دانشكده مهندسي كامپيوتر

هوش مصنوعي و سيستم‌هاي خبره

تمرين تشريحي چهارم

نام و نام خانوادگي . . . . . غزل زمانی نژاد .....

شماره دانشجويي . . . . . 97522166.....

مدرس محمدطاهر پيله‌ور - سيد صالح اعتمادي

طراحي و تدوين . . . . . سپهر باباپور (@Spr_Bpr)

تاريخ انتشار . . . . . ۵۱ آبان ۹۹۳۱

تاريخ تحويل . . . . . ۹۲ آبان ۹۹۳۱

فهرست مطالب

۱ سوالات بخش تئوري

** در اين بخش به سوالاتي كه داراي ⋆ هستند پاسخ دهيد **

۱.۱ سوال ۱

توضيح دهيد چرا در MDPها نمي‌توان از روش planning استفاده كرد. راه‌حل جايگزين را توضيح دهيد.

پاسخ:

.....

.....

.....

.....

.....

.....

۲.۱ سوال ۲

فرايندهاي ماركوفي را تعريف كرده و بگوييد كدام يك از فرايندهاي زير ماركوفي هستند.

- بازي سودوكو - بازي اسم - فاميل -بازي بي‌سوالي

پاسخ:

.....

.....

.....

.....

.....

۳.۱ ⋆ سوال ۳ (۰۲ نمره)

اگر به‌جاي ضريب تخفيف γt از توابع زير استفاده شود:

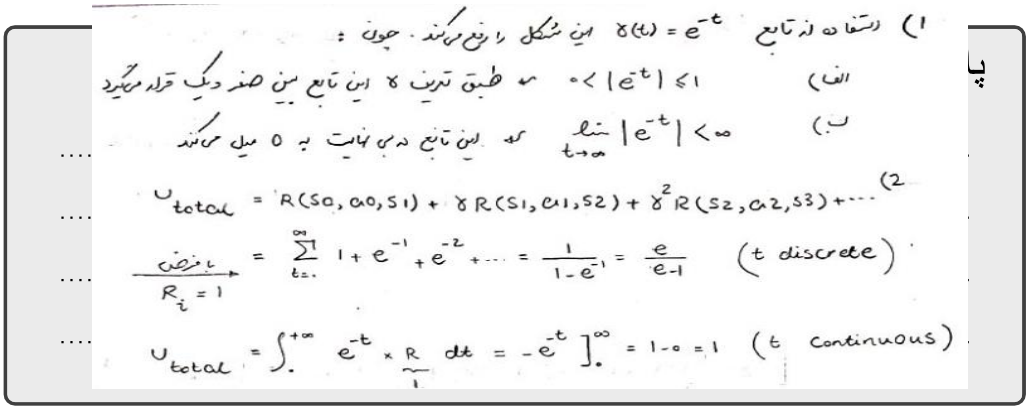
e^{-t} *

$|sin(t)|$ * $log(t)$ *

به سوالات زير پاسخ دهيد.

۱- كدام يك مشكل نامحدود شدن بازي را برطرف مي‌كنند؟ توضيح دهيد.

۲- براي تابع پاسخ قسمت اول، با فرض پاداش يك واحد در هر لحظه كوچك زماني (dt) پاداش كل را محاسبه كنيد.



۲ مسائل محاسباتي

** در اين بخش به سوالاتي كه داراي ⋆ هستند پاسخ دهيد **

۱.۲ سوال ۱: كارت بردار!!

فرض كنيد شما در يك مسابقه كارت بازي شركت كرده‌ايد كه در آن ۳ نوع كارت با شماره‌هاي ۲، ۳، ۴ وجود دارد. شما در هر مرحله از بازي تا زماني كه به مجموع امتياز ۶ نرسيده‌ايد مي‌توانيد يا يك كارت برداريد يا بازي را به اتمام برسانيد. احتمال آمدن هر كارت با هم برابر است. زماني كه مجموع امتيازات شما ۶ يا بيشتر شود امتيازات شما صفر مي‌شود و بازي تمام مي‌شود و زماني كه خودتان بازي را تمام كرده باشيد امتيازتان برابر مجموع كارت‌هايي كه كسب كرده‌ايد مي‌شود. همچنين برداشتن كارت را بدون هزينه در نظر بگيريد.
در اين سوال از شما خواسته شده است كه بازي فوق را به صورت يك مدل ماركوفي در نظر بگيريد و به سوالات زير پاسخ دهيد.

۱. ابتدا تابع انتقال (transition function) و تابع پاداش (reward function) را براي اين مدل محاسبه كنيد.

۲. سپس جدول زير را كامل كنيد.

۵	۴	۳	۲	۰	حالت
برداشت ۵ کارت	اتمام بازی	برداشت ۵ کارت	اتمام بازی	برداشت ۵ کارت	¡
					¡"
					¡#$

شکل ۱: جدول سوال کارت بردار!!

پاسخ:

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

۲.۲ ★ سوال ۲: تاس بريز!! (۵۴ نمره)

فرض كنيد در يك بازي ريختن تاس شركت كرده‌ايد كه هزينه هر بار ريختن تاس در آن ۱ سكه است و احتمال آمدن تمام اعداد در تاس با يكديگر برابر است. شما پس از ريختن تاس به اندازه عدد روي تاس سكه دريافت مي‌كنيد.
قانون بازي به اين شكل است كه شما موظف هستيد در بار اول يك تاس بريزيد، اما در ساير مراحل دو انتخاب داريد:

* اتمام بازي: با اين حركت شما به اندازه عدد روي تاس سكه دريافت مي‌كنيد.
* تاس ريختن: يك سكه هزينه مي‌كنيد و بار دگير تاس مي‌ريزيد.

لذا بازي را مي‌توان به اين صورت در نظر گرفت كه بازيكن در ابتداي بازي در حالت شروع قرار دارد و در حالت شروع فقط حركت ريختن تاس وجود دارد. در ساير حالات يك حركت اتمام بازي وجود دارد كه بازيكن را به حالت پاياني مي‌برد و در حالت پاياني حركتي وجود ندارد. هر حالت بين شروع و پايان با si نمايش داده مي‌شود كه بدين معني است كه عدد i در تاس آمده است.
با توجه به توضيحات فوق به سوالات زير پاسخ دهيد:

۱. فرض كنيد πiهاي زير در ابتدا وجود دارد، رديف $v\pi^i$ را كامل كنيد. ($\gamma = 1$)

حالت	S_1	S_2	S_3	S_4	S_5	s_6
□!	تاس ريخ &' €	تاس ريخ &' €	اتمام بازی	اتمام بازی	اتمام بازی	اتمام بازی
□¨!	3	3	3	4	5	6

شكل ۲: جدول قسمت اول سوال تاس بريز!!

۲. باتوجه به جدول فوق مقادير πi را بروزرساني كنيد و در جدول زير جايگذاري كنيد. اين مقادير مي‌تواند سه حالت تاس ريختن، اتمام بازي و تاس ريختن / اتمام بازي باشد. ($\gamma = 1$)

حالت	S_1	S_2	S_3	S_4	S_5	s_6
□!	تاس ريخ &' €	تاس ريخ &' €	اتمام بازی	اتمام بازی	اتمام بازی	اتمام بازی
□! ($	تاس ريختن	تاس ريختن	تاس ريختن / اتمام بازي	اتمام بازی	اتمام بازی	اتمام بازی

شكل ۳: جدول قسمت دوم سوال تاس بريز!!

۳. با توجه به مقادير جدول فوق آيا مي‌توان نتيجه گرفت كه مقادير بدست آمده بهينه هستند و ديگر نياز به بروزرساني ندارند؟ توضيح دهيد.

پاسخ:

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}	3	3	3	4	5	6

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}	تاس ریختن	تاس ریختن	تاس ریختن / اتمام بازی	اتمام بازی	اتمام بازی	اتمام بازی

1) $v_{\pi_3} = 3$ اتمام بازی دست اسیار $v_{\pi_4} = 4$ $v_{\pi_5} = 5$ $v_{\pi_6} = 6$

$$v_{\pi_1} = \tfrac{1}{6}(-1 + v_{\pi_1}) + \tfrac{1}{6}(-1 + v_{\pi_2}) + \tfrac{1}{6}(-1 + \underbrace{v_{\pi_3}}_{3}) + \tfrac{1}{6}(-1 + \underbrace{v_{\pi_4}}_{4}) + \tfrac{1}{6}(-1 + \underbrace{v_{\pi_5}}_{5}) + \tfrac{1}{6}(-1 + \underbrace{v_{\pi_6}}_{6})$$

$$v_{\pi_1} = -1 + \tfrac{1}{6}(v_{\pi_1} + v_{\pi_2} + 18)$$

$$v_{\pi_2} = -1 + \tfrac{1}{6}(v_{\pi_1} + v_{\pi_2} + 18)$$

$$\Rightarrow \begin{cases} \tfrac{5}{6} v_{\pi_1} - \tfrac{1}{6} v_{\pi_2} = 2 \\ \tfrac{5}{6} v_{\pi_2} - \tfrac{1}{6} v_{\pi_1} = 2 \end{cases} \Rightarrow v_{\pi_1} = v_{\pi_2} = 3$$

2) $\pi_{i+1}(s_1) = \text{argmax}\{\text{dice}:3, \text{stop}:1\} = \text{dice}$

$\pi_{i+1}(s_2) = \text{argmax}\{\text{dice}:3, \text{stop}:2\} = \text{dice}$

$\pi_{i+1}(s_3) = \text{argmax}\{\text{dice}:3, \text{stop}:3\} = \text{dice, stop}$

$\pi_{i+1}(s_4) = \text{argmax}\{\text{dice}:3, \text{stop}:4\} = \text{stop}$

$\pi_{i+1}(s_5) = \text{argmax}\{\text{dice}:3, \text{stop}:5\} = \text{stop}$

$\pi_{i+1}(s_6) = \text{argmax}\{\text{dice}:3, \text{stop}:6\} = \text{stop}$

3) converge می‌توانیم نتیجه بگیریم. π_{i+1} با π_i یکسان می‌شود. بله: اگر در π_3، stop را انتخاب کنیم، شده پس π_i بهینه است.

۳.۲ ⋆ سوال ۳: يك **MDP** ساده كه ديگر مثل قبل نيست؟ (۵۳ نمره)

يك مسئله MDP را تصور كنيد كه در آن تابع پاداش به جاي $R(s)$، $\eta R(s)$ باشد كه در آن η يك ثابت مثبت است. ساير خصوصيات اين مسئله MDP تغيير نكرده است.

ثابت كنيد راهبرد (policy) بهينه در مسئله MDP جديد مشابه راهبرد (policy) در مسئله اوليه است.

پاسخ:

3) باید ثابت کنیم اگر در M، reward ها را η برابر کنیم و M_η بدست بیاید، optimal policy تغییری نمیکند

می دانیم در M مجموع discounted reward های مربوط به π^* (optimal policy)، از سایر π ها بیشتر است یعنی:

$$\sum_{t=0}^{\infty} \gamma^t R_t(s,\pi^*,s') \geq \sum_{t=0}^{\infty} \gamma^t R_t(s,\pi,s') \qquad \forall \pi \in policies$$

می‌توانیم دو طرف نامعادله‌ی فوق را در $\eta > 0$ ضرب کنیم (جهت نامعادله به دلیل مثبت بودن η تغییر نمی‌کند):

$$\eta \sum_{t=0}^{\infty} \gamma^t R_t(s,\pi^*,s') \geq \sum_{t=0}^{\infty} \gamma^t R_t(s,\pi,s')$$

$$\Rightarrow \sum_{t=0}^{\infty} \gamma^t \eta R_t(s,\pi^*,s') \geq \sum_{t=0}^{\infty} \gamma^t \eta R_t(s,\pi,s')$$

مجموع discounted rewards در M_η

ثابت شد که optimal policy در M_η مشابه با M است.