به نام خدا

دانشکده مهندسی کامپیوتر

مبانی و کاربردهای هوش مصنوعی ترم پاییز ۱۴۰۰

تمرین سوم

مهلت تحویل ۱۰ دی ۱۴۰۰

سوال ۱ (۱۰ نمره)

در مورد فرآیند تصمیم گیری مارکوف[1] به سوالات زیر پاسخ دهید.

الف) **مقدار مورد انتظار یک وضعیت** یا $V^*(s)$ چیست؟

ب) **سیاست بهینه در یک وضعیت** یا $\pi^*(s)$ چیست؟

ج) ارتباط میان دو مورد بالا چیست؟ آیا می‌توان از یکی به دیگری رسید؟ دلیل خود را توضیح دهید.

سوال ۲ (۱۰ نمره)

در مورد جملات زیر صحیح یا غلط بودنشان را تعیین کنید و برای دلیل آن یک توضیح (مختصر) نیز ارائه دهید.

الف) در یک MDP برای تصمیم گیری در هر لحظه، هیچ نیازی به تاریخچه تصمیمات تا آن لحظه نداریم.

ب) با افزایش مقدار γ در معادله بلمن[2]، تمرکز ما روی رویداد های اخیر بیشتر می‌شود.

ج) اگر مقدار پاداش به ازای عمل a در وضعیت s برابر r باشد و هیچ عمل ممکن دیگری نداشته باشیم (تنها امکان انتخاب a را داریم) مقدار ارزش بهینه[3] آن وضعیت (s) برابر r خواهد بود.

د) به دست آوردن **مقادیر عمل-وضعیت بهینه**[4] برای یک MDP معادل **حل شدن** آن MDP است.

---

[1] Markov decision process

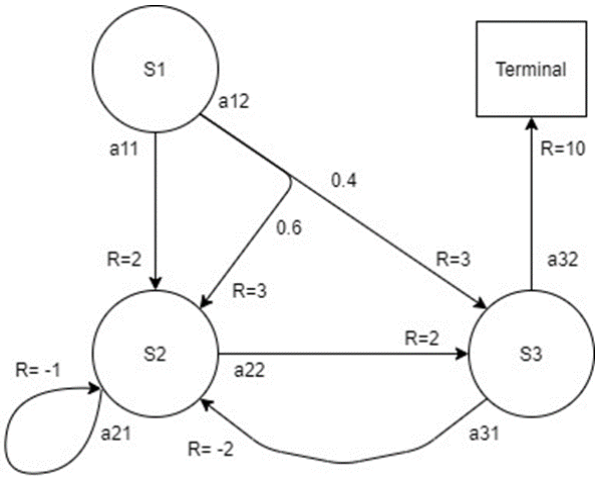
[2] Bellman equation

[3] V*(s)

[4] Q*(s,a)

سوال ۳ (۴۰ نمره)

برای این سوال، MDP زیر را در نظر بگیرید. (مقدار γ را ۰.۵ در نظر بگیرید.)



الف) فرض کنید در ابتدا مقدار ارزش[5] همه وضعیت‌ها (S) صفر است. مقادیر V وضعیت های مختلف را بر اساس فرمول به روز رسانی ارزش (که در زیر آمده است) در **سه** مرحله به روز کنید.

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s')\right]$$

ب) آیا انجام این عملیات به تعداد دفعات **متناهی** منجر به **همگرایی** مقادیر V خواهد شد؟ چرا؟

ج) بر اساس مقادیری که در بخش الف به دست آوردید، سیاست را تعیین کنید. این کار به معنی تعیین عمل[6] بهینه به ازای هر وضعیت، با استفاده از فرمول زیر است.

$$\pi\ (s) = \arg\max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V\ (s')]$$

(امتیازی) د) پس از محاسبه سیاست، آن را ارزیابی کنید. یعنی طبق فرمول زیر، مقدار ارزش هر وضعیت را تا زمانی که همگرا شود، تعیین کنید (حداقل ۳ تکرار).

$$V^{\pi}_{k+1}(s) \leftarrow \sum_{s'} T(s,\pi(s),s')[R(s,\pi(s),s') + \gamma V^{\pi}_k(s')]$$

(امتیازی) ه) پس از محاسبه بخش قبل، سیاست را (طبق معادله زیر) به روز کنید.

$$\pi_{i+1}(s) = \arg\max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^{\pi_i}(s')\right]$$

آیا تغییری در سیاست نسبت به بخش ج مشاهده می‌شود؟

نکته: محاسبات بخش الف، ج، د و ه را به صورت کامل بنویسید.

---

[5] value

[6] action

سوال ۴ (۱۰ نمره)

به سوالات زیر پاسخ کامل دهید.

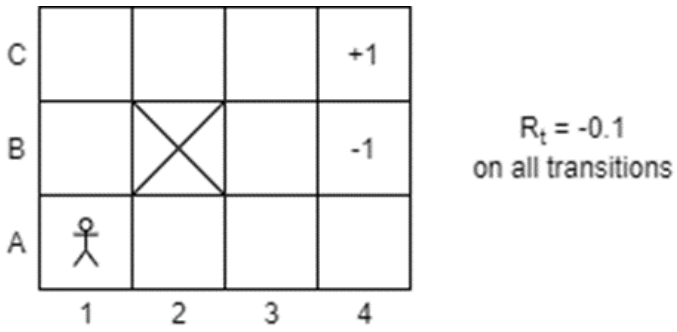
الف) از کاربردهای یادگیری تقویتی ۳ مثال زده و در هر کدام عامل[7] و محیط[8] را تعریف کنید.

ب) در کدام یک از این محیطها استفاده از یادگیری مبتنی بر مدل[9] و در چه محیطهایی یادگیری بدون مدل[10] مناسبتر است؟ چرا؟

ج) برای هر کدام از این محیطها دوراهی بین اکتشاف[11] و استخراج[12] را تعریف کنید.

سوال ۵ (۳۰ نمره)

در محیط زیر عامل ابتدا در خانه A1 قرار دارد. حرکاتی که عامل می‌تواند انجام دهد عبارت‌اند از شمال (N)، شرق (E)، جنوب (S) و غرب (W) که در هر حرکت عامل پاداشی برابر با -0.1 دریافت می‌کند. همچنین ارزش خانه C4 برابر با $+1$ و ارزش خانه B4 برابر با -1 می‌باشد. هرگاه عامل به یکی از این دو خانه برسد دنباله حرکات به اتمام رسیده و عامل باید مجددا از خانه A1 شروع به حرکت کند. در صورتی که عامل در زمان یادگیری به ترتیب دنباله‌های زیر را تجربه کرده باشد، به سوالات پاسخ دهید.



Episode 1: (A1, U, B1), (B1, U, C1), (C1, W, C1), (C1, U, C1), (C1, E, C1), (C1, E, C2), (C2, E, C3), (C3, E, C4)

Episode 2: (A1, E, A2), (A2, W, A2), (A2, S, A3), (A3, E, A4), (A4, U, B4)

Episode 3: (A1, S, A1), (A1, E, A2), (A2, W, A1), (A1, E, A2), (A2, U, A2), (A2, E, A3), (A3, U, B3), (B3, E, B4)

Episode 4: (A1, E, A2), (A2, E, A2), (A2, E, A3), (A3, U, B3), (B3, S, A3), (A3, U, B3), (B3, S, A3), (A3, U, B3), (B3, W, B3), (B3, E, B4)

Episode 5: (A1, E, A2), (A2, S, A2), (A2, U, A2), (A2, E, A3), (A3, E, A4), (A4, U, B4)

Episode 6: (A1, E, A2), (A2, S, A2), (A2, W, A1), (A1, E, A2), (A2, W, A3), (A3, U, B3), (B3, W, B3), (B3, S, C3), (C3, S, C3), (C3, E, C4)

---

[7] agent

[8] environment

[9] model-based learning

[10] model-free learning

[11] exploration

[12] exploitation

الف) با استفاده از روش ارزیابی مستقیم[13] و با فرض 1=γ، تابع ارزش را برای تمام حالات بدست آورید.

ب) با استفاده از یادگیری تفاوت زمانی[14] و با فرض 0.5 = α و 1=γ تابع ارزش Q را بدست آورید (فرمول زیر).

$$sample = R(s,a,s') + \gamma \max_{a'} Q(s',a')$$

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + (\alpha)\,[sample]$$

ج) تابع ارزش حالات را با استفاده از تابع ارزش Q که در قسمت قبل به دست آمد محاسبه کنید. آیا نتیجه مشابه با بخش الف می‌باشد؟ تفسیر شما چیست؟

نکته: محاسبات بخش الف، ب و ج را به صورت کامل بنویسید.

سوال ۶ (۱۰ نمره)

یادگیری Q[15] یک الگوریتم از نوع برون‌سیاست[16] می‌باشد. این دقیقا به چه معناست؟ یک الگوریتم یادگیری تفاوت زمانی از نوع برسیاست[17] مثال بزنید. این الگوریتم چه تفاوتی با یادگیری Q دارد؟

سوال ۷ (۱۰ نمره)

فرض کنید می‌خواهیم از روش یادگیری Q تخمینی[18] در یک ماشین خودران برای خودداری از برخورد با عابرین پیاده استفاده کنیم:

الف) شما از کدام ویژگی‌های محیط برای تشکیل تابع ارزش خطی[19] استفاده می‌کنید؟ چرا؟

ب) دو حالت که بر اساس این ویژگی‌ها مشابه هستند اما ارزش بسیار متفاوت دارند را مثال بزنید.

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) آپلود کنید.
- فرمت نامگذاری تمرین باید مانند AI_HW3_9931099.pdf باشد.
- در صورت هرگونه سوال یا ابهام از طریق ایمیل ai.aut.fall1400@gmail.com با تدریسیاران در تماس باشید، همچنین خواهشمند است در متن ایمیل به شماره دانشجویی خود اشاره کنید.
- ددلاین این تمرین **۱۰ دی ۱۴۰۰ ساعت ۲۳:۵۵** است و امکان ارسال با تاخیر وجود ندارد،بنابراین بهتر است انجام تکلیف را به روز های پایانی موکول نکنید.

---

[13] direct evaluation

[14] temporal difference learning

[15] Q-learning

[16] off-policy

[17] on-policy

[18] approximate Q-learning

[19] Linear value function