به‌نام خدا

دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

هوش مصنوعی

تمرین ۶: Markov Decision Process & RL

مهراد میلانلو

۹۹۱۰۵۷۷۵

سوال ۱

آ) درست

اگر ضریب تخفیف به مقدار کافی کوچک باشد، ایجنت ترجیح می‌دهد پاداش ابتدایی را جمع کند و حریصانه و کوته‌نظر عمل می‌کند. زیرا پاداش‌های بعدی علی‌رغم بزرگ‌تر بودن، وقتی در γ ضرب شود، می‌تواند کوچک و کوچک‌تر و نزدیک به صفر باشد.

ب) درست

هنگامی که پاداش منفی زندگی به‌اندازه‌ی کافی بزرگ باشد، ایجنت حریصانه عمل می‌کند و ترجیح می‌دهد یا زودتر به نقطه‌ی پایان برسد. زیرا با هر حرکتی که انجام می‌دهد، مقدار زیادی reward منفی دریافت می‌کند.

ج) نادرست

ضریب منفی هنگامی که به‌توان‌های زوج برسد، مثبت و در غیر این‌صورت منفی‌است که نشان می‌دهد نمی‌توانیم پاداش منفی زندگی را با آن مدل کنیم. زیرا یک‌درمیان مثبت و منفی می‌شود.

د) نادرست

دقیقا مشابه استدلال قبل، فرض کنید ضریب تخفیف منفی باشد. بنابراین در توان‌های زوج مثبت و در توان‌های فرد، منفی‌است. چون یکی در میان منفی و مثبت می‌شود، به‌وضوح نمی‌توان آن را با پاداش منفی زندگی مدل کنیم.

سوال ۲

آ) تابع ارزش حالت‌ها را این‌گونه در نظر می‌گیریم:

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

در این رابطه، $T(s, \pi(s), s')$ احتمال آن است که در صورت انتخاب اعمال سیاست $\pi(s)$ از حالت s به حالت s' برسیم. (به دلیل stochastic بودن فرآیند). $R(s, \pi(s), s')$ مقدار reward حاصل از اعمال $\pi(s)$ برای رسیدن از s به s' است. همین‌طور γ مقدار ضریب تخفیف می‌باشد. بنابراین واضح است که در صورت انتخاب سیاست بهینه‌ی π، $V^\pi(s)$ ماکسیمم مقدار utility با شروع از حالت s را نشان می‌دهد.

ب) رابطه‌ای که در قسمت قبل نوشتیم، همان رابطه‌ی بلمن برای تابع ارزش حالت‌هاست که روی سیاست‌های مختلف عمل می‌کند. برای آن که بتوانیم ارزش حالت‌ها را آپدیت کنیم، از روش Time-Limited استفاده می‌کنیم. در واقع داریم:

$$V^\pi_{k+۱}(s) \leftarrow \sum_{s'} T(s, \pi(s), s')[R(s, \pi(s), s') + \gamma V^\pi_k(s')]$$

ج) برای حل این قسمت، ابتدا از روی نمودار فرآیند به‌دست می‌آوریم:

$$T(A, ab, B) = ۱, T(B, ba, A) = ۱, T(B, bc, C) = ۱, T(C, ca, A) = ۰/۲۵, T(C, ca, C) = ۰/۷۵, T(C, cb, B) = ۱$$

$$R(A, ab, B) = -۸, R(B, bc, C) = -۲, R(C, ca, C) = ۴, R(B, ba, A) = ۲, R(C, ca, A) = ۴, R(C, cb, B) = ۸$$

$$\gamma = ۰/۵$$

بنابراین طبق رابطه‌ی قسمت قبل خواهیم داشت:

$$V^{\pi_۱}_۲(A) = \sum_{s'} T(A, \pi_۱(A), s')[R(A, \pi_۱(A), s') + \gamma V^{\pi_۱}_۱(s')] = T(A, ab, B)[R(A, ab, B) + \gamma V^{\pi_۱}_۱(B)]$$

$$= ۱ \times (-۸ + ۰/۵ \times ۲) \Rightarrow \boxed{V_۲^{\pi_۱}(A) = -۷}$$

$$V_۲^{\pi_۱}(B) = \sum_{s'} T(B, \pi_۱(B), s')[R(B, \pi_۱(B), s') + \gamma V_۱^{\pi_۱}(s')]$$

$$= \frac{۱}{۲} T(B, ba, A)[R(B, ba, A) + \gamma V_۱^{\pi_۱}(A)] + \frac{۱}{۲} T(B, bc, C)[R(B, bc, C) + \gamma V_۱^{\pi_۱}(C)]$$

$$= \frac{۱}{۲} \times ۱ \times (۲ + ۰/۵ \times ۲) + \frac{۱}{۲} \times ۱ \times (-۲ + ۰/۵ \times ۲) \Rightarrow \boxed{V_۲^{\pi_۱}(B) = ۱}$$

$$V_۲^{\pi_۱}(C) = \sum_{s'} T(C, \pi_۱(C), s')[R(C, \pi_۱(C), s') + \gamma V_۱^{\pi_۱}(s')]$$

$$= \frac{۱}{۲} \sum_{s'} T(C, ca, s')[R(C, ca, s') + \gamma V_۱^{\pi_۱}(s')] + \frac{۱}{۲} T(C, cb, B)[R(C, cb, B) + \gamma V_۱^{\pi_۱}(B)]$$

$$= \frac{۱}{۲}(\frac{۱}{۴} \times (۴ + ۰/۵ \times ۲) + \frac{۳}{۴} \times (۴ + ۰/۵ \times ۲)) + \frac{۱}{۲} \times ۱ \times (۸ + ۰/۵ \times ۲) \Rightarrow \boxed{V_۲^{\pi_۱}(C) = ۷}$$

د) اکنون با توجه به تابع ارزش‌گذاری جدید، با استفاده از مرحله‌ی extraction policy از رابطه‌ی زیر استفاده می‌کنیم:

$$\pi_{k+۱}(s) \leftarrow \arg\max_a \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V^{\pi_k}(s')]$$

$$\pi_۲(A) = \arg\max_a \sum_{s'} T(A, \pi_۱(A), s')[R(A, \pi_۱(A), s') + \gamma V_۲(s')]$$

$$= \arg\max_a \{۱ \times (-۸ + ۰/۵ \times ۱)\} \Rightarrow \boxed{\pi_۲(A) = ab}$$

$$\pi_۲(B) = \arg\max_a \sum_{s'} T(B, \pi_۱(B), s')[R(B, \pi_۱(B), s') + \gamma V_۲(s')]$$

$$= \arg\max_a \{۱ \times (۲ + ۰/۵ \times ۷), ۱ \times (-۲ + ۰/۵ \times -۷)\} \Rightarrow \boxed{\pi_۲(B) = bc}$$

$$\pi_۲(C) = \arg\max_a \sum_{s'} T(C, \pi_۱(C), s')[R(C, \pi_۱(C), s') + \gamma V_۲(s')]$$

$$= \arg\max_a \{\frac{۱}{۴} \times (۴ + ۰/۵ \times (-۷)) + \frac{۳}{۴} \times (۴ + ۰/۵ \times ۷), ۱ \times (۸ + ۰/۵ \times ۱)\} \Rightarrow \boxed{\pi_۲(C) = cb}$$

ه)

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

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

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

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

تهش به نتیجه می‌رسه احتمالا:) به‌طور شهودی، سیاست π' به‌صورت حریصانه از π به‌دست می‌اید. یعنی از بین اکشن‌ها ماکسیمم می‌گیریم. واضح است که مقدار جدید تابع ارزش بیشتر یا مساوی‌است. در حالت تساوی نیز نشان‌دهنده‌ی این است که π' همان π بهینه است.

سوال ۳

آ)

رابطه‌ی بلمن برای به‌روزرسانی هاQ-Value این‌گونه است:

$$Q(s,a) = \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma \max_{a'} Q(s',a')]$$

اکنون با توجه به این‌که ضریب تخفیف $\gamma = ۰/۹$ است، مقدار به‌روزشده‌ی $Q(۳, Left)$ را محاسبه می‌کنیم:

$$Q(۳, Left) = T(۳, Left, ۲)[R(۳, Left, ۲) + ۰/۹ \max_{a'} Q(۲, a')] = ۱ \times [-۱ + ۰/۹ \times max۳, ۶, ۸] = -۱ + ۷/۲$$

$$\Rightarrow Q(۳, Left) = -۶/۲$$

ب) در روش حریصانه در هر استیتی که باشیم، تنها حالت‌های همسایه که بیشترین مقدار reward را داشته باشند در نظر می‌گیریم در صورتی که ممکن است راهی وجود داشته باشد که در ابتدا پاداش کم و در ادامه پاداش بسیار بیشتری بدهد. از سوی دیگر عدم توجه به پاداش‌های همسایه‌ها و حرکات تصادفی بار محاسباتی را به‌شدت انجام می‌دهد. بنابراین ایجاد تعادل بین exploitation و exploration باعث سیاست‌های بهتر می‌شود.

ج) می‌دانیم برای آپدیت کردن π با استفاده از V^π نیاز به دانستن T و R داریم که هنگام یادگیری مسئله ممکن است به مقادیر آن‌ها دسترسی نداشته باشیم. اما به‌روزرسانی با استفاده از $Q(s,a)$ تنها نیاز به استیت و اکشن نیاز دارد و به همین دلیل از Q-Value ها استفاده می‌کنیم.

د)

در سیاست تصادفی softmax با استفاده از احتمالات محاسبه‌شده که در جدول زیر آورده شده‌است، اکشن مورد نظر را انتخاب می‌کنیم. همان‌طور که گفتیم، ایراد اصلی حریصانه عمل کردن، تصمیمات کوته‌نظرانه و عدم توجه به reward های مراحل جلوتر است. که در این روش این اتفاق نمی‌افتد و در محاسبه‌ی احتمالات، همسایه‌های دیگر نیز در نظر گرفته می‌شوند. بنابراین هم سنگینی محاسباتی روش رندوم را ندارد و سریع‌تر به هدف می‌رسد و هم مشکل حریصانه عمل کردن را ندارد.

$\pi(s,a) = \frac{e^{Q(s,a)}}{\sum_b e^{Q(s,b}}$			
$\pi(۱,U) = \frac{e^۴}{e^۳+e^۴} \approx ۰/۷۳$	$\pi(۱,R) = \frac{e^۳}{e^۳+e^۴} \approx ۰/۲۷$	–	–
$\pi(۲,U) = \frac{e^۶}{e^۶+e^۳+e^۸} \approx ۰/۱۱۸$	$\pi(۲,R) = \frac{e^۸}{e^۶+e^۳+e^۸} \approx ۰/۸۷۶$	–	$\pi(۲,L) = \frac{e^۳}{e^۶+e^۳+e^۸} \approx ۰/۰۰۶$
$\pi(۳,U) = \frac{e^۹}{e^۷+e^۹} \approx ۰/۸۸$	–	–	$\pi(۳,L) = \frac{e^۷}{e^۷+e^۹} \approx ۰/۱۲$
–	$\pi(۴,R) = \frac{e^۵}{e^۵+e^۲} \approx ۰/۹۵$	$\pi(۴,D) = \frac{e^۲}{e^۵+e^۲} \approx ۰/۰۵$	–
–	$\pi(۵,R) = \frac{e^۸}{e^۸+e^۵+e^۶} \approx ۰/۸۴$	$\pi(۵,D) = \frac{e^۶}{e^۸+e^۵+e^۶} \approx ۰/۱۲$	$\pi(۵,L) = \frac{e^۵}{e^۸+e^۵+e^۶} \approx ۰/۰۴$

ه)

با استفاده از رابطه‌ی داده‌شده، برای به‌روز رسانی مقادیر مورد نظر داریم: ($\alpha = ۰/۲$، $\gamma = ۰/۸$)

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

$$Q(۲,U) = Q(۲,U) + ۰/۲[R(۲,U,۵) + ۰/۸ \max_{a'} Q(۵,a') - Q(۲,U)] = ۶ + ۰/۲[-۱ + ۰/۸ \times ۸ - ۶]$$

$$\Rightarrow Q(۲,U) = ۵/۸۸$$

$$Q(۵,R) = Q(۵,R) + ۰/۲[R(۵,R,۶) + ۰/۸ \max_{a'} Q(۶,a') - Q(۵,R)] = ۸ + ۰/۲[۱۰ + ۰/۸ \times ۰ - ۸]$$

$$\Rightarrow Q(۵,R) = ۸/۴$$

سوال ۴

آ)

با استفاده از همان رابطه‌ای که در سوال قبل نوشتیم، مقدار $Q(s,a)$ را بعد از مشاهده‌ی هر نمونه به‌روزرسانی می‌کنیم: $(\alpha = ۰/۱, \gamma = ۰/۹)$

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

a:

$$Q(A,۱) \leftarrow Q(A,۱) + \alpha[R(A,۱,B) + \gamma \max_{a'} Q(B,a') - Q(A,۱)] = ۰ + ۰/۱[-۳ + ۰/۹ \times ۰ - ۰] = -۰/۳$$

b:

$$Q(B,۱) \leftarrow Q(B,۱) + \alpha[R(B,۱,A) + \gamma \max_{a'} Q(A,a') - Q(B,۱)] = ۰ + ۰/۱[۴ + ۰/۹ \times ۰ - ۰] = ۰/۴$$

c:

$$Q(A,۲) \leftarrow Q(A,۲) + \alpha[R(A,۲,A) + \gamma \max_{a'} Q(A,a') - Q(A,۲)] = ۰ + ۰/۱[-۴ + ۰/۹ \times ۰ - ۰] = -۰/۴$$

d:

$$Q(A,۱) \leftarrow Q(A,۱) + \alpha[R(A,۱,B) + \gamma \max_{a'} Q(B,a') - Q(A,۱)] = -۰/۳ + ۰/۱[-۳ + ۰/۹ \times ۰/۴ - ۰/۳] = -۰/۵۳۴$$

e:

$$Q(A,۲) \leftarrow Q(A,۲) + \alpha[R(A,۲,T) + \gamma \max_{a'} Q(T,a') - Q(A,۲)] = -۰/۴ + ۰/۱[۱ + ۰/۹ \times ۰ + ۰/۴] = -۰/۲۶$$

در نهایت مقادیر $Q(s,a)$ برابر است با:

$$Q(A,۱) = -۰/۵۳۴, Q(A,۲) = -۰/۲۶, Q(B,۱) = ۰/۴$$

ب)

سیاست زیر را در نظر می‌گیریم:

$$\pi(s) \leftarrow \arg\max_a Q(s,a)$$

که نتیجه می‌دهد:

$$\pi(A) = ۲, \pi(B) = ۱$$

این سیاست با توجه به مقادیر Q-Value به‌دست آمده، از سیاست تصادفی بهتر است و اکشن‌های بهتری را انتخاب می‌کند.

ج)

استفاده از سیاست π_{random} به وضوح امکان exploration بیشتری می‌دهد و احتمالا به جواب بهینه برای Q-Value ها خواهیم رسید. هرچند که از لحاظ زمانی ممکن است بیشتر طول بکشد. از طرف دیگر چون π^* از نمونه‌ها به‌دست آمده، اگه نمونه‌ها به مقدار میانگین نزدیک باشند و تعدادشان افزایش پیدا کند تغییر خاصی نداشته باشند، طبق محاسبات سیاست بهتری از رندوم است. اگرچه به‌دلیل حریصانه بودن، مانع exploration می‌شویم و ممکن است به جواب کاملا بهینه نرسیم.

سوال ۵

برای آن‌که بتوانیم مقدار نرخ اکتشاف را در طول زمان کاهش بدهیم، می‌توانیم آن را تابعی از زمان تعریف کنیم. مانند توابعی که در روش Simulated Anealing تعریف میکردیم ($\epsilon = e^{\frac{\Delta E}{T}}$) که به‌وضوح با گذر زمان مقدار نرخ اکتشاف کاهش می‌یابد. همین‌طور می‌توان از الگوریتم‌هایی مانند softmax که در سوال‌های قبلی نیز داشتیم، استفاده کنیم. در گذر زمان، اختلاف Q-Value استیت‌هایی که پاداش بیشتری می‌دهند، با دیگر استیت‌ها بیشتر می‌شود و چون موقع به‌روز رسانی، مقدار جدید برابر است با نسبت $Q(s, a)$ به دیگر Q-Value ها، بنابراین احتمال حرکت‌هایی که پاداش بیشتری دارند بیشتر می‌شود.

برای حل مشکل دیگر، دقت کنید با تغییر استراتژی حریف، محیط یعنی استیت‌ها و پاداش‌ها نیز تغییر می‌کنند. در این‌صورت می‌توان از رویکرد featured-based استفاده کرد. به این صورت که توابع ارزش‌گذاری را بر اساس تعدادی feature تعریف می‌کنیم:

$$V(s) = \sum_i w_i f_i(s), Q(s, a) = \sum_i w_i f_i(s, a)$$

برای به دست آوردن وزن‌ها نیز، می‌توانیم از روش‌های لرنینگ مانند Gradient-Descent استفاده کنیم. به این صورت با تغییر وزن‌ها می‌توان عملکرد مناسبی دربرابر تغییر استراتژی حریف داشت.