دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی

پاییز ۱۴۰۱

استاد: محمدحسین رهبان
گردآورندگان: محمدجواد هزاره‌ـ امیرحسین جوادی

تمرین ششم — فرآیندهای مارکف و یادگیری تقویتی — مهلت ارسال: ۵ بهمن

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- امکان ارسال با تاخیر برای این تمرین وجود ندارد.
- 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.

سوالات نظری (۱۰۰ نمره)

۱. (۱۰ نمره) درستی یا نادرستی گزاره‌های زیر را در رابطه با یک فرآیند تصمیم‌گیری مارکف[۱] مشخص کنید و توضیحی کوتاه در رابطه با آن ارائه دهید.

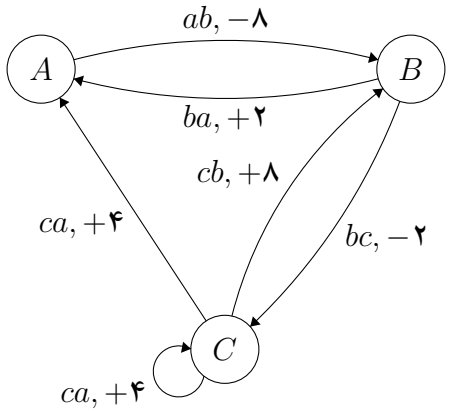
(آ) ضریب تخفیف[۲] کوچک و نزدیک به صفر به رفتار حریصانه و کوته‌نظر[۳] منجر می‌شود.

(ب) پاداش منفی زندگی[۴] با اندازه‌ی زیاد (بسیار منفی) به رفتار حریصانه و کوته‌نظر منجر می‌شود.

(ج) همواره می‌توان پاداش منفی زندگی را با استفاده از ضریب تخفیف منفی مدل کرد.

(د) همواره می‌توان ضریب تخفیف منفی را با پاداش منفی زندگی مدل کرد.

۲. (۲۵ نمره) فرآیند تصمیم‌گیری مارکف که در شکل ۱ آمده است را با ضریب تخفیف $\gamma = ۰/۵$ در نظر بگیرید که در آن حالت‌ها با حروف A، B و C نشان داده شده‌اند. روی هر یال حروف کوچک نوشته شده که یکی از کنش‌های موجود است و یال مربوطه گذار متناظر با انجام آن کنش را نشان می‌دهد. عدد صحیح روی هر یال نیز پاداش کسب شده از آن کنش است. تمام گذارها با احتمال ۱ به وقوع می‌پیوندد و تنها گذار از حالت C به A تصادفی است که احتمال رفتن به حالت A برابر $\frac{۱}{۴}$ و احتمال رفتن به حالت C برابر $\frac{۳}{۴}$ است.



شکل ۱: گراف فرآیند تصمیم‌گیری مارکف.

[۱] Markov Decision Process
[۲] discount
[۳] shortsighted
[۴] negative living reward

با در نظر گرفتن این فرآیند به سوال‌های زیر پاسخ دهید.

(آ) برای یک فرآیند تصمیم‌گیری مارکف به همراه ضریب تخفیف، تابع ارزش حالت‌ها[۵] یا همان $V^{\pi}(s)$ را توصیف کنید.

(ب) رابطه‌ی بلمن را برای تابع ارزش حالت‌ها بنویسید.

(ج) سیاست اولیه‌ی π_1 را در نظر بگیرید که به صورت تصادفی و با احتمال برابر در هر حالت یکی از کنش‌های موجود در آن حالت را انتخاب می‌کند. حال فرض کنید تابع ارزش‌گذاری اولیه را به صورت $V_1(A) = V_1(B) = V_1(C) = 2$ در نظر بگیریم. یک مرحله از الگوریتم ارزیابی سیاست[۶] را اجرا کنید تا به تابع ارزش $V_2(s)$ برای حالت‌های مختلف برسید.

(د) براساس تابع ارزش‌گذاری جدید و به صورت حریصانه سیاست قطعی جدید π_2 را بدست آورید.

(ه) سیاست قطعی π را در نظر بگیرید. اثبات کنید اگر سیاست جدید π' به صورت حریصانه از V^{π} بدست آمده باشد، آنگاه π' بهتر یا مساوی π است، یا به عبارتی برای تمام حالت‌ها داریم $V^{\pi'}(s) \geq V^{\pi}(s)$. همچنین اثبات کنید اگر تساوی برای تمام حالت‌ها رخ دهد آنگاه π' حتما سیاست بهینه است.

۳. (۲۰ نمره) صفحه‌ی ۳×۲ زیر را در نظر بگیرید. فرض کنید حرکت خود را از خانه‌ی شماره‌ی ۱ شروع می‌کنیم و با رسیدن به خانه‌ی شماره‌ی ۶ بازی تمام می‌شود و با رسیدن به این خانه ۱۰ امتیاز مثبت دریافت می‌کنیم. همچنین در تمام حرکت‌هایی که منجر به رسیدن به خانه‌ی شماره‌ی ۶ نمی‌شوند پاداش -۱ دریافت می‌کنیم.

۴	۵	۶
۱	۲	۳

شکل ۲: جدول بازی.

در هر خانه چهار کنش ممکن وجود دارد: بالا، پایین، چپ و راست. فرض کنید کنش‌هایی که باعث خارج شدن از صفحه می‌شوند مجاز نیستند. هر کنش نیز به صورت قطعی انجام شده و به خانه‌ی مربوطه می‌رویم. حال فرض کنید جدول زیر را برای $Q(s,a)$ داریم:

$Q(۱, \text{راست}) = ۳$			$Q(۱, \text{بالا}) = ۴$
$Q(۲, \text{راست}) = ۸$	$Q(۲, \text{چپ}) = ۳$		$Q(۲, \text{بالا}) = ۶$
	$Q(۳, \text{چپ}) = ۷$		$Q(۳, \text{بالا}) = ۹$
$Q(۴, \text{راست}) = ۵$		$Q(۴, \text{پایین}) = ۲$	
$Q(۵, \text{راست}) = ۸$	$Q(۵, \text{چپ}) = ۵$	$Q(۵, \text{پایین}) = ۶$	

شکل ۳: جدول Q-valueها

با درنظر گرفتن این جدول و توضیح مسئله به سوال‌های زیر پاسخ دهید.

(آ) باتوجه به داشتن دانش کامل در رابطه با محیط، می‌توان از رابطه‌ی بلمن برای بروزرسانی Q-valueها استفاده کرد. فرض کنید از سیاست حریصانه استفاده می‌کنیم و با درنظر گرفتن این سیاست، ابتدا رابطه‌ی بلمن برای بروزرسانی Q-valueها را نوشته و سپس مقدار بروز شده‌ی $Q(۳, \text{چپ})$ را حساب کنید.

(ب) حال فرض کنید مدل محیط را نداریم و جدول Q-valueهای داده شده از روش یادگیری تفاوت زمانی[۷] بدست آمده است. توضیح دهید چرا در این‌صورت استفاده از سیاست حریصانه هوشمندانه نیست و با برقراری تعادل بین چه مواردی می‌توان سیاست بهتری داشت؟

[۵] state-value function
[۶] policy evaluation
[۷] Temporal Difference Learning

(ج) توضیح دهید چرا به‌جای استفاده از ارزش حالت‌ها یا همان V-values[۱] از ارزش کنش‌ها یا همان Q-values استفاده شده است.

(د) یکی از روش‌های حل مشکل قسمت (ب) استفاده از سیاست تصادفی softmax است. در این روش احتمال انجام دادن کنش a از حالت s که آن را با $\pi(s,a)$ نشان می‌دهیم به صورت زیر محاسبه می‌شود:

$$\pi(s,a) = \frac{e^{Q(s,a)}}{\sum_b e^{Q(s,b)}}$$

با در نظر گرفتن این سیاست و جدول داده شده برای Q-valueها، احتمال انجام هر کنش در حالت‌های مختلف را بدست آورید. همین‌طور توضیح دهید چرا استفاده از این روش معقولانه است و مشکل قسمت (ب) را برطرف می‌کند.

(ه) حال می‌خواهیم با استفاده از الگوریتم SARSA[۸] مقدار Q-valueها را بروزرسانی کنیم. فرض کنید از خانه‌ی ۲ مسیر زیر را نمونه‌برداری کرده‌ایم.

$$۲ \rightarrow ۵ \rightarrow ۶$$

با درنظر گرفتن رابطه‌ی زیر برای بروزرسانی به روش SARSA مقدار $Q(\text{بالا}, ۱)$ و $Q(\text{راست}, ۵)$ را بروزرسانی کنید. $(\alpha = ۰/۲, \gamma = ۰/۸)$

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[R^a_{ss'} + \gamma Q(s',a') - Q(s,a)\right]$$

۴. (۲۰ نمره) یک MDP با دو استیت A و B، با دو اکشن (۱) و (۲)، و استیت ترمینال (T) با $V(T) = ۰$. transition function و reward function ناشناخته است اما نمونه‌های زیر را دیده‌ایم.

(a) $A \rightarrow B:\ a_1 = 1, r_1 = -3$

(b) $B \rightarrow A:\ a_2 = 1, r_2 = 4$

(c) $A \rightarrow A:\ a_3 = 2, r_3 = -4$

(d) $A \rightarrow B:\ a_4 = 1, r_4 = -3$

(e) $A \rightarrow T:\ a_5 = 2, r_5 = 1$

که هر $\rightarrow$ یک تغییر از حالت مبدا به مقصد با انجام action و reward مشخص شده است.

(آ) مقدار $Q(s,a)$ را بعد از مشاهده این نمونه‌ها تعیین کنید.

(ب) یک سیاست deterministic با توجه به سمپل‌ها معرفی کنید که از سیاست رندوم بهتر است. توضیح دهید.

(ج) سیاست رندوم را با π_{random} و سیاست طراحی شده را با π^* نام‌گذاری کنید. چه انتظار در مورد مقدار نهایی value estimation در زمانی که الگوریتم Q-Learning با سیاست π^* شروع شود نسبت به وقتی با π_{random} شروع شود دارید؟ هر کدام از این سیاست‌ها به چه مشکلاتی ممکن است بینجامد؟

۵. (۲۵ نمره) فرض کنید ما با نرخ اکتشاف ϵ شروع می کنیم. به این معنی که هرگاه مدل یک action را انتخاب کند، با احتمال ϵ به صورت تصادفی و با احتمال $۱ - \epsilon$ action انتخاب شده انجام می‌شود. اگر فرض کنیم که محیط به اندازه کافی کاوش شده‌است، ممکن است بخواهیم پس از مدتی میزان اکتشاف را کاهش دهیم. یک الگوریتم برای کاهش این نرخ اکتشاف ارائه دهید. اگر حریف استراتژی‌اش را تغییر دهد، آیا روش شما کار می کند؟ چرا؟ اگر نه، یک heuristic ارائه دهید که بتواند با تغییرات در استراتژی حریف سازگار شود.

[۸] state-action-reward-state-action