دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی

پاییز ۱۴۰۱

استاد: محمدحسین رهبان

گردآورندگان: محمدجواد هزاره‌ـ امیرحسین جوادی‌ـ علی ثالثی‌ـ فریدون مهری

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

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین تا سقف ۱۰ روز و در مجموع ۲۰ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۰ درصد از نمره تمرین به صورت ساعتی کسر خواهد شد.
- 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.

سوالات (۹۰ 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.

حل.

(آ) درست است. هر چه ضریب تخفیف کوچک‌تر باشد تاثیر پاداش‌هایی که در آینده می‌گیریم کم‌تر شده و در نتیجه عامل فقط حال و آینده‌ی نزدیک را برای تصمیم‌گیری‌هایش مورد توجه قرار می‌دهد.

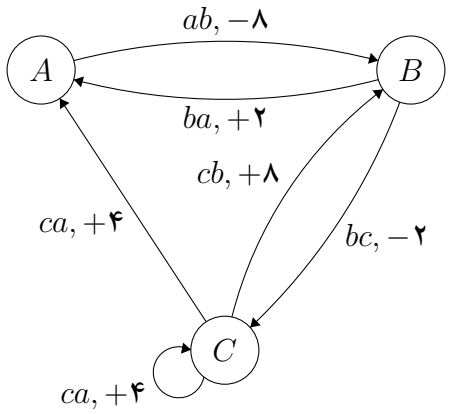
(ب) درست است. هر چه پاداش منفی زندگی از لحاظ اندازه بزرگ‌تر باشد، با زنده ماندن در دنیا عامل پاداش منفی بیش‌تری دریافت می‌کند پس سعی دارد هر چه سریع‌تر کار خود را تمام کند. بنابراین باز هم حال و آینده‌ی نزدیک را معیار تصمیم‌گیری‌هایش قرار خواهد داد.

(ج) نادرست است. با استفاده از ضریب تخفیف کوچک‌تر از ۱ نمی‌توان پاداش‌هایی با مقدار منفی تولید کرد، پس نمی‌توان پاداش منفی زندگی را با استفاده از ضریب تخفیف کوچک‌تر از ۱ مدل کرد. (با فرض ضریب تخفیف منفی نیز در توان‌های فرد مقدار پاداش منفی خواهد شد اما باز هم نمی‌توان پاداش منفی زندگی را با این روش مدل کرد چرا که هر چه دورتر برویم پاداشی که از ضریب تخفیف منفی بدست می‌آید کم‌تر می‌شود اما پاداش منفی زندگی تغییری نخواهد کرد، علاوه بر اینکه در توان‌های زوج مقدار پاداش مثبت خواهد بود و باز هم نمی‌توان پاداش منفی زندگی را مدل کرد.)

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

(د) نادرست است. مشابه قسمت قبل نمی‌توان با استفاده از پاداش منفی زندگی ضریب تخفیف منفی/کوچک‌تر از ۱ را 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.

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 $\pi_۱$ 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 $V_۱(A) = V_۱(B) = V_۱(C) = ۲$ در نظر بگیریم. یک مرحله از الگوریتم ارزیابی سیاست[۶] را اجرا کنید تا به تابع ارزش $V_۲(s)$ برای حالت‌های مختلف برسید.

(د) براساس تابع ارزش‌گذاری جدید و به صورت حریصانه سیاست قطعی جدید $\pi_۲$ 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.

حل.

(آ) تابع ارزش حالت‌ها امید ریاضی پاداشی را که عامل با دنبال کردن سیاست π بدست خواهد آورد محاسبه می‌کند. به نوعی این تابع مشخص می‌کند که هر حالت در فضای حالت‌های مسئله با در نظر گرفتن سیاست π و پیروی کردن از آن خوب و ارزشمند است. همچنین از رابطه‌ی زیر می‌توان این تابع را بدست آورد:

$$V^{\pi}(s) = \mathbb{E}[r_{t+۱} + \gamma r_{t+۲} + \gamma^{۲} r_{t+۳} + \cdots | s_t = s]$$

[۵] state-value function
[۶] policy evaluation

(ب) رابطه‌ی بلمن برای تابع V^π به شکل زیر است:

$$V^\pi(s) = \sum_{s'} T(s, \pi(s), s') \left[R(s, \pi(s), s') + \gamma V^\pi(s') \right]$$

(ج) اگر سیاست ما انتخاب تصادفی بین کنش‌های موجود در یک حالت و تابع ارزش اولیه برای تمام حالت‌ها برابر ۲ باشد داریم:

$$\begin{cases} V_۲(A) = -۸ + ۰/۵ \times ۲ = -۷ \\ V_۲(B) = ۰/۵(۲ + ۰/۵ \times ۲) + ۰/۵(-۲ + ۰/۵ \times ۲) = ۱ \\ V_۲(C) = ۰/۵(۸ + ۰/۵ \times ۲) + ۰/۵(۰/۲۵(۴ + ۰/۵ \times ۲) + ۰/۷۵(۴ + ۰/۵ \times ۲)) = ۷ \end{cases}$$

(د) با استفاده از الگوریتم policy improvement می‌توان سیاست جدید را بدست آورد:

$$\pi_۲(A) = ab$$

$$\left.\begin{aligned} Q(B, ba) &= ۲ + ۰/۵V_۲(A) = -۱/۵ \\ Q(B, bc) &= -۲ + ۰/۵V_۲(C) = ۱/۵ \end{aligned}\right\} \implies \pi_۲(B) = bc$$

$$\left.\begin{aligned} Q(C, ca) &= ۴ + ۰/۵(۰/۲۵V_۲(A) + ۰/۷۵V_۲(B)) = ۵/۷۵ \\ Q(C, cb) &= ۸ + ۰/۵V_۲(B) = ۸/۵ \end{aligned}\right\} \implies \pi_۲(C) = cb$$

(ه) پس از بهبود حریصانه‌ی سیاست خواهیم داشت

$$\pi'(s) = \arg\max_{a \in A} Q^\pi(s, a)$$

از طرفی برای سیاست π داریم

$$V^\pi(s) = Q^\pi(s, \pi(s))$$

بنابراین با توجه به انتخاب π' داریم:

$$\begin{aligned} V^\pi(s) &\le Q^\pi(s, \pi'(s)) \\ &= \mathbb{E}_{\pi'}[r_{t+۱} + \gamma V^\pi(s_{t+۱}) | s_t = s] \\ &\le \mathbb{E}_{\pi'}[r_{t+۱} + \gamma Q^\pi(s_{t+۱}, \pi'(s_{t+۱})) | s_t = s] \\ &= \mathbb{E}_{\pi'}[r_{t+۱} + \gamma \mathbb{E}_{\pi'}[r_{t+۲} + \gamma V^\pi(s_{t+۲}) | s_{t+۱}] | s_t = s] \\ &= \mathbb{E}_{\pi'}[r_{t+۱} + \gamma r_{t+۲} + \gamma^۲ V^\pi(s_{t+۲}) | s_t = s] \\ &\le \mathbb{E}_{\pi'}[r_{t+۱} + \gamma r_{t+۲} + \gamma^۲ Q^\pi(s_{t+۲}, \pi(s_{t+۲})) | s_t = s] \\ &= \mathbb{E}_{\pi'}[r_{t+۱} + \gamma r_{t+۲} + \gamma^۲ r_{t+۲} + \gamma^۲ V^\pi(s_{t+۳}) | s_t = s] \\ &\vdots \\ &\le \mathbb{E}_{\pi'}[r_{t+۱} + \gamma r_{t+۲} + \gamma^۲ r_{t+۲} + \gamma^۳ r_{t+۳} + \cdots | s_t = s] \\ &= V^\pi(s) \end{aligned}$$

و اگر بهبود سیاست متوقف شود و حالت تساوی در نابرابری‌های بالا برقرار باشد داریم:

$$Q^\pi(s, \pi'(s)) = Q^\pi(s, \pi(s)) = V^\pi(s)$$

در نتیجه $V^\pi(s)$ در رابطه‌ی بلمن برای بهینه‌بودن صدق خواهد کرد و بنابراین 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؟

(ج) توضیح دهید چرا به‌جای استفاده از ارزش حالت‌ها یا همان V-values 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.

[۷] Temporal Difference Learning

(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.

$$۲ \to ۵ \to ۶$$

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

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

حل.

(آ) رابطه‌ی بلمن به صورت زیر خواهد بود:

$$Q_{k+۱}(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} Q_k(s',a')\right]$$

اگر بخواهیم $Q(۳, \text{چپ})$ را آپدیت کنیم، با توجه به قطعی بودن کنش‌ها داریم $s = ۳$، $a = \text{چپ}$، $s' = ۲$ و راست $= a'$ که a' با استفاده از جدول و حریصانه عمل کردن بدست آمده است. بنابراین:

$$Q(۳, \text{چپ}) = ۱ \times (-۱ + ۰/۹ \times ۸) = ۶/۲$$

(ب) در حالتی که اطلاعاتی از محیط نداریم استفاده از سیاست حریصانه هوشمندانه نیست چرا که ممکن است تمام محیط به طور کامل دیده نشود و به جواب بهینه‌ی مسئله نرسیم. دنبال کردن سیاست بهینه به ما در انتخاب کنش‌های جدید کمکی نمی‌کند و همین باعث می‌شود بخشی از محیط دیده نشود. برای حل این مشکل باید بین exploration و exploitation تعادل برقرار کنیم که یکی از روش‌ها می‌تواند استفاده از روش ϵ-greedy باشد.

(ج) استفاده از Q-valueها از این جهت سودمند است که می‌توان به راحتی در هر حالتی که هستیم بهترین کنش را انتخاب کرده و سیاست خود را مشخص کنیم. اما با استفاده از V-value علاوه بر اینکه نیاز به محاسبه‌ی زیاد برای پیدا کردن سیاست بهینه داریم ، به مدلی برای احتمالات گذار محیط یا به عبارتی مدلی برای تغییرات محیط نیاز داریم که این می‌تواند چالش برانگیز باشد.

(د) محاسبه‌ی احتمال‌های مربوطه سر راست است. اما چرا این روش در حل مشکل قسمت (ب) مفید است؟ در این روش کنش‌هایی که Q-value کوچکی دارند به طور کامل دور ریخته نمی‌شوند و عامل ممکن است با احتمالی هر چند کوچک این کنش‌ها را انجام دهد. این مسئله به برقراری تعادل بین exploration و exploitation کمک کرده و می‌تواند قسمت‌های ناشناخته‌ی فضای حالت را نیز جست‌وجو کند و به جواب بهینه برسد.

(ه) با استفاده از رابطه‌ی داده شده:

$$\begin{cases} Q(۲, \text{بالا}) = ۶ + ۰/۲ \left[-۱ + ۰/۸ \times ۸ - ۶\right] = ۵/۸۸ \\ Q(۵, \text{راست}) = ۸ + ۰/۲ \left[۱۰ + ۰/۸ \times ۰ - ۸\right] = ۸/۴ \end{cases}$$

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) $A \to B: \ a_1 = 1, r_1 = -3$

[۸] state-action-reward-state-action

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

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

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

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

که هر $\to$ 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؟

حل.

(آ) Q-learning: $Q(A,1) = -0.534, Q(A,2) = -0.26, Q(B,1) = 0.4, Q(B,2) = 0$

(ب) یک سیاست Deterministic می‌تواند به صورت زیر حریصانه باشد:
۲ action $\pi(a|s = A) = \text{argamax}(Q(A, ۱), Q(A, ۲)) =$
۱ action $\pi(a|s = B) = \text{argamax}(Q(B, ۱), Q(B, ۲)) =$

(ج) با π_{random} ما می‌توانیم همواره تخمین‌های واقعی را بدست آوریم ولی همگرایی به تخمین‌های واقعی زمان می‌برد. در حالی که در مورد π^* (سیاست حریصانه) ممکن است به دلیل رفتارهای حریصانه هرگز به تخمین واقعی همگرا نشویم.

سیاست π_{random} از مقادیر یادگرفته شده بهره نمی‌برد و به بیان بهتر Exploit نمی‌کند. از طرفی استفاده از سیاست حریصانه نیز خوب نیست چون سیاست π^* هم از خاصیت Exploring بی‌بهره است. عامل در این حالت کوته‌بین است و پاداش‌های خوب آنی را در نظر می‌گیرد که لزوما منجر به پاداش بیشنه/اپتیمال نمی‌شود.

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.

حل.

این کار از طریق کاهش احتمال Explore کردن یعنی ϵ از طریق تعیین مقدار آن به صورت $\epsilon = \frac{۱}{ct}$ امکان پذیر است که t شماره گام زمانی بوده و c یک ثابت (پیش‌فرض ۰) است.

خیر اگر استراتژی حریف تغییر کند، فرض ما در مورد این که مشاهدات از محیط (که شامل اعمال حریف هم می‌شود) یکسان یا Deterministic باقی می‌ماند معتبر نخواهد بود. یک راه این است که سیاست را به صورت $\epsilon - \text{greedy}$ نگاه داریم یا به شکلی تغییر در رفتار را وارد مدل محیط کرده و براساس آن یادگیری را پیش از انجام حرکت واقعی انجام بدهیم.