به نام خدا

دانشکده‌ی مهندسی کامپیوتر

یادگیری تقویتی

نیم‌سال بهار ۴۰۲-۱۴۰۱

اساتید: دکتر رهبان، آقای حسنی

تمرین سری چهارم | **مباحث پیشرفته در یادگیری تقویتی** | زمان تحویل: ۲۲ خرداد

لطفا نکات زیر را رعایت کنید:

- سوالات خود را از طریق پست مربوط به تمرین در Quera مطرح کنید.
- پاسخ ارسالی واضح و خوانا باشد.
- در هر کدام از سوالات، اگر از منابع خاصی استفاده کرده‌اید، آن را ذکر کنید.
- اگر با افرادی همفکری کرده‌اید، نام ایشان را ذکر کنید.
- پاسخ ارسالی باید توسط خود شما نوشته شده باشد. به اسکرین‌شات از منابع یا پاسخ افراد دیگر نمره‌ای تعلق نمی‌گیرد.
- تمام پاسخ‌های خود را در یک فایل با فرمت RL_HW#_[SID]_[Fullname].zip روی کوئرا قرار دهید.
- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. علاوه بر آن، در هر تمرین می توانید تا سقف پنج روز از تأخیر مجاز باقیمانده‌ی خود استفاده کنید.

سوال ۱: کشف مهارت (۳۰ نمره)

در این سوال می‌خواهیم به بررسی روش‌های کشف مهارت و چگونگی کارکرد آن‌ها بپردازیم. استفاده از این روش‌ها که مبتنی بر معیارهایی بر پایه نظریه یادگیری اطلاعات هستند، منجر به کشف مهارت‌های مختلف توسط عامل یادگیرنده بدون در نظر گرفتن تابع پاداش مجزا می‌شود. یکی از روش‌های ارائه شده برای این امر روش DIAYN می‌باشد. در این سوال به بررسی بیشتر این روش و روش‌های مشابه خواهیم پرداخت. فرض کنید با در نظر گرفتن شروط زیر می‌خواهیم یک معیار برای کشف مهارت‌های جدید و متنوع در نظر بگیریم:
- مهارت‌های مختلف برای اینکه قابل تمییز از یکدیگر باشند، باید منجر به مشاهده حالت‌های مختلفی از محیط شوند.
- برای تمییز بین مهارت‌های مختلف ما تنها به حالت‌های محیط نیاز داریم نه عمل‌هایی که توسط عامل انجام می‌شود.
- مهارت‌های مختلف تا حد امکان باید متفاوت از یکدیگر باشند.

در نظر داشته باشید که S و A به عنوان یک متغیر تصادفی که به ترتیب نشانگر حالت‌های محیط و عمل‌هایی که عامل انجام می‌دهد، هستند. همچنین $Z \sim p(z)$ به عنوان یک متغیر latent می‌باشد، که ما به سیاستی که مشروط به یک z ثابت باشد مهارت می‌گوییم. همچنین $H[.]$ نشانگر Shannon Entropy و $I(.;.)$ را به عنوان Mutual Information با پایه e در لگاریتم در نظر بگیرید.

(آ) توضیح دهید هر کدام از ترم‌های زیر نشانگر چه چیزی هستند؟

۱. $I(S;Z)$

۲. $I(A;Z|S)$

۳. $H(A|S)$

(ب) با در نظر گرفتن شروط گفته شده می‌توان تابع هدف زیر را برای کشف مهارت‌های مختلف در نظر گرفت:

$$F(\theta) \triangleq I(S;Z) + H(A|S) - I(A:Z|S) \quad (۱)$$

توضیح دهید که معیار در نظر گرفته شده چگونه شروط گفته شده را در بر می‌گیرد.

(ج) تابع هدف گفته شده را تا حد امکان بسط دهید و توضیح دهید ترم‌های معیار نهایی به دست آمده نشان دهنده چه چیزی هستند؟

(د) با در نظر گرفتن $q_\phi(z|s)$ به عنوان تقریبی از $p(z|s)$ و با استفاده از نامساوی Jensen

$$E[f(x)] \geq f(E[x]) \quad (۲)$$

یک حد پایین برای تابع هدف به دست آمده در قسمت قبل به دست آورید.

(ه) مصالحه بین exploration و discrimination را در حد پایین به دست آمده شرح دهید.

(و) با در نظر گرفتن این نکته که بیشینه کردن حد پایین منجر به بیشینه کردن تابع هدف اصلی خواهد شد، و همچنین با در نظر گرفتن استفاده از الگوریتم‌هایی مانند SAC برای یادگیری، به نظر شما از جه تابع پاداشی به جای تابع پاداش تسک موجود می‌توان استفاده کرد؟

(ز) یک تابع هدف دیگر برای فراگیری مهارت‌ها در محیط که در روش DADS استفاده شده است، بیشینه کردن عبارت زیر است:

$$I(s^{'}; z|s) \tag{۳}$$

با بسط دادن عبارت اشاره شده، مفهوم این تابع هدف و تفاوت آن با تابع هدف در قسمت ب را توضیح دهید.

سوال ۲: یادگیری سلسله مراتبی (۳۰ نمره)

هنگام استفاده از Options Framework در روش‌های یادگیری سلسله مراتبی به جای MDP مسئله را با Semi-MDP مدل‌سازی می‌کنند. در نظر داشته باشید که برای مدل کردن این مسائل و اجرای الگوریتم‌هایی مانند value iteration همواره به اطلاعات زیر نیاز است.
- $\varepsilon(o; s; t)$ نشان دهنده رخداد انتخاب Option به اسم o در حالت s و در زمان t می‌باشد.
- احتمال جابجایی بین حالت‌ها از رابطه زیر پیروی می‌کند که $p(s^{'}, k)$ بیانگر احتمال اتمام Option به اسم o در حالت s پس از k گام می‌باشد

$$p^{o}_{s^{'}s} = \Sigma^{\infty}_{k=1} p(s^{'}, k)\gamma^{k} \tag{۴}$$

- برای هر حالت $s \in S$ پاداش مدل از رابطه زیر پیروی می‌کند. در این رابطه $t + k$ نشانگر زمان پایان Option به اسم o است

$$r^{o}_{s} = E[r_{t+1} + \gamma r_{t+2} + ... + \gamma^{k-1} r_{t+k} | \varepsilon(o; s; t)] \tag{۵}$$

- همچنین در نظر داشته باشید که سیاست انتخاب کننده Option در هر حالت است $\mu(s, o)$ بیانگر احتمال انتخاب یک Option مانند o در حالت s می‌باشد.
با در نظر گرفتن فرض‌های بالا به سوالات زیر پاسخ دهید:

(آ) دلیل استفاده از semi-mdp هنگامی که می‌خواهیم temporal abstraction داشته باشیم چیست؟

(ب) با در نظر گرفتن فرض‌های گفته شده، معادله‌ی بلمن را برای ارزش هر حالت در semi-mdp به دست آورید.

(ج) برای قسمت‌های ب و ج معادلات بهینگی بلمن را به دست آورید.

سوال ۳: مروری بر یادگیری تقویتی معکوس (۲۵ نمره)

می‌دانیم که یکی از روش‌های یادگیری تقلیدی تابع سیاست، یادگیری تقویتی معکوس است. در این روش به طور کلی سعی داریم ابتدا یک تابع پاداش را با استفاده از رفتارهای خبره یاد بگیریم، و سپس از آن برای یادگیری تابع سیاست استفاده کنیم. به سوالات زیر پاسخ دهید:

(آ) چرا به جای یادگیری مستقیم سیاست از رفتار خبره به سراغ یادگیری تابع پاداش می‌رویم؟ آیا این مسئله به اندازه‌ی کافی خوش تعریف است؟ توضیح دهید و مثال بزنید.

(ب) فرض کنید برای شروع یادگیری، تابع پاداش از فرمول خطی زیر استفاده کنیم

$$r_{\psi}(s, a) = \sum_{i=1}^{n} \psi_i f_i(s, a) \tag{۶}$$

که در آن ψ_iها پارامترهای تابع پاداش هستند و f_iها یک سری ویژگی مشخص‌اند. اگر π^{r_ψ} را سیاست بهینه برای r_ψ فرض کنیم، می‌توان ψ را طوری انتخاب کرد که تساوی زیر برقرار باشد:

$$\mathbb{E}^{r_\psi}_{\pi}[f(s, a)] = \mathbb{E}_{\pi^*}[f(s, a)] \tag{۷}$$

۱. تفسیر این شیوه‌ی انتخاب ψ چیست؟

۲. این روش چه مشکلاتی دارد؟

۳. برای بهبود این مشکلات، تابع هدف زیر پیشنهاد داده شده است. نحوه‌ی عملکرد آن را توضیح دهید. مشکلات این تابع هدف چیست؟

$$\max_{\psi,m} \quad m \tag{۸}$$

$$s.t. \quad \psi^T \mathbb{E}_{\pi^*}[f(s,a)] \geq \max_{\pi} \psi^T \mathbb{E}_{\pi}[f(s,a)] + m \tag{۹}$$

(ج) یک روش دیگر برای حل مشکلات مذکور، یادگیری تابع پاداش با استفاده از متغیر بهینگی در مدل گرافی است. توزیع متغیر بهینگی با فرض پارامتریزه بودن تابع پاداش بصورت زیر تعریف می‌شود :

$$p(\mathcal{O}_t|s_t,a_t) = \exp\left(r_\psi(s_t,a_t)\right) \tag{۱۰}$$

۱. با فرض این که مسیرهای $\tau_1, ..., \tau_N$ از رفتار خبره را در اختیار داریم، نشان دهید که تابع هدف بیشینه‌سازی درست‌نمایی رفتارهای خبره، L_ψ، در معادله‌ی (۱۲) صدق می‌کند:

$$L_\psi := \frac{1}{N}\sum_{i=1}^{N} \log p(\tau_i|\mathcal{O}_{1:T},\psi) \tag{۱۱}$$

$$\max_{\psi} \quad L_\psi = \max_{\psi} \quad \frac{1}{N}\sum_{i=1}^{N} r_\psi(\tau) - \log Z \tag{۱۲}$$

که در آن داریم

$$Z := \int p(\tau)\exp\left(r_\psi(\tau)d\tau\right) \tag{۱۳}$$

و منظور از $r_\psi(\tau)$ جمع پاداش مسیر است.

۲. ضرورت وجود $\log Z$ در رابطه‌ی بالا چیست و چه تأثیری در انتخاب سیاست بهینه دارد؟

۳. حال نشان دهید که رابطه‌ی زیر برقرار است:

$$\nabla_\psi L_\psi = \mathbb{E}_{\tau\sim\pi^*(\tau)}[\nabla_\psi r_\psi(\tau)] - \sum_{t=1}^{T}\int\int \mu(s_t,a_t)\nabla_\psi r_\psi(s_t,a_t)ds_t da_t \tag{۱۴}$$

که در آن $\mu(s_t,a_t) \propto \beta(s_t,a_t)\alpha(s_t,a_t)$ است و نمادهای α و β به ترتیب متغیرهای رو به جلو و رو به عقب در تحلیل مدل گرافی رفتار بهینه هستند و در اسلایدهای مبحث RL as inference بررسی شده‌اند.

۴. مزیت رابطه‌ی (۱۴) نسبت به رابطه‌ی (۱۲) چیست؟

۵. با توجه به نحوه‌ی محاسبه‌ی بخش دوم عبارت سمت راست، تنها در فضای حالت و کنش محدود می‌توانیم محاسبات سرراستی داشته باشیم. همچنین نحوه‌ی محاسبه‌ی متغیرهای رو به جلو و رو به عقب به گونه‌ای است که نیاز به دانستن دینامیک محیط داریم. توضیح دهید که چگونه می‌توان با Importance Sampling بر این مشکلات غلبه کرد. وزن‌های مورد نظر را با ذکر دقیق روابط بدست آورید.

سوال ۴: یادگیری برون خط (offline) (۲۵ نمره)

با توجه به اینکه در یادگیری برون خط از یک دیتاست ثابت $D = \{(s_t^i, a_t^i, s_{t+1}^i, r_t^i)\}$ برای آموزش یک عامل استفاده می‌کنیم، یکی از مشکلاتی که ممکن است در هنگام تست کردن عامل در تعامل با محیط رخ دهد، مسئله drift data می‌باشد. این مسئله زمانی رخ می‌دهد که عامل در حالاتی از محیط مانند $s \in S$ قرار بگیرد که اطلاعات قابل تعمیمی نسبت به این حالات جدید در دیتاستی که از آن برای آموزش استفاده شده، موجود نباشد یا مدل آموزش داده شده در تعمیم ضعیف باشد. انتخاب عمل اشتباه در این حالت جدید منجر به اتخاذ یک سیاست غلط در ادامه trajectory و ناکامی در به دست آوردن سیاست بهینه می‌شود. در مسائل یادگیری از این مشکل با عنوان shift distribution نیز یاد شده است. در این سوال می‌خواهیم آنالیزی بر روی این قبیل خطاها داشته باشیم. برای این منظور فرض کنید تابع هزینه ما معادل تابع زیر تعداد عمل‌های اشتباهی است که عامل یادگیرنده در یک اپیزود با افق H انجام می‌دهد:

$$\delta(a,a^*) = \begin{cases} 0 & \text{if } a \in \pi^*(s) \\ 1 & Otherwise \end{cases} \tag{۱۵}$$

در این صورت تابع هدف ما معادل کمینه کردن رابطه زیر خواهد بود که معادل کمینه کردن امید ریاضی تعداد دفعات انتخاب اعمال غیر بهینه در هر اپیزود با افق H می‌باشد:

$$\mathcal{L}(\pi) = \mathbb{E}_{\rho_\pi(\tau)}[\sum_{t=0}^{H} \delta(a_t, a_t^*)] \tag{۱۶}$$

و با فرض اینکه احتمال انتخاب عمل غیر بهینه در یک حالت (خطای تعمیم) حداکثر برابر با ϵ باشد، یعنی در هر حالت عامل حداکثر با احتمال ϵ عمل غیر بهینه را انتخاب می‌کند که می‌تواند منجر به رفتار غیر بهینه و انحراف عامل از مسیر بهینه تا انتهای اپیزود گردد. همچنین عامل با احتمال $1-\epsilon$ عمل بهینه را در هر حالت برمی‌گزینند. این نکته را می‌توان در قالب رابطه زیر در نظر گرفت:

$$\pi(a \neq a^*|s) \leq \epsilon \tag{۱۷}$$

با در نظر گرفتن فرض‌های گفته شده و همچنین این نکته که $\pi(a|s)$ از آموزش یک مدل Learning Supervised یعنی emprical standard minimization risk بر روی دیتاست D به دست آمده است، به سوالات زیر پاسخ دهید.

(آ) ثابت کنید در صورت آموزش مدل در حالت گفته شده خطای یادگیری دارای حد بالای $O(H^2\epsilon)$ خواهد بود.

(ب) اگر فرض کنیم که عامل امکان اضافه کردن دیتای trajectory جدید خود که از تعامل با محیط به صورت on policy به دست آورده است را به دیتاست D دارد و هر حالت جدید $s \sim d^\pi(s)$ به همراه عمل بهینه آن حالت به دیتاست D اضافه گردد آنگاه ثابت کنید حد بالا برای خطای یادگیری $O(H\epsilon)$ خواهد بود.

سوال ۵: سناریوهای برخط و برون‌خط برای SAC (۳۰ نمره)

در این تمرین قصد داریم یک عامل SAC را در محیط `CartPole-v1` آموزش دهیم. برای این کار، هم در تنظیمات برخط و هم در تنظیمات برون‌خط عملکرد عامل را خواهیم سنجید. لطفا نوت‌بوک Soft_Actor_Critic.ipynb را طبق توضیحات و با رعایت قالب پیشنهادی تکمیل کنید. می‌توانید مقادیر هایپرپارامترها را برای رسیدن به نتیجه‌ی بهتر تغییر دهید. عمده‌ی نمره به پیاده‌سازی صحیح تعلق می‌گیرد.