دانشکده مهندسی کامپیوتر

هوش مصنوعی و سیستم‌های خبره

تمرین تشریحی چهارم

نام و نام خانوادگی . . . .

شماره دانشجویی . . . .

مدرس . . . . محمدطاهر پیله‌ور - سید صالح اعتمادی

طراحی و تدوین . . . . سپهر باباپور (@Spr_Bpr)

تاریخ انتشار . . . . ۱۵ آبان ۱۳۹۹

تاریخ تحویل . . . . ۲۹ آبان ۱۳۹۹

فهرست مطالب

۱ سوالات بخش تئوری

در این بخش به سوالاتی که دارای ⋆ هستند پاسخ دهید

۱.۱ سوال ۱

توضیح دهید چرا در MDPها نمی‌توان از روش planning استفاده کرد. راه‌حل جایگزین را توضیح دهید.

پاسخ:

از آنجایی که در مسائله MDP انتخاب یک action لزوما به معنای انجام آن action نیست و به نوعی قطعیتی در انجام تصمیم‌گیری‌ها وجود ندارد، نمی‌توان از روش planning در این مسائل استفاده کرد.

راه‌حلی که برای اینگونه مسائل ارائه می‌شود، محاسبه راهبرد (policy) بهینه برای هر حالت می‌باشد. توضیح بیشتر آنکه در هر خانه باید محاسبه شود کدام action بهترین نتیجه را برای ما دارد.

۲.۱ سوال ۲

فرایندهای مارکوفی را تعریف کرده و بگویید کدام یک از فرایندهای زیر مارکوفی هستند.

- بازی سودوکو
- بازی اسم – فامیل
- بازی بی‌سوالی

پاسخ:

فرایندهای مارکوفی به فرایندهایی می‌گویند که حالت آینده تنها به حالت فعلی وابسته است و حالت‌های پیشین در حالت آینده تاثیری ندارند.

طبق این تعریف بازی‌های فوق را دسته‌بندی می‌کنیم:

- بازی سودوکو: فرانید مارکوفی است.
- بازی اسم – فامیل: فرایند مارکوفی است.
- بازی بی‌سوالی: فرایند مارکوفی نیست.

۳.۱ ⋆ سوال ۳ (۲۰ نمره)

اگر به جای ضریب تخفیف γ^t از توابع زیر استفاده شود:

* $|sin(t)|$
* e^{-t}
* $log(t)$

به سوالات زیر پاسخ دهید.

۱- کدام یک مشکل نامحدود شدن بازی را برطرف می‌کنند؟ توضیح دهید.

۲- برای تابع پاسخ قسمت اول، با فرض پاداش یک واحد در هر لحظه کوچک زمانی (dt) پاداش کل را محاسبه کنید.

پاسخ:

۱- مشکل نامحدود شدن بازی زمانی بوجود میاید که مقدار ضریب تخفیف کاهش نیابد. از بین توابع فوق، تنها تابعی که در طول زمان کاهش می‌یابد و به صفر میل می‌کند، تابع e^{-t} می‌باشد.

۲- برای محاسبه پاداش کل ضریب تخفیف تابع فوق داریم:

$$\int_0^{\infty} e^{-t}\,dt = -[e^{-t}]_0^{\infty} = -[e^{-\infty} - e^0] = -[0-1] = 1 \qquad (۱)$$

۲ مسائل محاسباتی

** در این بخش به سوالاتی که دارای ⋆ هستند پاسخ دهید **

۱.۲ سوال ۱: کارت بردار!!

فرض کنید شما در یک مسابقه کارت بازی شرکت کرده‌اید که در آن ۳ نوع کارت با شماره‌های ۲، ۳، ۴ وجود دارد. شما در هر مرحله از بازی تا زمانی که به مجموع امتیاز ۶ نرسیده‌اید می‌توانید یا یک کارت بردارید یا بازی را به اتمام برسانید. احتمال آمدن هر کارت با هم برابر است. زمانی که مجموع امتیازات شما ۶ یا بیشتر شود امتیازات شما صفر می‌شود و بازی تمام می‌شود و زمانی که خودتان بازی را تمام کرده باشید امتیازتان برابر مجموع کارت‌هایی که کسب کرده‌اید می‌شود. همچنین برداشتن کارت را بدون هزینه در نظر بگیرید.
در این سوال از شما خواسته شده است که بازی فوق را به صورت یک مدل مارکوفی در نظر بگیرید و به سوالات زیر پاسخ دهید.

۱. ابتدا تابع انتقال (transition function) و تابع پاداش (reward function) را برای این مدل محاسبه کنید.

۲. سپس جدول زیر را کامل کنید.

حالت	۰	۲	۳	۴	۵
π_i	برداشتن کارت	اتمام بازی	برداشتن کارت	اتمام بازی	برداشتن کارت
v^{π_i}					
π_{i+1}					

شکل ۱: جدول سوال کارت بردار!!

پاسخ:

۱- تابع انتقال (transition function) برای این مسئله به شرح زیر می‌باشد:

$$T(s, Stop, Done) = 1$$

$$T(s, Draw, s') = \begin{cases} 1/3 & \text{If } s' \in \{2, 3, 4\} \\ 1/3 & \text{If } s \in \{2\} \text{ and } s' \in \{4, 5, Done\} \\ 1/3 & \text{If } s \in \{3\} \text{ and } s' \in \{5\} \\ 2/3 & \text{If } s \in \{3\} \text{ and } s' \in \{Done\} \\ 1 & \text{If } s \in \{4, 5\} \text{ and } s' \in \{Done\} \end{cases}$$

و حالت‌هایی خارج از حالات فوق دارای تابع انتقالی برابر با صفر می‌باشند.
برای تابع پاداش (reward function) این مسئله داریم:

$$R(s, Stop, Done) = s,\ \ s \leq 5$$

$$R(s, a, s') = 0,\ \text{Otherwise}$$

۲- ابتدا تمامی ارزش‌های حالات مختلف را محاسبه می‌کنیم. در حالت ۵ از آنجایی که action برابر با برداشتن کارت می‌باشد، هر کارتی که انتخاب شود، مجموع امتیازات را به بالای عدد ۵ می‌رساند، ارزش این حالت برابر با صفر می‌شود.

$$v_5^{\pi_i} = 0$$

برای دو حالت ۲ و ۴ نیز از آنجایی که action برابر با اتمام بازی می‌باشد، ارزش این حالات به ترتیب برابر با ۲ و ۴ می‌باشد.

$$v_2^{\pi_i} = 2$$
$$v_4^{\pi_i} = 4$$

حال برای محاسبه ارزش حالت ۳ داریم:

$$v_3^{\pi_i} = \frac{1}{3} \times v_5^{\pi_i} = 0$$

و برای محاسبه ارزش حالت صفر داریم:

$$v_0^{\pi_i} = \frac{1}{3} \times v_2^{\pi_i} + \frac{1}{3} \times v_3^{\pi_i} + \frac{1}{3} \times v_5^{\pi_i} = \frac{1}{3} \times (2+4) = \frac{6}{3} = 2$$

حال با توجه به ارزش‌های محاسبه شده در فوق به بروزرسانی راهبرد (policy) می‌پردازیم: برای این کار از حالت ۵ آغاز می‌کنیم. همانطور که در بالا محاسبه شد، در صورتی که در این حالت کارت برداشته شود به طور متوسط صفر امتیاز گرفته می‌شود، اما در صورتی که در این حالت بازی تمام شود، امتیاز نهایی برابر با ۵ می‌شود. لذا:

$$\pi_5 = \text{Stop}$$

برای حالت ۴ در صورتی که کارتی انتخاب شود، مجموع امتیازات بیشتر از ۵ می‌شود و در نتیجه امتیاز گرفته شده برابر با صفر می‌شود، لذا معقول است در این حالت بازی را پایان دهیم و امتیاز ۴ داشته باشیم:

$$\pi_4 = \text{Stop}$$

در حالت ۳ امتیاز گرفته شده برای حالتی که کارت برداشته می‌شود، در فوق محاسبه شده است. از آنجایی که با اتمام بازی در این حالت، امتیاز ۳ کسب می‌شود، به نسبت امتیاز صفر گرفته شده در فوق معقول است تا در این حالت نیز بازی پایان یابد:

$$\pi_3 = \text{Stop}$$

برای حالت ۲ باید مقداری محاسبه انجام شود. حالتی را فرض کنید که در این حالت کارت برداشته شود. برای این action داریم:

$$v_2^{\pi_{i+1}} = \frac{1}{3} \times v_4^{\pi_i} + \frac{1}{3} \times v_5^{\pi_i} = \frac{1}{3} \times (4+0) = \frac{4}{3}$$

از آنجایی که در صورتی که در این حالت بازی را پایان دهیم، امتیازی برابر با ۲ کسب می‌کنیم،

به نسبت امتیازی که در فوق محاسبه کردیم، معقول است در این حالت بازی را پایان دهیم:

$$\pi_2 = \text{Stop}$$

در حالت آخر یعنی حالت صفر واضح است با اتمام بازی، صفر امتیاز کسب می‌شود که در مقایسه با امتیاز محاسبه شده در فوق برای برداشتن کارت عملکرد ضعیف‌تری است، لذا:

$$\pi_0 = \text{Draw}$$

۲.۲ ⋆ سوال ۲: تاس بریز!! (۴۵ نمره)

فرض کنید در یک بازی ریختن تاس شرکت کرده‌اید که هزینه هر بار ریختن تاس در آن ۱ سکه است و احتمال آمدن تمام اعداد در تاس با یکدیگر برابر است. شما پس از ریختن تاس به اندازه 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 دو انتخاب دارید:

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

لذا بازی را می‌توان به این صورت در نظر گرفت که بازیکن در ابتدای بازی 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 نمایش داده می‌شود که بدین معنی است که عدد i در تاس آمده است.
با توجه به توضیحات فوق به سوالات زیر پاسخ دهید:

۱. فرض کنید π_iهای زیر در ابتدا وجود دارد، ردیف v^{π_i} را کامل کنید. $(\gamma = 1)$

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}						

شکل ۲: جدول قسمت اول سوال تاس بریز!!

۲. باتوجه به جدول فوق مقادیر π_i را بروزرسانی کنید و در جدول زیر جایگذاری کنید. این مقادیر می‌تواند سه حالت تاس ریختن، اتمام بازی و تاس ریختن / اتمام بازی باشد. $(\gamma = 1)$

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}						

شکل ۳: جدول قسمت دوم سوال تاس بریز!!

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

پاسخ:

۱- از آنجایی که در سوال گفته شده است اتمام بازی به معنای کسب امتیازی برابر با عدد روی تاس می‌باشد، برای حالات ۳ تا ۶ داریم:

$$v_3^{\pi_i} = 3 \qquad v_4^{\pi_i} = 4 \qquad v_5^{\pi_i} = 5 \qquad v_6^{\pi_i} = 6$$

اما برای محاسبه ارزش حالات ۱ و ۲ راه‌حل مقداری پیچیده‌تر است. به روابط که برای حالات ۱ و ۲ نوشته شده‌اند، توجه کنید:

$$v_1^{\pi_i} = \frac{1}{6}(v_1^{\pi_i} + v_2^{\pi_i} + v_3^{\pi_i} + v_4^{\pi_i} + v_5^{\pi_i} + v_6^{\pi_i})$$

$$v_2^{\pi_i} = \frac{1}{6}(v_1^{\pi_i} + v_2^{\pi_i} + v_3^{\pi_i} + v_4^{\pi_i} + v_5^{\pi_i} + v_6^{\pi_i})$$

با جایگذاری مقادیر محاسبه شده آغاز راه‌حل برای حالات ۳ تا ۶، در روابط فوق داریم:

$$v_1^{\pi_i} = -1 + \frac{1}{6}(v_1^{\pi_i} + v_2^{\pi_i} + 3 + 4 + 5 + 6)$$

$$v_2^{\pi_i} = -1 + \frac{1}{6}(v_1^{\pi_i} + v_2^{\pi_i} + 3 + 4 + 5 + 6)$$

از حل دو معادله - دو مجهول فوق داریم:

$$v_1^{\pi_i} = v_2^{\pi_i} = 3$$

۲- با توجه به توصیف بازی‌ای که در سوال ارائه شده است، می‌توان متوجه شد که ارزش هر حالت در صورتی که تاس ریخته شود مستقل از دیگری و برابر است. این ارزش برابر است با:

$$v_s^{\pi_i} = -1 + \frac{1}{6}(3 + 3 + 3 + 4 + 5 + 6) = -1 + \frac{24}{6} = 3$$

همچنین ارزش هر حالت در صورتی که بازی پایان یابد برابر با عدد روی تاس می‌باشد:

$$v_s^{\pi_i} = s$$

لذا می‌توان نتیجه گرفت در حالت‌های ۴، ۵ و ۶ که در آن‌ها امتیاز کسب شده در حالت اتمام بازی بیشتر از ریختن تاس می‌باشد، راهبرد برابر با اتمام بازی می‌باشد:

$$\pi_4 = \pi_5 = \pi_6 = \text{Stop}$$

در حالت ۳ این دو امتیاز با یکدیگر برابر است، لذا هر دو راهبرد بهینه می‌باشند:

$$\pi_3 = \text{Roll/Stop}$$

و در نهایت برای دو حالت ۱ و ۲ درست برخلاف حالت‌های ۴، ۵ و ۶ ریختن تاس امتیاز بیشتری نسبت به اتمام بازی کسب می‌کند و لذا ریختن تاس راهبرد بهینه می‌باشد:

$$\pi_1 = \pi_2 = \text{Roll}$$

۳- از آنجایی که راهبرد محاسبه شده در گام $i+1$ مشابه با گام i می‌باشد، می‌توان نتیجه گرفت همگرایی صورت گرفته است. این همگرایی نشان می‌دهد که ما به راهبرد (policy) بهینه دست یافته‌ایم.

۳.۲ ⋆ سوال ۳: یک **MDP** ساده که دیگر مثل قبل نیست؟ (۳۵ نمره)

یک مسئله MDP را تصور کنید که در آن تابع پاداش به جای $R(s)$، $\eta R(s)$ باشد که در آن η یک ثابت مثبت است. سایر خصوصیات این مسئله MDP تغییر نکرده است.
ثابت کنید راهبرد (policy) بهینه در مسئله MDP جدید مشابه راهبرد (policy) در مسئله اولیه است.

پاسخ:

از آنجایی که هر دو مسئله اصلی و تغییر داده شده یک مسئله MDP معتبر هستند، هر مسئله دارای یک تابع بهینه برای $Q^*(s,a)$ خود می‌باشند.
با این تفسیر، تابع Q را برای مسئله اصلی $Q_o^*(s,a)$ و برای مسئله تغییر داده شده، $Q_m^*(s,a)$ می‌نامیم.

حالتی را در نظر بگیرید که در آن (s,a) جفت state-action گام آخر باشد. در این حالت داریم:

$$\begin{aligned}
Q^*_m(s,a) &= \sum_{s'} T(s,a,s')[\eta R(s,a,s') + \gamma \max_{a'} Q^*_m(s',a')] \\
&= \sum_{s'} T(s,a,s')[\eta R(s,a,s') + \gamma \max_{a'} \sum_{s''} T(s',a',s'')\eta R(s',a',s'')] \\
&= \eta \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma \max_{a'} \sum_{s''} T(s',a',s'') R(s',a',s'')] \\
Q^*_m(s,a) &= Q^*_o(s,a)
\end{aligned}$$

حال فرض کنید در گام یکی مانده به آخر هستیم:

$$\begin{aligned}
Q^*_m(s,a) &= \sum_{s'} T(s,a,s')[\eta R(s,a,s') + \gamma \max_{a'} Q^*_m(s',a')] \\
&= \sum_{s'} T(s,a,s')[\eta R(s,a,s') + \gamma \max_{a'} \eta Q^*_o(s',a')] \\
&= \eta \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma \max_{a'} Q^*_o(s',a')] \\
Q^*_m(s,a) &= Q^*_o(s,a)
\end{aligned}$$

بنابراین با توجه به استقرا می‌توان گفت $Q^*_m(s,a) = Q^*_o(s,a)$ برای تمام گام‌ها صادق است.

حال برای راهبرد بهینه مسئله MDP تغییر داده شده می‌توان نوشت:

$$\begin{aligned}
\pi^*_m(s) &= \operatorname*{argmax}_a Q^*_m(s,a) \\
&= \operatorname*{argmax}_a \eta Q^*_o(s,a) \\
&= \operatorname*{argmax}_a Q^*_o(s,a) \\
\pi^*_m(s) &= \pi^*_o(s)
\end{aligned}$$