

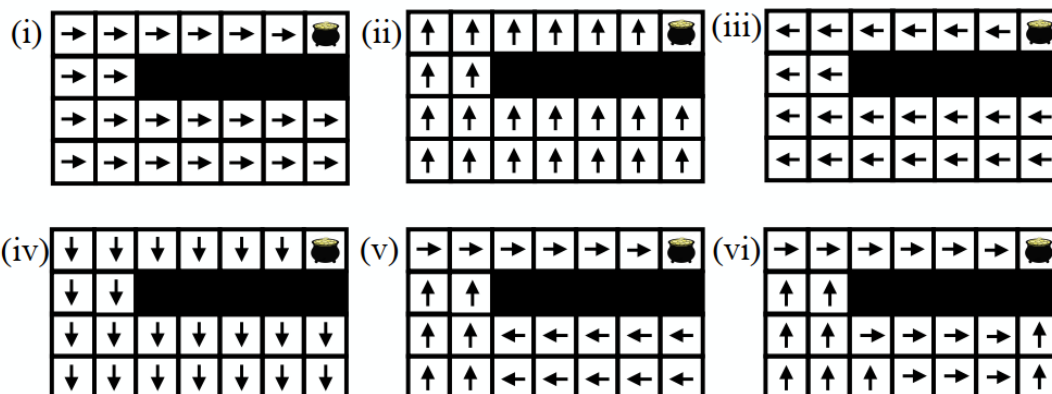
به نام خدا



مبانی و کاربردهای هوش مصنوعی ترم بهار 1402

تمرین سوم

مهلت تحویل: ۲۴ اردیبهشت 1402 ساعت 23:55



سیاست های بالا در صفحه ای شامل یک گنج را در نظر بگیرید. خانه های سیاه دیوار هستند. در این صفحه value هر خانه به شکل $V(x,y) = w^T f(x,y)$ محاسبه میشود. f برداری از ویژگی ها و w نیز برداری از وزن هاست. فرض کنید موقعیت گنج با (x_*, y_*) مشخص شده است. هر تابع f زیر کدام سیاست های بالا را نتیجه میدهد؟ (هیچ وزنی صفر نیست و می تواند مثبت یا منفی باشد)

الف) $f(x,y) = |x - x_*|$

ب) $f(x,y) = |x - x_*| + |y - y_*|$

ج) $f(y,x) =$ کوتاه ترین مسیر ممکن تا گنج

د) برای سیاست های باقی مانده تابع f را پیشنهاد دهید.

سوال 2) ۱۰ امتیاز

یک MDP با حالت های X_0 تا X_3 را در نظر بگیرید. به هر کدام از این حالت ها یک عدد نسبت داده شده است که با value آن حالت هیچ نسبتی ندارد. حرکت های مجاز در این MDP عبارتند از $\{right, left, stay\}$. عمل stay باعث میشود عامل در حالتی که هست باقی بماند، همیشه موفقیت آمیز است و پاداشی به اندازه عددی که به آن حالت نسبت داده شده است، به عامل برمی گرداند. اعمال چپ و راست در نیمی از اوقات عامل را به جهتی که قصد آن را داشتیم انتقال میدهند. در باقی اوقات عامل در حالتی که هست باقی می ماند. اگر اعمال چپ و راست موفقیت آمیز باشند (نیمی از اوقات) هیچ پاداشی به عامل تعلق نمیگیرد اما اگر موفق نباشد (در جای خود باقی بماند) مانند عمل stay به عامل پاداشی برابر با عدد نسبت داده شده تعلق میگیرد. در حالت X_3 عمل right و در حالت X_0 عمل left ممنوع می باشد. اگر $V_0(S)$ برای هر چهار حالت را برابر صفر در نظر بگیریم، value iteration را تا دو مرحله انجام دهید و راه حل خود را شرح دهید. مقدار γ را برابر 0.5 در نظر بگیرید.

6	0	5	24
X_0	X_1	X_2	X_3

سوال (3) ۱۰ امتیاز

فرض کنید در یک بازی ریختن تاس شرکت کرده اید که هزینه هر بار ریختن تاس در آن 1 سکه است و احتمال آمدن تمام اعداد در تاس با هم برابر است. شما پس از ریختن تاس به اندازه عدد روی تاس سکه دریافت می کنید. قانون بازی به این شکل است که شما موظف هستید در بار اول یک تاس بریزید. اما در سایر مراحل دو انتخاب دارید:

1- اتمام بازی (finish): با این حرکت به اندازه عدد روی تاس سکه دریافت کنید و به حالت نهایی می روید که در آن حرکتی وجود ندارد.

2- تاس ریختن (roll): یک سکه هزینه می کنید و بار دیگر تاس می ریزید.

هر حالت بین شروع و پایان با S_i نشان داده می شود که بدین معناست که عدد i در پرتاب تاس آمده است. ضریب تخفیف را برابر 1 در نظر بگیرید.

برای حل این مسئله از MDP و policy iteration استفاده می کنیم. ارزش ابتدایی همه حالت ها را برابر صفر در نظر می گیریم.

الف) در قسمت اول، از policy evaluation استفاده می کنیم. با در نظر گرفتن سیاست های ابتدایی زیر، ارزش هر حالت با توجه به سیاست گفته شده بدست آورید. (۴ نمره)

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) اکنون با بدست آوردن ارزش هر حالت، از policy improvement استفاده کنید و سیاست ها را به روز کنید.

در صورتی که در یک حالت می توان از هر دو عمل استفاده کرد، هر دوی آن ها را بنویسید. (۴ نمره)

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) آیا سیاست ها همگرا شدند؟ توضیح دهید. (۲ نمره)

سوال 4) (۲۰ نمره)

درست یا نادرستی موارد الف و ب را با ذکر دلیل مشخص کنید و به موارد ج و د پاسخ دهید.

الف) سیاست بهینه ای که در مسائل mdp به آنها دست پیدا میکنیم یکتا هستند.

ب) دو mdp با discount factor های متفاوت ممکن است سیاست های بهینه یکسان ندهند.

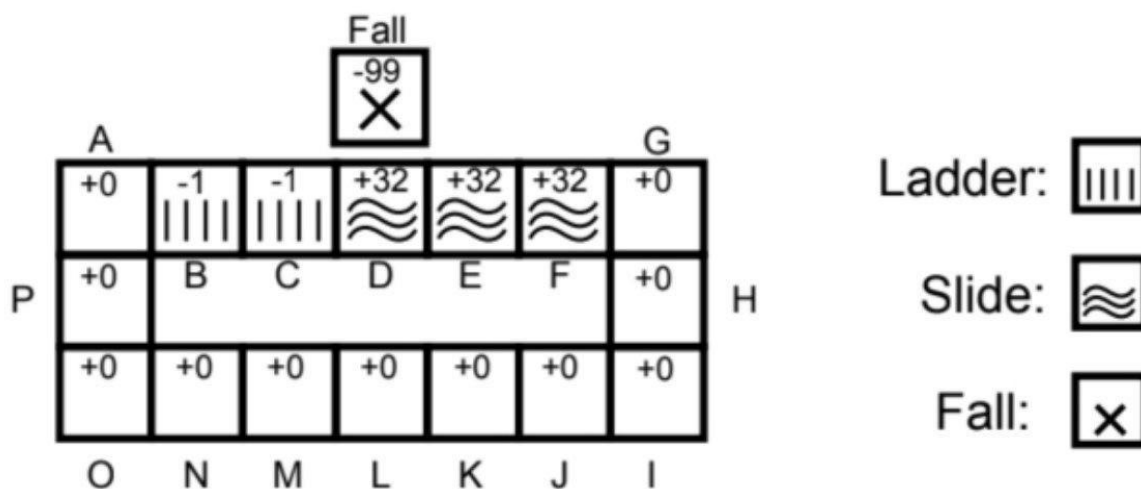
ج) هنگام یادگیری با روش epsilon-greedy action selection، ایده خوبی است که با گذشت زمان اپسیلون را به 0 کاهش داد؟ چرا؟

د) از کاربرد های یادگیری تقویتی 3 مثال زده و در هر کدام محیط و عالم را تعریف کنید و با دلیل مشخص کنید در کدام محیط استفاده از یادگیری مبتنی بر مدل و در چه محیط یادگیری بدون مدل مناسبتر است؟

سوال (5) (۳۰ نمره)

از سال‌ها پیش بر روی دریاچه هیرکانی پلی زده شده بود که روستاییان اطراف دریاچه از آن برای عبور و مرور محلی استفاده می‌کنند. این پل که از به هم پیوستن قطعات چوب تشکیل شده است، با اینکه مسیر رفت و آمد را خیلی کوتاه می‌کند اما به دلیل آن که توسط افراد محلی و بدون دانش تخصصی بنا شده، بعضی از قسمت‌های آن در حال تخریب بوده و خطرناک است.

مهسا که دانشجوی رشته کامپیوتر است و هر روز برای رفتن به دانشگاه باید از این پل عبور کند، می‌داند که عبور از این پل می‌تواند سریع‌تر و به سلامت به کلاس برسد یا یکی از چوب‌ها بلغزد و بیفتد و آسیب ببیند. هزینه افتادن از پل ۹۹- است. از آنجایی که احتمال افتادن از پل مشخص نیست، او قصد دارد با آموخته‌های خود از درس مبانی هوش مصنوعی، این مسئله را به روش یادگیری تقویتی حل کند اما در حل آن به مشکل خورده و از شما درخواست کمک کرده است.



فرض کنید مقدار گاما برای 1 است. دو مسیر زیر را برای هر رشته نمونه در نظر بگیرید.

مسیر 1: (A, East, B, -1), (B, East, C, -1), (C, East, D, +32)

مسیر 2: (A, East, B, -1), (B, East, Fall, -99)

الف) مقدار ارزش در حالت های A, B, C بعد از اعمال temporal difference learning با نرخ آلفا برابر 0.5 در مسیر 1 چقدر خواهد بود؟

ب) مقدار ارزش در هریک از حالت های A, B, C بعد از اعمال temporal difference learning با نرخ آلفا برابر 0.5 در مسیر 1 و مسیر 2 چقدر خواهد بود؟

ج) مقدار دوتایی حالت و عمل در هریک از حالت های زیر بعد از اعمال Q learning با نرخ آلفا برابر 0.5 در مسیر 1 و مسیر 2 چقدر خواهد بود؟


$Q(A, \text{South}):$

$Q(A, \text{East}):$

$Q(B, \text{East}):$

سوال 6) (۲۰ نمره)

دنیایی که در شکل زیر ارائه شده است را در نظر بگیرید، Pacman که در تلاش است سیاست بهینه را بیاموزد را در نظر بگیرید. اگر عملی منجر به ورود در یکی از خانه های رنگی شود، پاداش مربوطه در طول آن انتقال تعلق می گیرد. همه حالت های رنگی پایانی هستند. در سایر حالت ها حرکات شمال، شرق، جنوب، غرب در دسترس هستند که به طور قطعی Pacman را به خانه همسایه مربوطه منتقل می کنند (یا اگر نتیجه عمل باعث شود پکمن از جدول خارج شود، Pacman در جای خود باقی بماند). ضریب تخفیف $\gamma = 0.5$ و نرخ یادگیری $\alpha = 0.5$ را برای همه محاسبات فرض کنید. Pacman در حالت (1، 3) شروع می کند.

3		-80	+100
2			
1	+25	-100	+80
	1	2	3

الف) مقدار V^* را برای خانه های زیر بدست آورید:

$$V^*(3, 2) =$$

$$V^*(2, 2) =$$

$$V^*(1, 3) =$$

ب) جدول زیر حرکت های پکمن را در فضای بالا نشان می دهد هر خط دارای tuple شامل (s,a,s',r) است. با استفاده از Q-learning مقادیر Q-value زیر را بدست آورید.

$$Q((3,2),N) =$$

$$Q((1,2),S) =$$

$$Q((2, 2), E) =$$

$$Q((3, 2), N) =$$

Episode 1	Episode 2	Episode 3
(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
(2,2), S, (2,1), -100	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0
	(3,2), N, (3,3), +100	(3,2), S, (3,1), +80

پ) نمایشی مبتنی بر ویژگی از عملکرد Q-value را در نظر بگیرید:

$$Q_f(s, a) = w_1 f_1(s) + w_2 f_2(s) + w_3 f_3(a)$$

$F_1(S)$: حالت X مختصات

$F_2(S)$: حالت Y مختصات

$$f_3(N) = 1, f_3(S) = 2, f_3(E) = 3, f_3(W) = 4$$

1- با توجه به اینکه همه w_i در ابتدا 0 هستند ، مقدار آنها بعد از ایزود اول چیست؟

2- فرض کنید بردار وزن W برابر است با (2 ، 1 ، 1-). اقدامی که توسط Q-function انتخاب میشود زمانی که در

حالت (2 ، 1) قرار داریم چیست؟

سوال 7 (۱۰ نمره امتیازی)

فرض کنید می‌خواهیم از روش یادگیری تخمینی برای یک ماشین خودران استفاده کنیم، کدام یک از ویژگی‌های محیط برای تشکیل تابع ارزش خطی را استفاده می‌کنید؟ دو حالت که بر اساس این ویژگی‌ها مشابه هستند اما ارزش بسیار متفاوت دارند را مثال بزنید.

توضیحات تکمیلی

- پاسخ به تمرین‌ها باید به صورت انفرادی انجام شود. در صورت مشاهده تقلب برای همه افراد نمره صفر لحاظ خواهد شد.
- پاسخ خود را در قالب یک فایل PDF به صورت تایپ شده و یا دست نویس (مرتب و خوانا) در سامانه کورسز آپلود کنید.
- فرمت نام گذاری تمرین باید مانند AI_HW3_9931062 باشد.
- در صورت هرگونه سوال یا ابهام از طریق ایمیل aipring1401@gmail.com با تدریس‌یاران در ارتباط باشید. همچنین خواهشمند است در متن ایمیل به شماره دانشجویی خود نیز اشاره کنید.
- همچنین می‌توانید از طریق تلگرام نیز با آیدی‌های زیر در تماس باشید و سوالاتان را مطرح کنید:
- @Hosna_oyar
- @eeajohsehale
- @Ali_nrb
- @Nika_ST
- دلایل این تمرین 24 اردیبهشت 1402 ساعت 23:55 است. بهتر است انجام تکلیف را به روزهای پایانی موکول نکنید.