دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی

زمستان ۱۴۰۱

استاد: محمدحسین رهبان
گردآورندگان: حسین گلی، محمد مشتاقی و علی ثالثی

تمرین سوم — **فرآیندهای مارکوف و یادگیری تقویتی** — مهلت ارسال: ۲۹ فروردین

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین تا سقف ۷ روز و در مجموع ۱۵ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۲ 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 نظری (۱۵۰ نمره)

۱. (۲۰ نمره) موارد صحیح و غلط را مشخص کنید و در صورت غلط بودن استدلال کنید.

الف) در مسائل یادگیری تقویتی عامل لزوما باید با محیط به صورت برخط تعامل کند و با استفاده از پاداش هایی که به صورت برخط می‌گیرد سیاست خود را بهبود دهد.

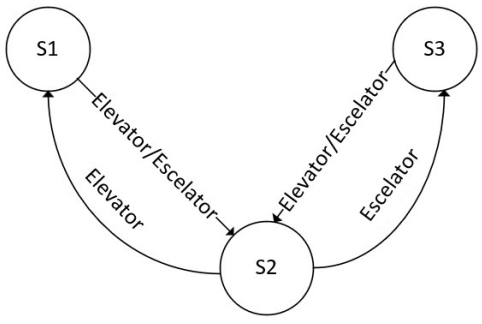
ب) سیاست بهینه در MDP یکتا است.

ج) الگوریتم REINFORCE یک الگوریتم Model Based است.

د) در یک Infinite Horizon MDP به ازای تمامی سیاست و حالات داریم: $V^*(s) \geq V^\pi(s)$

۲. (۳۰ نمره) سو که در یک فروشگاه زنجیره‌ای گم شده است و با ۲ تصمیم مواجه است. به سمت بلوک شرقی یا بلوک غربی فروشگاه برود. برای فهم بهتر مسئله به MDP زیر توجه کنید که حالت‌ها و اکشن‌های مسئله را بهتر توصیف می‌کند. $S_۲$ مکان فعلی او در فروشگاه است. $S_۱, S_۳$ به ترتیب حالات رفتن به بلوک شرقی و غربی هستند. همچنین او اکشن‌های پله برقی و آسانسور را در اختیار دارد.

شکل ۱: MDP



در واقع ما فرض کرده‌ایم سو مدت طولانی در فروشگاه سرگردان خواهد بود و مسئله را عملا می‌توان با یک Infinite Horizon MDP مدل‌سازی کرد.

(آ) فرض کنید جدول Q-values به شما داده شده است. مطابق جدول زیر سیاست بهینه را بدست آورید.

Q-values	Elevator	Escalator
$S_۱$	$+۱/۵$	$-۰/۵$
$S_۲$	$+۲/۰$	$-۰/۳$
$S_۳$	$+۰/۸$	$+۰/۹$

(ب) یک مدل‌سازی ساده می‌تواند MDP دو حالته که سو در حالت $S_۲$ است یا خیر باشد. برای این MDP یک state-diagram رسم کنید. فرض کنید با اجرا الگوریتم Q-learning به مقادیر زیر برای Q-values رسیده‌ایم.

Q^*	Elevator	Escalator
$S_۲$	$+۰/۰۸$	$+۲/۰۸$
$S_۴$	$+۱/۱۸$	$+۱/۳۸$

(ج) سیاست بهینه را بدست آورید.

(د) به چه دلیل سیاست بهینه در این مدل‌سازی با مدل‌سازی قبلی تفاوت کرده است؟ راهنمایی: از فرض مارکو کمک بگیرید.

۳. (۵۰ نمره) در بسیاری از کابردهای یادگیری تقویتی ما نمی‌توانیم داده با استفاده از سیاست‌های مختلف جمع‌آوری کنیم (مانند سیستم‌های درمانی).

به همین دلیل می‌خواهیم در این سوال روش‌های off policy evaluation را باهم بررسی کنیم. فرض کنید در مسئله ما سیاست احتمالاتی است.

$$\pi(s,a) = P(a|s,\pi)$$

تابع هدف ما $R(s,a)$ با استفاده از یک سیاست $\pi_۱$ به‌جای $\pi_.$ است.

$$\begin{aligned}\mathbb{E}_{s\sim p(s),a\sim\pi_۱(s,a)}R(s,a) &= \sum R(s,a)p(s)p(a|s)\\ &= \sum R(s,a)p(s)\pi_۱(a|s)\end{aligned}$$

مشکل ما اینجا است که با روش‌هایی که در کلاس بررسی کردیم (روش‌های نمونه‌گیری) می‌توانیم R را برای سیاست $\pi_.$ به‌دست آوریم. می‌خواهیم روش‌های که به ما این قابلیت را می‌دهد با استفاده از داده‌های جمع‌آوری شده از سیاست $\pi_.$ تابع R را برای $\pi_۱$ تخمین بزند استفاده کنیم.

(آ) تخمین‌گر زیر را در نظر بگیرید:

$$\mathbb{E}_{s\sim p(s),a\sim\pi_.(s,a)}\frac{\pi_۱(s,a)}{\hat{\pi}_.(s,a)}R(s,a)$$

نشان دهید اگر $\pi_. = \hat{\pi}_.$ باشد، این تخمین‌گر سازگار است.

(ب) مورد بالا را برای تخمین‌گر زیر نشان دهید.

$$\frac{\mathbb{E}_{s\sim p(s),a\sim\pi_.(s,a)}\frac{\pi_۱(s,a)}{\hat{\pi}_.(s,a)}R(s,a)}{\mathbb{E}_{s\sim p(s),a\sim\pi_.(s,a)}\frac{\pi_۱(s,a)}{\hat{\pi}_.(s,a)}}$$

(ج) مثالی نقضی برای این‌که نشان دهد تخمین‌گر بالا می‌تواند بایاس باشد ارائه کنید.

(د) اختیاری: درباره تخمین‌گر Doubly Robust تحقیق کنید و برای چه از این تخمین‌گر به جای تخمین‌گرهایی که بررسی کردیم استفاده می‌شود.

۴. (۵۰ نمره) حسین که به تازگی به ماشین زمان دست یافته است به چین قدیم سفر کرده و یک فرمانده محلی شده است. تابع پاداش به‌ازای هر حالت را در جدول زیر برای فرمانروایی حسین مشخص کرده‌ایم. (هر حالت مکان فرمانروایی اوست)

State	Reward
Mountain	۱/۰
Riverside	+۲/۰
Desert	−۱/۰

همچنین در جدول زیر transition function توسط مشاور کاردرستش به او داده شده است.

Start State	Action	Probability	State End
Mountain	Peace	۰/۵	Riverside
Mountain	Peace	۰/۵	Desert
Mountain	War	۰/۱	Mountain
Mountain	War	۰/۲	Desert
Mountain	War	۰/۷	Riverside
Riverside	Peace	۱/۰	Riverside
Riverside	War	۰/۲	Riverside
Riverside	War	۰/۸	Mountain
Desert	Peace	۱/۰	Desert
Desert	War	۱/۰	Mountain

(آ) حسین تصمیم گرفته است که صلح بهترین رویکرد برای موفقیت است. از این رو سیاست او به این صورت است که هیچ‌گاه به جنگ نمی‌رود. تابع ارزش این سیاست را به‌دست آورید. (بر حسب γ)

(ب) حسین پس از شکست‌های پی‌درپی متوجه شده است که باید سیاستش را عوض کند. با استفاده از الگوریتم Policy iteration یک گام سیاست حسین را بهبود دهید. (می توانید فرض کنید $\gamma = ۰/۹$ است)

(ج) مشاور حسین خائن بوده است. او دیگر از transition function هایی که او داده است نمی‌تواند استفاده کند و مجبور است از الگوریتم‌های RL استفاده کند. با توجه به عملکرد اخیر حسین Q Table زیر را پر کنید. (فرض کنید $\gamma = ۱, \alpha = ۰/۵$)

Start State	Action	State End	Reward
Desert	War	Desert	−۲/۰
Desert	War	Riverside	۳/۰
Riverside	Peace	Mountain	۱/۰
Mountain	Peace	Riverside	۱/۰

Mountain-peace	Riverside-peace	Desert-war
۰	۰	۰

۵. (اختیاری) پاسخ به این سوال اختیاری است. حل آن در این لینک قرار دارد. پیشنهاد می‌کنیم آن را مشاهده کنید.

در این تمرین ما با استفاده از cauchy sequence می‌خواهیم اثبات کنیم value iteration به یک جواب یکتا مستقل از نقطه شروع همگرا می‌شود.

Bellman Backup Operator مطابق زیر تعریف می‌شود.

$$V_{k+1} = BV_k = max(R(s,a) + \gamma \sum p(s'|s,a)V_k^{\pi}(s')$$

همچنین می‌دانیم که این اپراتور یک Contraction mapping است.

$$||BV' - BV''||_\infty \leq \gamma ||V' - V''||_\infty$$

(آ) اثبات کنید $||V_{n+1} - V_n||_\infty \leq \gamma^n ||V_1 - V_0||_\infty$

(ب) اثبات کنید به ازای هر $c > 0$ $||V_{n+c} - V_n||_\infty \leq \frac{\gamma^n}{1-\gamma} ||V_1 - V_0||_\infty$ برقرار است.

(ج) cauchy sequence یک سری است که عناصر آن به با جلو رفتن سری به یک دیگر نزدیک می‌شوند. درواقع به ازای هر $\epsilon > 0$ یک k وجود دارد به طوری که اگر $m, n > k$ آنگاه $d(a_m, a_n) < \epsilon$ و در نتیجه سری همگرا است. با استفاده از توصیف گفته‌شده نشان دهید سری $V_1, V_2, V_3, ...$ همگرا است.

(د) نشان دهید که جواب Value iteration یکتا است.

سوالات عملی (۲۰۰ نمره)

۱. (۱۰۰ نمره) برای حل این سوال به نوت‌بوک Q_Tabular.ipynb مراجعه کنید.

۲. (۱۰۰ نمره) برای حل این سوال به نوت‌بوک RL_Chat.ipynb مراجعه کنید. پیشنهاد می‌کنیم از google colab استفاده کنید.