به نام خدا

هوش مصنوعی

بهار ۱۴۰۲

مدرس: محمدمهدی سمیعی

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

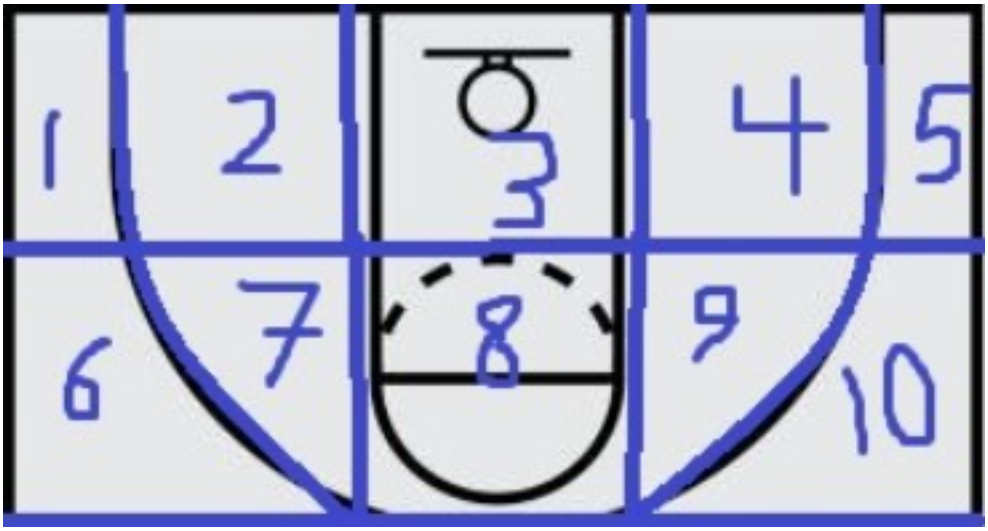
گردآورندگان: پردیس زهرایی، بنیامین ملکی، امیرحسین رازلیقی

تمرین سوم — فرایند تصمیم مارکوف و یادگیری تقویتی — مهلت ارسال: ۱۲ فروردین

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ وجود ندارد و پاسخ هایی که بعد از زمان تعیین شده ارسال شوند، پذیرفته نخواهند شد.
- هم‌کاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.
- توجه داشته باشید که تمرین نمره امتیازی ندارد.

سوالات نظری (۱۱۵ نمره)

۱. (۵۰ نمره) فرض کنید یک ربات بسکتبالیست داریم که در زمین بازی زیر قرار دارد.



شکل ۱: زمین بازی

این ربات، در هر لحظه، می‌تواند در یکی از ۱۰ خانه‌ی مشخص شده در زمین بازی حضور داشته باشد.
همچنین، در هر مرحله، ربات می‌تواند یا به یکی از خانه‌های مجاور برود و یا اینکه در همان خانه‌ای که قرار دارد بماند و اقدام به پرتاب توپ به سمت حلقه کند. منظور از خانه‌ی مجاور نیز، خانه‌ای است که یک ضلع مشترک با خانه‌ی فعلی داشته باشد. فرضا، از خانه‌ی ۹ می‌تواند به خانه‌های ۴، ۸ یا ۱۰ برود.
در صورتی که تصمیم بگیرد به یکی از خانه‌های چپ، راست، بالا و یا پایینش برود، به احتمال ۱۰۰ درصد این کار با موفقیت انجام می‌شود. از طرفی اگر action ربات ما، حرکت کردن باشد، امتیاز (reward) 1- را

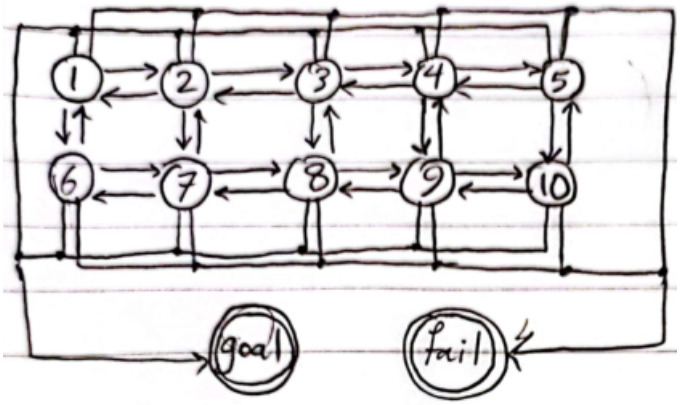
دریافت می‌کند (مستقل از وضعیتی که در آن قرار دارد). از طرفی اگر انتخاب ربات، اقدام به پرتاب (shoot) باشد، بازی با تمام می‌رسد. همچنین، با توجه به خانه‌ای که در آن قرار دارد (فرضا s)، به احتمال $P_{goal}(s)$ پرتابش گل می‌شود. متناظرا، به احتمال $1 - P_{goal}(s)$ اقدامش با شکست مواجه می‌شود.
برای وضعیت‌های مختلف، احتمالات زیر را داریم:

$$P_{goal}(s_1) = P_{goal}(s_5) = P_{goal}(s_6) = P_{goal}(s_{10}) = 0.6$$
$$P_{goal}(s_2) = P_{goal}(s_7) = P_{goal}(s_8) = P_{goal}(s_9) = P_{goal}(s_4) = 0.75$$
$$P_{goal}(s_3) = 0.9$$

در صورتی که پرتابش گل نشود، امتیاز 10- را دریافت می‌کند. اما اگر گل شود، بسته به اینکه از کجا اقدام به پرتاب کرده باشد، امتیاز متفاوتی دریافت می‌کند:

$$\forall i \in \{1, 5, 6, 10\} : R(s_i, shoot, goal) = 3$$
$$\forall i \in \{2, 3, 4, 7, 8, 9\} : R(s_i, shoot, goal) = 2$$

برای واضح‌تر شدن صورت سوال، شماتیک MDP شرح‌داده‌شده، در زیر آمده است:



شکل ۲: MDP

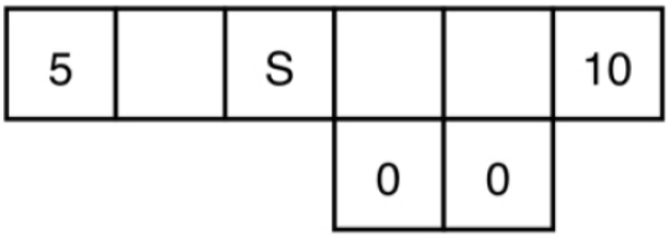
الف) به کمک روش Value Iteration، تابع ارزش وضعیت‌ها ($V(s_i)$) را برای تمامی استیت‌های MDP محاسبه کنید. توجه کنید که برای محاسبات خود، صرفا ۲ بار ایتریشن انجام دهید، به این معنا که Valueهایی که محاسبه می‌کنید، مقادیر درخت Expectimax تا عمق ۲ باشند. ($\gamma = 1$) (۳۰ نمره)
ب) حال فرض کنید که از توابع T و R اطلاعی نداریم و به جای آن، تجربه‌هایی از محیط به دست آورده ایم که در جدول ۱ (در صفحه‌ی بعد) آمده‌اند.
حال الگوریتم Q-learning را اجرا نموده و Q(s,a) را به ازای stateها و actionهایی که در تجربه‌های ما دیده‌شده‌اند و تغییر می‌کنند، محاسبه کنید. در حقیقت، در محاسبات مربوط به هر Episode، صرفا (s,a)های حاضر در آن Episode را محاسبه کنید و اگر Q(s,a) یک (s,a)ی خاص دچار تغییر نمی‌شود، نیازی به ذکر کردن مجدد مقدار آن نیست. با این فرض که همه‌ی Q(s,a)ها در ابتدا صفر باشند و با نرخ یادگیری $\alpha = 0.5$ و ضریب تخفیف $\gamma = 1$، محاسبات خود را انجام دهید. (۲۰ نمره)

۲. (۱۵ نمره) عبارات صحیح و غلط را مشخص و به صورت مختصر برای عبارات غلط دلیل بیاورید.
الف) در فرایند value iteration اگر مقدار تخفیف بین ۰ و ۱ باشد valueها حتما همگرا می‌شوند. (۳ نمره)
ب) سیاست‌های پیداشده توسط policy iteration از سیاست‌های پیداشده توسط value iteration ضعیف‌تر هستند. (۳ نمره)
پ) در Q-learning می‌توان بدون رسیدن به سیاست بهینه، به *Q بهنیه رسید. (۳ نمره)
ت) یک تخفیف کوچک‌تر از ۱ همواره می‌تواند به عنوان یک پاداش منفی در نظر گرفته شود. (۳ نمره)
ث) یک پاداش بزرگ منفی می‌تواند باعث رفتار حریصانه شود. (۳ نمره)

	S	A	S'	Reward
Episode 1	$s_۸$	move R	$s_۹$	-1
	$s_۹$	move U	$s_۴$	-1
	$s_۴$	shoot	goal	+1
Episode 2	$s_۸$	move R	$s_۷$	-1
	$s_۷$	move R	$s_۸$	-1
	$s_۸$	shoot	goal	+1
Episode 3	$s_۸$	move R	$s_۹$	-1
	$s_۹$	move R	$s_{۱۰}$	-1
	$s_{۱۰}$	shoot	goal	+2
Episode 4	$s_۸$	move R	$s_۷$	-1
	$s_۷$	move L	$s_۸$	-1
	$s_۸$	shoot	fail	-5
Episode 5	$s_۸$	move R	$s_۹$	-1
	$s_۹$	move U	$s_{۱۰}$	-1
	$s_{۱۰}$	shoot	fail	-5

جدول ۱: تجربه‌های بدست آمده از محیط

۳. (۲۵ نمره) در یک شب سرد زمستانی، شما توسط گروه مافیایی Tarasht، اسیر می‌شوید. رییس گروه، که از قضا انسان فرهیخته‌ای است، پس از بررسی لپ‌تاپ شما و پاسخ‌هایتان برای تمرین هوش مصنوعی، از توانایی شما در حل مسئله‌های قبلی خوشش می‌آید، و با طراحی چالشی به شما فرصتی می‌دهد تا فرار کنید. ابتدا در زندانی مانند شکل زیر قرار دارید:



اگر بر روی یک سلول شماره‌گذاری‌شده (با شماره مثبت) باشید، تنها اقدام موجود، خروج از آن است و زمانی که از آن خارج می‌شوید، پاداشی برابر با عدد روی سلول دریافت می‌کنید (در صورتی که از خانه با پاداش ۱۰ خارج شوید، دری به رویتان باز می‌شود که سر از خیابان تیموری در می‌آورید). در هر سلول دیگر (بدون شماره)، اقدامات موجود، حرکت به سمت شرق یا غرب است. رییس مافیا، برای اینکه کار را سخت‌تر کند، چندین چاله هم در مسیر قرار داده است (خانه‌های با شماره صفر). اگر در سلول‌های بالایی این چاله‌ها قرار داشته باشید، با حرکت کردن به شرق یا غرب، به احتمال 1-p در چاله‌ها می‌افتید و به احتمال p حرکتی که قصد داشتید انجام دهید، به درستی انجام می‌شود.
سیاست‌های مختلفی به ذهن شما می‌رسند. در هر حالت، $V_\pi(s)$ را برای هر سلول بدون شماره به‌دست آورید.

الف) همواره به سمت شرق حرکت کنید. (۵ نمره)
ب) همیشه به سمت غرب حرکت کنید. (۵ نمره)
ج) برای هر کدام از سیاست های بالا دامنه p را تعیین کنید. (۱۰ نمره)
د) در ادامه تفاوت policy iteration با value iteration را بنویسید. (۵ نمره)

۴. (۲۵ نمره) می‌دانیم که اگر بخواهیم یک تابع utility برای ارزش‌دهی به دنباله‌ای از stateها ارائه دهیم که به صورت stationary (مستقل از زمان رسیدن به state)، آن دنباله را ارزش‌گذاری کند، باید یکی از دو روش زیر باشد:

$$(۱)\ Additive : U_۱[s_۰, s_۱, s_۲, \dots] = \sum_{i=۰}^{\infty} R(s_i)$$

$$(۲)\ Discounted : U_۲[s_۰, s_۱, s_۲, \dots] = \sum_{i=۰}^{\infty} \gamma^i . R(s_i)$$

الف) دو مورد از مشکلات تابع $U_۱$ را ذکر کنید. (۵ نمره)
ب) اگر دنباله وضعیت‌های ما محدود باشد (مثلا $[s_0, s_1, \dots, s_n]$)، آیا همچنان روش‌های با stationary preference می‌توانند خوب باشند؟ (۵ نمره)
ج) فرض کنیم تمامی $R(s_i)$ها متعلق به بازه $[R_{min}, R_{max}]$ باشند. حال، برای تابع $U_۲$، کران بالا و پایینی بیابید و نشان دهید که همواره کران‌دار است. (۱۵ نمره)

سوالات عملی (۸۵ نمره)

سوالات عملی در فایل جوپیتر نوتبوک موجود هستند.

- تمرین MDP (۴۵ نمره)
- تمرین RL (۴۰ نمره)