دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی
بهار ۱۴۰۳
استاد: محمدحسین رهبان
گردآورندگان: مریم توسلی - آیلین رسته - رضا حیدری

تمرین پنجم | **یادگیری تقویتی** | مهلت ارسال: ۲۵ خرداد

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ هر تمرین تا سقف ۴ روز و در مجموع ۱۰ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر ساعت تأخیر غیر مجاز ۰.۵ درصد از نمره تمرین به صورت ساعتی کسر خواهد شد.
- تاخیر سوالات نظری و عملی با یکدیگر محاسبه می‌شوند. به عبارتی تاخیر شما در هر تمرین معادل تاخیر بیشتر بین ارسال جواب‌های تئوری و عملی است.
- 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.
- در کنار هر سوال عددی به عنوان درجه سختی قرار گرفته است. درجه سختی برای مقایسه میزان سختی و وقت‌گیری سوالات و برنامه ریزی بهتر شما برای حل سوالات قرار گرفته است. هر درجه تقریبا معادل ۵ دقیقه وقت برای حل است. البته این اعداد به هیچ وجه دقیق نیست چرا که سرعت حل افراد متفاوت است، اما می‌توانید فرض کنید که اگر سرعت عملی مشابه با درجه سختی‌های داده شده دارید، با اطمینان بالایی در امتحانات به مشکل نخواهید خورد.

سوالات نظری (۱۴۰ نمره)

۱. (۲۰ نمره، درجه سختی ۵) درستی یا نادرستی جملات زیر را با ذکر دلیل مشخص کنید.

الف) در الگوریتم Value Iteration در صورتی که، سیاست بدست آمده برای یک استیت در T مرحله تغییر نکند، سیاست بدست آمده در این T مرحله همان سیاست بهینه خواهد بود.(به طور دقیق‌تر اگر داشته باشیم $\pi_k = \pi_{k+1} = ... = \pi_{k+T}$، به ازای یک T بزرگ، آنگاه π_{k+i}به ازای هر i همان سیاست بهینه است.)

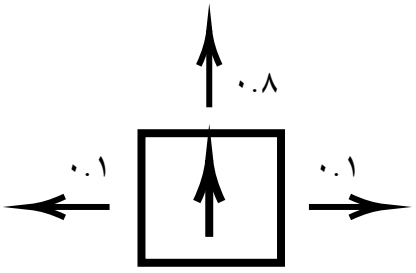
ب) Q-learning تنها زمانی مقادیر بهینه Q-values را یاد می‌گیرد که کنش‌هایی (actions) که در نهایت انتخاب می‌شود، براساس سیاست(policy) بهینه باشد.

ج) در یک MDP قطعی (یعنی به ازای هر state و action به یک state مشخص بصورت قطعی می‌رویم.) الگوریتم Q-learning با استفاده از نرخ یادگیری(learning rate) $\alpha = 1$ در هر بروزرسانی، به طور صحیح مقدارهای بهینه Q-values را یاد می‌گیرد.

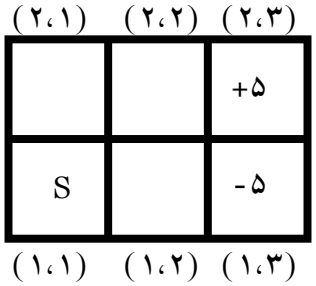
د) اگر بدانیم که $|S| >> |A|$ آنگاه پیچیدگی زمانی اجرای هر ایتریشن در policy iteration و value iteration با هم برابر میشود.

ه) در یک MDP با حالات محدود و پاداش با باند مشخص و $\gamma < ۱$ در صورتی همه پاداش‌ها با عدد ثابت c جمع شود، سیاست بهینه تغییر نمی‌کند.

۲. (۱۵ نمره، درجه سختی ۶) در شکل زیر یک محرک قرار دارد که همواره از خانه $(۱, ۱)$ که با S نشان داده شده است، شروع به حرکت می‌کند. دو تا از خانه‌های جدول خانه‌های پایانی هستند، خانه $(۲, ۳)$ با جایزه به مقدار $+۵$ و خانه $(۱, ۳)$ به مقدار -۵. در خانه‌هایی که پایانی نیستند، جایزه‌ای وجود ندارد.(جایزه برای یک خانه در صورتی دریافت می‌شود که محرک به آن خانه برود). تابع احتمال برای حرکت از هر خانه به خانه‌های مجاور به این صورت است که: برای هر حرکت به سمت بالا، پایین، چپ و راست با احتمال ۰/۸ این حرکت انجام می‌شود و با احتمال ۰/۱ حرکتی عمود بر حرکت در نظر گرفته شده انجام می‌شود. اگر محرک با دیوار برخورد کند، محرک در همان خانه‌ای که قرار داشته است می‌ماند.



تابع احتمال حرکت



شکل مربوط به سوال

الف) فرض کنید محرک احتمال‌هایی که برای حرکت‌ها وجود دارد را می‌داند. سه مرحله‌ی اولیه الگوریتم Value Iteration را برای هر وضعیت(خانه) انجام دهید. فرض $\gamma = ۰/۹$ و $V_۰ = ۰$برای هر وضعیتی در نظر بگیرید.

ب) فرض کنید محرک احتمال‌های مربوط به حرکت را نمی‌داند. چه کاری باید انجام دهد تا سیاست بهینه را یاد بگیرد؟

ج) با استفاده از $\alpha = ۰/۱$ و مقدارهای اولیه صفر، آپدیت‌های Temporal Difference-Learning را بعد از تجربه‌ی $(۱,۳) - (۱,۲) - (۱,۱)$ و $(۲,۳) - (۲,۲) - (۱,۲) - (۱,۱)$ برای $V(s)$ ها بنویسید($\gamma = ۰/۹$)
(توجه داشته باشید که در هر یک از مسیرهای گفته شده، محرک به ترتیب از چپ به راست خانه‌ها را طی کرده‌است.)

۳. (۱۵ نمره، درجه سختی ۶)

(آ) توضیح دهید که الگوریتم Q-Learning چگونه در محیط grid world ۴ × ۴ کار می‌کند، جایی که عامل از موقعیت $(۰, ۰)$ شروع می‌کند و هدف در $(۳, ۳)$ است. نحوه استفاده از قاعده Bellman update در Q-Learning را شرح دهید.

(ب) با توجه به شرایط زیر در grid world ۴ × ۴، مقادیر Q را برای دو iteration به صورت دستی محاسبه کنید:

- عامل از $(۰, ۰)$ شروع می‌کند و هدف در $(۳, ۳)$ است.
- پاداش‌ها: -۱ برای حالات غیرنهایی، $+۱۰$ برای رسیدن به هدف.
- اعمال: بالا، پایین، چپ، راست.
- ضریب تخفیف :(gamma) ۰/۹.
- نرخ یادگیری :(alpha) ۰/۱.
- عامل در این هاiteration به طور تصادفی اقدامات را انتخاب می‌کند.

محاسبات به‌روزرسانی مقادیر Q را برای هر گامی که عامل برمی‌دارد تا به هدف برسد برای دو iteration نشان دهید.

(ج) در زمینه استراتژی epsilon-greedy در Q-Learning، نقش‌های اکتشاف و استفاده از اطلاعات را توضیح دهید. چگونه باید epsilon را بر اساس زمان تنظیم کرد و چرا؟

۴. (۲۰ نمره، درجه سختی ۸) یک MDP متناهی مانند $M = (S, A, T, R, \gamma)$ در نظر بگیرید، به طوری که S فضای وضعیت، A فضای عمل، T احتمالات انتقال، R تابع پاداش و γ ضریب تخفیف (discount factor) است. Q^* را به صورت $Q^*(s,a) = Q_{\pi^*}(s,a)$ تعریف کنید بطوریکه π^* سیاست بهینه است. فرض کنید ما یک تخمین $\tilde{Q}$ از Q^* داشته باشیم، و $\tilde{Q}$ بوسیله نرم l_∞ با توجه به تعریف زیر محدود شده است:

$$||\tilde{Q} - Q^*||_\infty \leq \varepsilon$$

که در آن $||x||_\infty = max_{s,a}|x(s,a)|$ است.
فرض کنید ما به دنبال سیاست حریصانه برای $\tilde{Q}$ هستیم، $\pi(s) = argmax_{a\in A}\tilde{Q}(s,a)$. می‌خواهیم نشان دهیم که عبارت زیر برقرار است:

$$V_\pi(s) \geq V^*(s) - \frac{۲\varepsilon}{۱-\gamma}$$

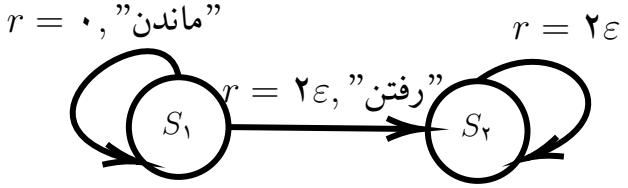
که در آن $V_\pi(s)$ تابع ارزش سیاست حریصانه π است و $V^*(s) = max_{a\in A}Q^*(s,a)$ تابع ارزش بهینه است. این نشان می‌دهد که اگر ما یک تابع ارزش state-action تقریباً بهینه را محاسبه کنیم و سپس سیاست حریصانه را برای آن تابع ارزش state-action تقریبی استخراج کنیم، سیاست نتیجه‌گرفته همچنان در MDP واقعی خوب عمل می‌کند. حال با کمک دو قسمت ابتدایی سوال زیر عبارت گفته شده را اثبات کنید.

الف) فرض کنید π^* سیاست بهینه باشد، V^* تابع ارزش بهینه و همانطور که در بالا تعریف شده است $\pi(s) = \arg\max_{a\in A} Q(s,a)$. نشان دهید نامساوی زیر برای تمام وضعیت‌های $s \in S$ برقرار است.

$$V^*(s) - Q^*(s, \pi(s)) \leq ۲\varepsilon$$

ب) با استفاده از نتیجه قسمت قبل اثبات کنید: $V_\pi(s) \geq V^*(s) - \frac{۲\varepsilon}{۱-\gamma}$

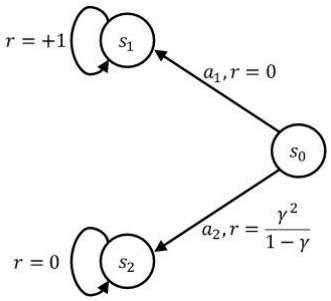
اکنون نشان می‌دهیم که حالت تساوی این نامساوی برقرار است. یک MDP دو وضعیتی را که در شکل زیر نشان داده شده است، در نظر بگیرید. وضعیت $s_۱$ دو عمل دارد، "ماندن" که با پاداش ۰ به خودش منتقل می‌شود و "رفتن" که به وضعیت $s_۲$ با پاداش ۲ε منتقل می‌شود. وضعیت $s_۲$ نیز به خودش با پاداش ۲ε منتقل می‌شود.



ج) مقدار ارزش بهینه $V^*(s)$ برای هر وضعیت و $Q^*(s,a)$ را برای وضعیت $s_۱$ و هر عمل محاسبه کنید.

د) نشان دهید که یک تابع ارزش state action تقریبی $\tilde{Q}$ با خطا ε (با استفاده از نرم l_∞) وجود دارد، به طوری که $V_\pi(s_۱) - V^*(s_۱) = -\frac{۲\epsilon}{۱-\gamma}$، که در آن $\pi(s) = \arg\max_{a\in A}\tilde{Q}(s,a)$ است.

۵. (۲۰ نمره، درجه سختی ۸) در این مسئله، یک مثال برای محدود کردن تعداد گام‌های لازم برای یافتن سیاست بهینه با استفاده از Value Iteration مشاهده می‌کنید. یک MDP با ضریب تخفیف (discount factor) $\gamma < ۱$ که در شکل زیر نشان داده شده است را در نظر بگیرید. این MDP شامل ۳ وضعیت است، و پاداش‌ها به محض انجام یک عمل از وضعیت داده می‌شود. در وضعیت $s_.$، عمل $a_۱$ دارای پاداش فوری صفر است و باعث یک انتقال قطعی به وضعیت $s_۱$ می‌شود که پاداش $+۱$ برای هر گام بعدی دارد (بدون توجه به عمل). از وضعیت $s_.$، عمل $a_۲$ باعث یک انتقال قطعی به وضعیت $s_۲$ با پاداش فوری $\frac{\gamma^۲}{(۱-\gamma)}$ می‌شود، اما وضعیت $s_۲$ برای هر گام بعدی (بدون توجه به عمل) پاداش صفر دارد.



شکل ۱: MDP با سه وضعیت

الف) مجموع پاداش تخفیف داده شده $(\sum_{t=۰}^{\infty} \gamma^t r_t)$ با انجام عمل $a_۱$ از وضعیت s. در مرحله زمانی $t = ۰$ چقدر است؟

ب) مجموع پاداش تخفیف داده شده $(\sum_{t=۰}^{\infty} \gamma^t r_t)$ با انجام عمل $a_۲$ از وضعیت s. در مرحله زمانی $t = ۰$ چقدر است؟
عمل بهینه در این وضعیت کدام است؟

ج) فرض کنید هر وضعیت را صفر مقداردهی اولیه کنیم.
(یعنی در مرحله $n = ۰$، $\forall s : V_{n=۰}(s) = ۰$). نشان دهید که الگوریتم Value Iteration تا زمانی ادامه می‌دهد که عمل sub-optimal را در n^* مرحله پیدا کند به صورتیکه:

$$n^* \geq \frac{\log(۱-\gamma)}{\log \gamma} \geq \frac{۱}{۲} \log\left(\frac{۱}{۱-\gamma}\right) \frac{۱}{۱-\gamma}.$$

بنابراین، زمان الگوریتم Value Iteration سریع‌تر از $\frac{۱}{(۱-\gamma)}$ رشد می‌کند. (شما فقط باید درستی نامساوی اول را نشان دهید)