بسمه تعالی

تمرین چهارم

یادگیری تقویتی[1]

استاد درس: دکتر آرش عبدی هجراندوست

تدریسیار آموزشی: آرمین توکلی

✓ نکات تمرین :

- مهلت تحویل ۱۴۰۲/۰۲/۲۶ ساعت ۲۳:۵۵
- مهلت ارسال به هیچ وجه قابل تغییر نیست .
- مواردی که بعد از تاریخ فوق ارسال شوند قابل قبول نبوده و نمره ای نخواهند داشت .
- انجام تمرین تک نفره است. لطفا به تنهایی انجام شود، در غیر اینصورت نمره منفی در نظر گرفته خواهد شد .

[1] Reinforcement Learning

- کل محتوای ارسالی زیپ شود و نام فایل زیپ ارسالی HW4_StudentNumber_YourName باشد .
- محتوای ارسالی دارای راهنما (read me) جهت تسهیل اجرا باشد .
- زبان برنامه نویسی دلخواه است. (پیشنهاد : پایتون)
- در صورت استفاده از زبان پایتون فایل کد ترجیحا به فرمت **.ipynb** بوده و **فایل کد حتما به صورت اجرا شده آپلود گردد و از وجود خروجی سلول ها اطمینان حاصل نمایید .**
- موارد ارسال شده در تاریخی که بعدا مشخص خواهد شد و متعاقبا اعلام می‌گردد به صورت آنلاین نیز تحویل گرفته خواهند شد (صرفا آنچه در LMS طبق تاریخ فوق تحویل داده شده است بعدا به صورت حضوری تست شده و توضیح داده می شود).
- تنها تکالیفی که به LMS و قبل از مهلت ارسال، فرستاده می شوند بررسی خواهند شد .
- در صورت داشتن هر گونه سوال می‌توانید سوال خود را در گروه تلگرامی درس مطرح کنید .
- حداقل **یک ساعت قبل از مهلت ارسال** را احتیاط هدف قرار دهید، تا مشکلات غیرقابل پیش بینی مانند موارد زیر باعث عدم آپلود پاسخ ها در LMS و ارسال آنها از طریق ایمیل نشوند :

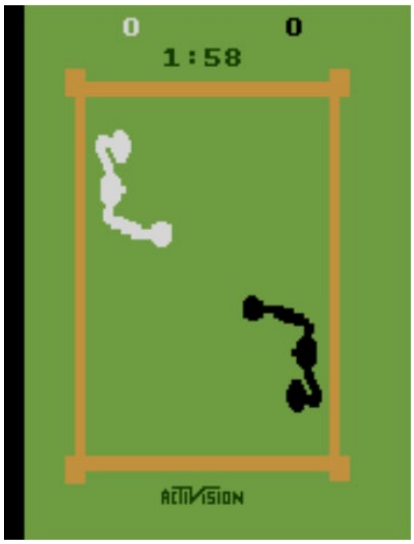
(قطعی اینترنت، تنظیم نبودن دقیق ساعت سایت با ساعت گرینویچ، کرش سیستم عامل و نیاز به فرمت، بارش زیبای شهاب سنگ از آسمان و ...)

شرح تمرین :

این تمرین شامل دو بخش عملی خواهد بود.

استفاده از کتابخانه های آماده مجاز نمی‌باشد، حتی شما دانشجوی عزیز!

بخش اول) یادگیری در محیط چندعامله



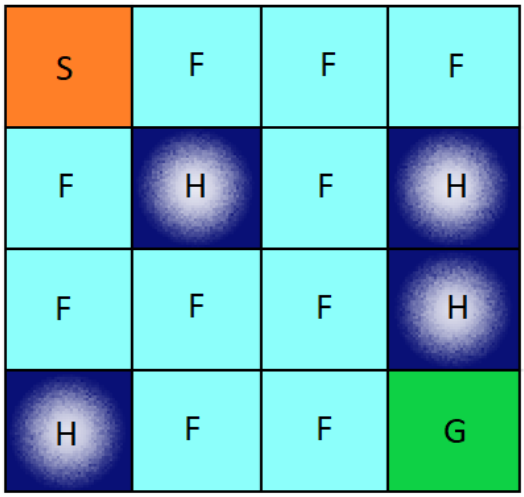
شکل (۱-۱) تصویری از محیط بوکس

بوکس یک بازی رقابتی است که در آن کنترل دقیق و واکنش‌های مناسب به حریف شما بسیار مهم است. بازیکنان دو دقیقه (حدود ۱۲۰۰ حرکت) برای مبارزه در رینگ زمان دارند. در هر حالت محیط، آن‌ها می‌توانند حرکت کنند و مشت بزنند. مشت‌های موفق امتیاز کسب می‌کنند، ۱ امتیاز برای یک ضربه جب (long range jab)، ۲ امتیاز برای یک ضربه قدرتی (close power punch) نزدیک و ۱۰۰ امتیاز برای knock out (که همچنین بازی را به پایان می‌رساند). زمانی که شما تعدادی امتیاز کسب می‌کنید، متناسب با تعداد امتیاز کسب شده پاداش دریافت خواهید کرد و حریف شما نیز با آن تعداد امتیاز مجازات می‌شود. بازی در یک محیط دو بعدی با اندازه کوچک اجرا می‌شود که در آن بوکسورها در حال حرکت بوده و سعی بر وارد کردن ضربات به حریف خود دارند و در عین حال می‌بایست از ضربات حریف جلوگیری کنند تا بتوانند امتیازات بیشتری را کسب کنند و در نهایت برنده بازی شوند.

فرض می‌کنیم که عامل‌ها محیط را نمی‌دانند. با استفاده از الگوریتم SARSA که می‌بایست توسط خودتان پیاده شود، یکی از عامل‌ها را به گونه‌ای آموزش دهید که توانایی تعامل با محیط و برد مسابقه را داشته باشد. در نهایت نمودار متوسط مجموع پاداش دریافتی (average reward) را رسم کنید. برای اطلاعات بیشتر از محیط بازی می‌توانید به این لینک مراجعه نمایید.

بخش دوم) عامل‌های model-based و model-free

در فصل زمستان در یک دریاچه یخی مسیر های مختلفی برای پیمایش بین نقطه شروع و پایان وجود دارد. عامل هوشمندی قرار است طی روز های مختلف این مسیر را بپیماید. نکته مهم برای این پیمایش این است که در هر منطقه از این دریاچه یک احتمال برای شکستن یخ وجود دارد (H). نقطه‌ی شروع و پایان مسیر در شکل ۲-۱ نشان داده شده است. خانه آبی (F) نشان دهنده ی مکان های بی خطر برای حرکت توسط agent است.



شکل (۲-۱) محیط دریاچه یخی

عامل در هر زمان می‌تواند از بین چهار حرکت چپ (۰)، پایین (۱)، راست (۲) و بالا (۳) یکی را انتخاب کند. البته توجه کنید که در حالت های مرزی در صورت انتخاب حرکت غیرمجاز عامل در سر جای خود باقی می

ماند . عامل به ازای هر حرکت به دلیل زمان از دست رفته پاداش ۱- دریافت می‌کند. در صورت سقوط در هر یک از خانه ها، بازی خاتمه یافته و عامل پاداش ۱۰- دریافت می‌کند و در صورت رسیدن به هدف عامل پاداش ۵۰+ دریافت می‌کند. هم چنین نکته قابل توجه دیگر آن است که حرکات عامل به دلیل سر بودن سطح دریاچه به صورت قطعی نیست به این معنا که عامل ما با احتمال ۰.۹ حرکت انتخاب شده را انجام می‌دهد و در غیر این صورت یکی از ٤ حرکت را به صورت تصادفی انجام می‌دهد. محیط این مسئله در پوشه تمرین پیوست گردیده است.

در این تمرین میخواهیم روشهای model-based و model-free و هم چنین ترکیب آنها را بررسی کنیم:

<u>برای تمامی بخش‌های زیر، هنگام تعریف محیط، متغییر studentNum را در کد های ارائه شده برابر شماره دانشجویی خودتان قرار دهید.</u>

۱. بررسی روشهای model-based :

ابتدا فرض می‌کنیم که عامل مدل محیط را می‌داند. با کمک الگوریتم value iteration و با در نظر گرفتن مقدار discount factor =۰.۹ ارزش هر خانه را پیدا کنید و به کمک آن Q-value های هر حالت-عمل (state-action) را محاسبه کنید . هم چنین نهایتا سیاست بهینه را بیابید . سپس بر روی نقشه برای هر خانه عدد عمل بهینه را نمایش دهید.

۲. بررسی روشهای model-free :

در این بخش میخواهیم عامل model-free را بررسی کنیم سیاست مورد استفاده برای عامل را epsilon-greedy در نظر بگیرید. مقدار اپسیلون را به صورت کاهشی و مقدار discount factor را ۰.۹ و همچنین مقدار نرخ یادگیری را برابر ۰.۱ در نظر بگیرید. برای تمامی روش‌های زیر مسئله را

برای تعداد اپیزود خواسته شده برای ۲۰ بار تکرار انجام دهید و متوسط مجموع پاداش دریافتی در هر اپیزود را در طول یادگیری را رسم نمایید.

الگوریتم q-learning و SARSA را به ازای نرخ یادگیری ۰.۱ و کاهشی پیاده‌سازی نمایید و نتایج بدست آمده را با استفاده از نمودار میانگین پاداش در افق ۲۰۰۰ اپیزود نمایش دهید.

آنچه تحویل داده می‌شود:

۱) کداجرایی برنامه با توضیحات لازم برای اجرا

۲) **گزارش کاملی از مسیر انجام کار، چالش‌هایی که مواجه شده‌اید، اجراهایی که گرفتید و نتایجی که حاصل شده است. گزارش کار از اهمیت بالایی برخوردار است، حجم آن و فرمت استاندارد آن اهمیت ندارد، اما باید نشان دهنده مسیر انجام پروژه، چالشها، راه حلها و نتایج کار شما باشد.**