HW06 _ Artificial Intelijence

Saeed Razavi

Student ID : 98106542

---

۱. (۱۰ نمره) درستی یا نادرستی گزاره‌های زیر را در رابطه با یک فرآیند تصمیم‌گیری مارکف[1] مشخص کنید و توضیحی کوتاه در رابطه با آن ارائه دهید.

(آ) ضریب تخفیف[2] کوچک و نزدیک به صفر به رفتار حریصانه و کوته‌نظر[3] منجر می‌شود.

(ب) پاداش منفی زندگی[4] 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.

آ) هر چقدر γ کوچک و نزدیک به صفر باشد اثر reward های آینده کمتر می‌شود ← منجر به رفتار حریصانه و کوته‌نظری می‌شود.

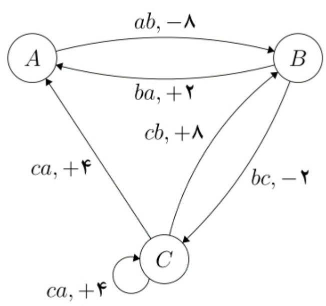
ب) هر چقدر living reward بزرگتر باشد، agent سعی می‌کند زودتر به هدف برسد یا زودتر terminate شود. که در هر دو صورت منجر به رفتار حریصانه و کوته‌نظری می‌شود.

ج) خیر ـ لزوماً نمی‌توان مدل کرد به دلیل اینکه living reward تأثیر خود را به صورت جمع می‌گذارد اما discount reward به صورت ضرب [یک چند حالتی در رتبه] را این می‌توانند باعث شود که شرایطی نتوانیم living reward منفی را با آن مدل کنیم.

د) خیر ـ پاسخ‌ها شبیه قسمت قبل می‌باشد.

---

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.

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

$$V_k^{\pi}(s) = \mathbb{E}\left[r_{k+1} + \gamma r_{k+2} + \dots \mid s_k = s\right] \triangleq \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i r_{k+i} \mid s_t = s\right]$$

توصیف کنید.

(ب) رابطه بلمن برای تابع ارزش حالت‌ها بنویسید.

$$\text{Bellman Eq} \triangleq V_{k+1}(s) = \max_a \sum T(s,a,s')\left[R(s,a,s') + \gamma V_k(s')\right]$$

(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 $V_1(A) = V_1(B) = V_1(C) = ۲$ در نظر بگیریم. یک مرحله از الگوریتم ارزیابی سیاست[۶] را اجرا کنید تا به تابع ارزش $V_2(s)$ برای حالت‌های مختلف برسید.

$$V_2(A) = -8 + 0.5\,V_1(B) = -7$$

$$V_2(B) = 0.5\left[-2 + 0.5 V_1(C) + 2 + 0.5 V_1(A)\right] = 1$$

$$V_2(C) = 0.5\left[4 + 0.5\left(\tfrac{3}{4} V_1(C) + \tfrac{1}{4} V_1(A)\right) + 8 + 0.5 V_1(B)\right] = 7$$

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

$Q_2(B, ba) = 2 + 0.5(-7) = -1.5$

$Q_2(B, bc) = -2 + 0.5(7) = 1.5$

$\Downarrow$

$\pi_2(B) = bc$ (2)

برای A یکی action بیشتر نداریم ← (1) $\pi_2(A) = ab$

$Q_2(C, ca) = 4 + 0.5\left[\tfrac{1}{4} V_2(A) + \tfrac{3}{4} V_2(B)\right] = 5.75$

$Q_2(C, cb) = 8 + 0.5(1) = 8.5 \Rightarrow \pi_2(C) = cb$ (3)

(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.

we know that for $Q^{\pi}(s, \pi'(s)) = \max_a Q^{\pi}(s,a) \geq Q^{\pi}(s, \pi(s)) = V^{\pi}(s)$

$\rightarrow Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s)$ $\xrightarrow{V^{\pi'}(s) = Q(s, \pi'(s))}$ $V^{\pi'}(s) \geq V^{\pi}(s)$

If $V^{\pi'}(s) = V^{\pi}(s) \Rightarrow Q^{\pi'}(s, \pi'(s)) = Q^{\pi}(s, \pi(s))$ $\xrightarrow{Q^{\pi'}(s, \pi'(s)) = \max_a Q^{\pi}(s,a)}$

$$\max_a Q^{\pi}(s,a) = Q^{\pi}(s, \pi(s))$$

$\Rightarrow \pi(s) = \text{optimal policy}$

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؟

آ) $Q_{k+1} = \sum T(s,a,s')[R(s,a,s') + \gamma \max_a Q_k(s',a')]$

$\Rightarrow Q(۳,\text{چپ}) = \frac{۱}{۲}[-۱ + \gamma \max(۵,۳,۸)] = ۸\gamma - ۱$

ب) این کار باعث می‌شود تا خیلی از خانه‌های مسئله explore نشود زیرا ما با استفاده از یک سری randomness تربیت غیر optimal نمیم.

ج) چون می‌خواهیم action بهینه در هر مرحله را حساب کنیم که از آن زیاد مانی نشود. اما طبق رابطه‌ی $\pi(s) = \arg\max_a Q(s,a)$ این policy زیاد می‌شود. در نتیجه محاسبه‌ی Q-value ها منطقی می‌باشد.

د) ۴ | ۵ | ۶
۱ | ۲ | ۳

$s:1 \rightarrow \pi(1,\text{بالا}) = \frac{e^9}{e^4+e^3}$ ، $\pi(1,\text{راست}) = \frac{e^3}{e^3+e^4}$

$s:2 \rightarrow \pi(2,\text{بالا}) = \frac{e^6}{e^6+e^8}$ ، $\pi(2,\text{راست}) = \frac{e^8}{e^6+e^8+e^3}$

$s:3 \rightarrow \pi(3,\text{بالا}) = \frac{e^9}{e^3+e^7}$ ، $\pi(3,\text{چپ}) = \frac{e^7}{e^3+e^7}$

$\varepsilon:4 \rightarrow \pi(4, \text{پایین}) = \frac{e^2}{e^2+e^5}$, $\pi(4, \text{راست}) = \frac{e^5}{e^2+e^5}$

$\varepsilon:5 \rightarrow \pi(5, \text{پایین}) = \frac{e^6}{e^6+e^8+e^5}$, $\pi(5, \text{راست}) = \frac{e^8}{e^6+e^8+e^5}$, $e(5, \text{چپ}) = \frac{e^5}{e^6+e^8+e^5}$

ما باستفاده از این مسئله درباره ی روش exploration را انجام میدهیم.

$Q(1, \text{پایین}) = 4 + 0.2\left[R_{1\to4} + 0.8\, Q(4, a') - Q(1, \text{پایین})\right] = 3.8$ (ف

$Q(5, \text{راست}) \leftarrow Q(5, \text{راست}) + 0.2\left[R_{5\to6} + 0.8 Q(6, a') - Q(5, \text{راست})\right] = 10$

۴. (۲۰ نمره) یک MDP با دو استیت A و B، با دو اکشن (۱) و (۲)، و استیت ترمینال (T) با $V(T) = 0$.
transition function و reward function ناشناخته است اما نمونه‌های زیر را دیده‌ایم.

(a) $A \to B:\ a_1 = 1, r_1 = -3$
(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؟

(الف)

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

Sample

$Q(A,1) \Rightarrow samples = \{a, d\} = \frac{-3-3}{2} = -3$

$Q(A,2) \Rightarrow samples = \{c, e\} = \frac{-4+1}{2} = -1.5$

$Q(B,1) \Rightarrow samples = \{b\} = 4$

$Q(B,2) = 0 \to initial\text{-}value$

(ب)

$V^{\pi}(s) = \max_a Q(s,a)$ مناسب با Q-value های بدست آمده از تست قبل رابطه‌ی زیر را:

$\pi^*(A) = 2$ $\quad$ $\pi^*(B) = 1$ ← سیاست اتخاذ شده

(ج)

← اگر با سیاست π_{random} پیش برویم مطمئنیم که پس از گذشتن تعداد نهایی Iteration، جواب ما به مقدار واقعی converge میشود. هرچند ممکن است زمان بسیاری طول بکشد تا converge شود.

← اگر با π^* شروع کنیم امکان دارد چون یک سیاست را متناسب با نقطه 5، Episode اتخاذ کردیم، exploration به خوبی صورت نگیرد و باعث شود به value های واقعی converged نشود. اما سرعت همگرایی آن بیشتر است.

← مزیت π_{random} ← اطمینان از converge اما کمی time consuming باشد.

← مزیت π^* ← سریعتر converge میکند اما احتمال هست به مقدار واقعی converge نکند

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 $1-\epsilon$ 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.

← برای بخش اول، یک روش naive و ساده این است که یک decay rate انتخاب کنم و هر بار (بعد از k حرکت مختص) ϵ را در روش ϵ-greedy با فرمول $\epsilon \leftarrow \epsilon - \text{decay rate}$ آپدیت کنم.

به ظاهر روش بدی نیست و به خصوص اگر حریف استراتژی خود را تغییر ندهد روش خوبی است.

← برای بخش دوم که حریف استراتژی خود را تغییر می‌دهد بهتر است که دو بار آپدیت کردن میزان error یعنی $e = \sum_s |Q^{\pi}(s) - \text{sample}|^2$ را حساب کنم و اگر دیدیم این میزان error از یک Threshold پایین‌تر است، هر بار ϵ را با فرمول $\epsilon \leftarrow \epsilon + \text{decay rate}$ آپدیت کنم.

← در حقیقت در این روش، هر بار میزان اختلاف میان sample، Q پیش‌بینی شده از قبل را حساب می‌کنم تا هیچ‌گاه از حالت convergence خارج نشویم ← زیرا تغییر حالت convergence به صورت ناگهانی شان از تغییر استراتژی حریف است.