به نام او...

تمرین سوم : مسائل Model Free

سؤال یک

فرض کنید که یک عامل، یک مسئله MDP را که در تعامل با انسان است، با استفاده از Q-learning حل می‌کند. با توجه به اینکه در مسائل انسان در حلقه[1]، به طور معمول پاداش به صورت تُنُک[2] داده می‌شود، این مورد چه تاثیری در نرخ همگرایی دارد؟ چه روش‌هایی را برای حل مشکلات ناشی از تنک بودن پاداش پیشنهاد می‌کنید؟ دیگر مشکلات مربوط به مسائل انسان در حلقه چیست و چه راه‌حلی برای آن‌ها پیشنهاد می‌کنید؟

سؤال دو

دو الگوریتم SARSA و Expeted-SARSA را از نظر روش به‌روزرسانی مقدار Q مقایسه کنید. از لحاظ پایداری و سرعت همگرایی، کدام الگوریتم عملکرد بهتری دارد؟ استفاده از هر الگوریتم در چه شرایطی بهتر است؟

سؤال سه

الگوریتم‌های on-policy و off-policy، از منظر بهره‌وری نمونه[3]، اکتشاف[4] محیط و مقاومت[5] در برابر تصادفی بودن محیط، چه تفاوت‌هایی دارند؟ با توجه به این تفاوت‌ها، در مسائل دنیای واقعی، کدام الگوریتم عملکرد بهتری دارد؟ (یک مسئله خاص در دنیای واقعی را در نظر بگیرید و بر روی آن بحث کنید)

سؤال چهار

در این سوال قصد داریم از کتابخانه gymnasium استفاده کنیم. برای این مسئله، از محیط Cliff Walking استفاده می‌کنید. ابتدا لینک‌های قرار داده شده در بالا را مشاهده کنید تا با کتابخانه و محیط مورد استفاده آشنا شوید. سپس به سوالات داده شده پاسخ دهید.

نکات پیاده‌سازی

- سیاست مورد استفاده برای عامل را epsilon-greedy در نظر بگیرید.
- در تمامی سوالات به جز ذکر صریح در صورت سوال مقدار اپسیلون را به صورت کاهشی مناسب و مقدار discount factor را 0.9 در نظر بگیرید. همچنین مقدار نرخ یادگیری را برابر 0.1 در نظر بگیرید.

---

[1] Human in the loop
[2] Sparse
[3] Sample efficiency
[4] Exploration
[5] Robustness

- دقت شود که پارامترهای داده شده صرفا به عنوان یک گزینه‌ی اولیه بوده و ممکن است پارامترها را بتوان طوری تنظیم کرد که یادگیری بهتر شود. در صورتی که در صورت سوال به صورت مستقیم قید نشده باشد، شما میتوانید این پارامترها را تغییر دهید.
- در صورتی که پارامتری برای حل سوال مشخص نشده‌است، با ذکر دلیل مشخص کنید که آن پارامتر را چگونه انتخاب کرده‌اید.

سوالات پیاده‌سازی

الف) الگوریتم Q-learning را یکبار به ازای نرخ یادگیری 0.1 و بار دیگر به ازای نرخ یادگیری کاهشی پیاده‌سازی نمایید و نتایج بدست آمده را از حیث میزان حسرت (سرعت همگرایی و مقدار همگرا شده) با یکدیگر مقایسه کنید. روش انتخابی خود برای کاهش مقدار اپسیلون در طی فرآیند یادگیری را توضیح دهید.

ب) الگوریتم‌های Sarsa و Tree Backup n-step را به ازای سه مقدار n پیاده سازی کنید و نتایج بدست آمده را از حیث میزان حسرت (سرعت همگرایی و مقدار همگرا شده) با یکدیگر مقایسه کنید و در تحلیل نتایج علت عملکرد بهتر به ازای یک مقدار n مشخص را تحلیل نمایید.

اگر رقم آخر شماره دانشجویی شما زوج است:

ج) با استفاده از روش On-Policy MC مسئله را حل کنید و موارد خواسته شده را یک بار برای اپسیلون کاهشی و همچنین برای اپسیلون 0.1 انجام دهید و نتایج به دست آمده را از منظر میزان حسرت (سرعت همگرایی و مقدار همگرا شده) با یکدیگر مقایسه کنید.

اگر رقم آخر شماره دانشجویی شما فرد است:

ج) با استفاده از روش Off-Policy MC مسئله را حل کنید و موارد خواسته شده را یک بار برای اپسیلون کاهشی و همچنین برای اپسیلون 0.1 انجام دهید و نتایج به دست آمده را از منظر میزان حسرت (سرعت همگرایی و مقدار همگرا شده) با یکدیگر مقایسه کنید.

نکات تمرین

- استفاده از LLM ها در این تمرین مشکلی ندارد. اما در صورت استفاده لطفاً منبع و prompt خود را ذکر نمایید تا تقلب محسوب نشود.
- مهلت ارسال این تمرین تا پایان روز یکشنبه ۱۸ آذرماه خواهد بود.
- انجام این تمرین به‌صورت یک‌نفره است. اما بحث و گفت‌وگو در پیامرسان درس مانعی ندارد.
- برای سوالات پیاده‌سازی، حتما گزارش کامل در خصوص کد پیاده‌سازی شده را قرار دهید. همچنین تحلیل نمودارهای به دست آمده را در گزارش ذکر کنید.
- لطفاً گزارش و کد تمرین را در قالب یک فایل zip در سامانه ایلرن بارگذاری کنید.
- برای قسمت‌های مختلف کد، حتما از کامنت‌های مناسب برای توضیح استفاده کنید.
- در صورت وجود سؤال و یا ابهام می‌توانید از طریق پیام‌رسان درس با دستیاران آموزشی در ارتباط باشید