ۰.۱ در نظر بگیرید)

سوال ۴ (۲۰ نمره)

فرض کنید در یک بازی ریختن تاس شرکت کرده‌اید که هزینه‌ی هر بار ریختن تاس در آن ۱ سکه است و احتمال آمدن تمام اعداد در تاس با هم برابر است. شما پس از ریختن تاس به اندازه‌ی عدد روی تاس سکه دریافت می‌کنید. قانون بازی به این شکل است که شما موظف هستید در بار اول یک تاس بریزید. اما در سایر مراحل ۲ انتخاب دارید:

1. اتمام بازی: شما با این حرکت به اندازه‌ی عدد روی تاس سکه دریافت می‌کنید.
2. تاس ریختن: یک سکه هزینه می‌کنید و بار دیگر تاس می‌ریزید.

بنابراین بازی را می‌توان بدین صورت در نظر گرفت که بازیکن ابتدا در حالت شروع قرار دارد و در حالت شروع، فقط حرکت ریختن تاس وجود دارد. در سایر حالات یک حرکت اتمام بازی وجود دارد که بازیکن را به حالت پایانی می‌برد و در حالت پایانی حرکتی وجود ندارد. هر حالت بین شروع و پایان با S_i نشان داده می‌شود که بدین معناست که عدد i در پرتاب تاس آمده است. ضریب تخفیف را ۱ در نظر گرفته و به سوالات زیر پاسخ دهید:

الف) فرض کنید i π های زیر در ابتدا وجود دارند. ردیف دوم جدول را کامل کنید.

حالت	s_1	s_2	s_3	s_4	s_5	s_6
π_i	تاس ریختن	تاس ریختن	اتمام بازی	اتمام بازی	اتمام بازی	اتمام بازی
v^{π_i}						

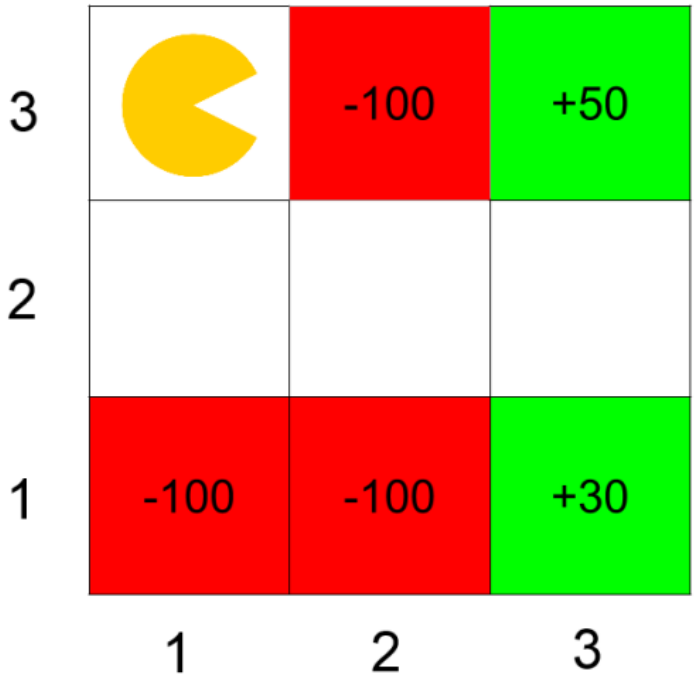
ب) با توجه به جدول فوق مقادیر i π را بروزرسانی کنید. این مقادیر می‌توانند سه حالت اتمام بازی، تاس ریختن ، و اتمام بازی/ تاس ریختن باشند.

حالت	s_1	s_2	s_3	s_4	s_5	s_6
π_i	تاس ریختن	تاس ریختن	اتمام بازی	اتمام بازی	اتمام بازی	اتمام بازی
π_{i+1}						

ج)با توجه به مقادیر جدول فوق آیا می‌توان نتیجه گرفت که مقادیر بدست آمده بهینه بوده و دیگر نیاز به بروزرسانی ندارند؟ توضیح دهید.

سوال ۵ (۲۰ نمره)

Grid-world داده شده در زیر را در نظر بگیرید. عاملی در آن قرار دارد که سعی دارد سیاست بهینه را یاد بگیرد. پاداش‌ها تنها به حالتی که در یکی از استیت‌های رنگی، اکشن Exit انجام شود، تعلق می‌گیرد. انجام این اکشن عامل را به استیت Done برده، و MDP به اتمام می‌رسد. مقادیر $\gamma = 1$ و $\alpha = 0.5$ را برای تمامی محاسبات در نظر بگیرید. در تمامی روابط باید γ و α در صورت نیاز، لحاظ شوند.



الف) عامل از قسمت بالا و سمت چپ شروع می‌کند. در ادامه، شما می‌توانید اپیزودهای حرکت عامل در grid-world را مشاهده کنید. هر خط در هر اپیزود تشکیل شده از یک تاپل شامل $(s,\ a,\ s',\ r)$ می‌باشد.

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5
(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0
(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0
(2,2), E, (3,2), 0	(2,2), S, (2,1), 0	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0
(3,2), N, (3,3), 0	(2,1), Exit, D, -100	(3,2), S, (3,1), 0	(3,2), N, (3,3), 0	(3,2), S, (3,1), 0
(3,3), Exit, D, +50		(3,1), Exit, D, +30	(3,3), Exit, D, +50	(3,1), Exit, D, +30

مقادیر Q-valueهای زیر را که از ارزیابی مستقیم نمونه‌ها به دست آمده اند را به دست آورید:

1. $Q((3,\ 2),\ N)$
2. $Q((3,\ 2),\ S)$
3. $Q((2,\ 2),\ E)$

ب) Q-learning الگوریتمی آنلاین است که مقادیر بهینه برای Q-valueها را در یک MDP با پاداش‌ها و توابع انتقال نامشخص را یاد می‌گیرد. رابطه‌ی به روز رسانی به صورت زیر می‌باشد:

$$Q(s_t, a_t) = (1 - \alpha)Q(s_t, a_t) + \alpha(r_t + \gamma \max_{a'} Q(s_{t+1}, a'))$$

که γ ضریب تخفیف، α ضریب یادگیری و دنباله‌ی مشاهدات به صورت $(\ldots, s_t, a_t, s_{t+1}, r_t, \ldots)$ می‌باشد. با توجه به اپیزودهای داده شده در قسمت الف، زمان‌هایی را که در آن‌ها برای اولین بار مقدار Q-value داده شده، غیر صفر شده است مشخص کنید. پاسخ شما باید به صورت دوتایی (episode, iter) باشد که iter در واقع Q-learning update iteration در آن اپیزود است. اگر مقدار Q-value مشخص شده هیچگاه غیر صفر نشود، عبارت هرگز را برای آن قرار دهید.

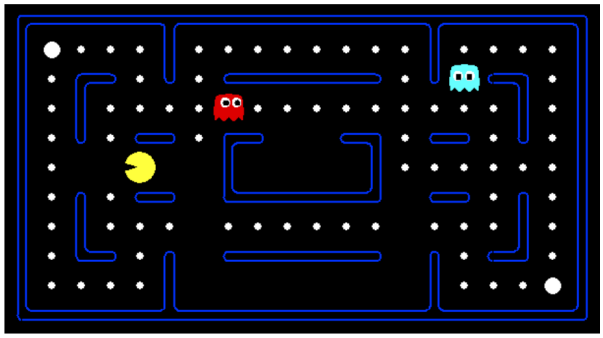
1. $Q((1, 2), E)$
2. $Q((2, 2), E)$
3. $Q((3, 2), S)$

ج) در Q-learning، ما از پنجره‌ی (s_t, a_t, s_{t+1}, r_t) جهت به روزرسانی Q-valueها استفاده می‌کنیم. می‌توان این نظر را داشت که از قانون به روز رسانی‌ای استفاده کنیم که از یک پنجره‌ی بزرگ‌تر برای آپدیت این مقادیر، بهره ببرد. قانون به روز رسانی‌ای را برای $Q(s_t, a_t)$ با استفاده از پنجره ی زیر به دست آورید:

$$(s_t, a_t, s_{t+1}, r_t, s_{t+1}, a_{t+1}, r_{t+1}, s_{t+2})$$

سوال ۶ (۲۰ نمره)

می‌خواهیم از یک عامل یادگیری Q برای پکمن استفاده کنیم، اما اندازه‌ي فضای حالات برای یک شبکه بزرگ آنقدر بزرگ است که نمی‌توانیم آن را در حافظه نگه داریم. برای حل این مشکل، به سراغ نمایش مبتنی بر ویژگی(feature-based) وضعیت پکمن می‌رویم. در اینجا یک صفحه‌ي بازی پکمن وجود دارد:



الف) برای قضاوت در مورد نتیجه مورد انتظار بازی، چه ویژگی‌هایی را از صفحه‌ی بازی پکمن استخراج می‌کنید؟

ب) فرض کنید در اینجا دو ویژگی تعداد ارواح در یک قدمی پکمن (Fg) و تعداد غذاها در یک قدمی پکمن (Fp) را در نظر داریم. برای صفحه‌ی زیر، مقادیر این ویژگی ها را استخراج کنید.

ج) با استفاده از Q-learning بعد از چند اپیزود تمرین کردن، وزن های ما شروع به مقدار گرفتن می کنند. در حال حاضر wg = 100 و wp = -10 است. مقدار Q را برای حالت بالا محاسبه کنید.

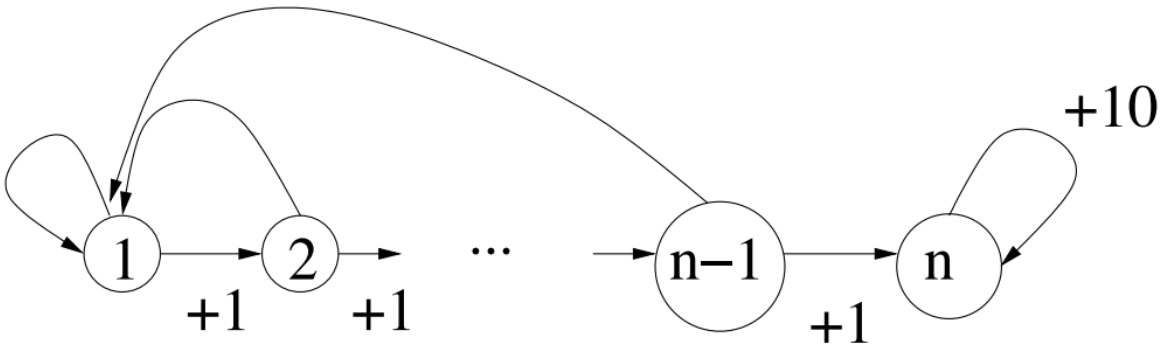
د) ما یک اپیزود دریافت می کنیم، بنابراین اکنون باید مقادیر خود را بروزرسانی کنیم. حالت شروع اپیزود حالت فوق است (که قبلاً مقادیر ویژگی و مقدار Q مورد انتظار را محاسبه کرده اید). حالت بعدی دارای مقادیر ویژگی Fg = 0 و Fp = 2 است و پاداش آن 50 است. با فرض تخفیف 0.5، تخمین جدید مقدار Q را برای این وضعیت بر اساس این اپیزود محاسبه کنید.

ه) با این تخمین جدید و نرخ یادگیری (α) برابر با 0.5، وزن ها را برای هر ویژگی به روز کنید.

ی) بیایید به کل این فرآیند یک قدم به عقب فکر کنیم. در این فرآیند (با فرض تعریف ویژگی ها) چه مواردی را به عنوان ارزش می آموزیم؟ وقتی یادگیری را به پایان رساندیم، چگونه متوجه می شویم که پکمن عملکرد درستی دارد؟

سوال ۷ (۱۰ نمره - امتیازی)

MDP با n حالت که در شکل زیر آورده شده است را در نظر بگیرید. در استیت n، تنها یک اکشن وجود دارد که پاداشی معادل 10+ را جمع‌آوری کرده و به حرکات خود پایان می‌دهد. در تمام استیت‌های دیگر، دو اکشن وجود دارد: حرکت، که به صورت قطعی یک قدم به سمت راست حرکت می‌کند، و بازنشانی؛ که به صورت قطعی عامل را به استیت 1 برمی‌گرداند. برای حرکت، میزان پاداش 1+ و برای بازنشانی هم پاداش برابر با 0 در نظر گرفته شده است. ضریب تخفیف نیز $\gamma = 0.5$ می‌باشد.



الف) سیاست بهینه در این مساله به چه صورت است؟

ب) مقدار بهینه در استیت n چه خواهد بود؟ $V^*(n)$

ج) مقدار بهینه تابع $V^*(k)$ را برای همه $k = 1, \ldots, n-1$ محاسبه کنید.

د) فرض کنید که شما دارید value iteration را برای به دست آوردن این مقادیر، انجام می‌دهید. در ابتدا، شما با تخصیص مقادیر اولیه برابر با 0، این فرایند را آغاز می‌کنید. تمامی مقادیر غیر صفر را بعد از iteration اول و دوم نشان دهید.

توضیحات تکمیلی

- 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.spring1401@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:
    - @sarvenaz_srv
    - @anotherbrickinthwall
    - @Mohadeseh_atyabi
    - @elhamrazi
- ددلاین این تمرین **۲۴ اردیبهشت ۱۴۰۰ ساعت ۲۳:۵۵** است و امکان ارسال با تاخیر وجود ندارد،بنابراین بهتر است انجام تکلیف را به روز های پایانی موکول نکنید.