تمرین سری ششم

درس: یادگیری ماشین- پاییز ۱۴۰۳

استاد درس: دکتر فاطمه زمانی

دستیاران آموزشی: حسین آقاگل زاده، ابوالفضل حسینی فر

دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی نوشیروانی بابل

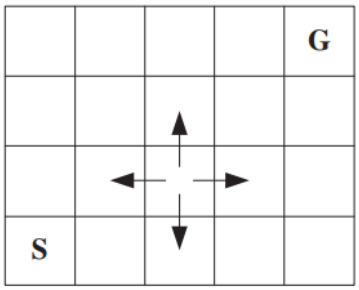
- این تمرین دارای سه بخش **تئوری**، **پیاده سازی** و **مطالعاتی** است.
- پاسخنامه بخش تئوری می تواند به صورت دست نویس یا تایپی باشد. در هر صورت اما می بایست در قالب **یک فایل pdf** تحویل گردد.
- پاسخ بخش پیاده سازی می بایست حاوی **کد پیاده سازی** شده **به همراه گزارش آن در قالب فایل pdf** باشد. از آوردن کد در گزارش اجتناب کنید (مگر آوردن بخشی از آن از نظر شما ضرورت داشته باشد).
- پاسخ بخش مطالعاتی می بایست به صورت تایپ شده در قالب یک فایل pdf باشد.
- گزارش بخش پیاده سازی می بایست میزان تلاش شما رو با ٤ اولویت مهم **۱- شفافیت ۲- درستی ۳-زیبایی ٤- کوتاهی**، نشان بدهبخش تئوری بدون استفاده از کد انجام بشه.
- کل پاسخ (کد و فایل های pdf) در **یک فایل zip** با فرمت زیر ارسال شود.

HW6_شماره دانشجویی.zip

تئوری

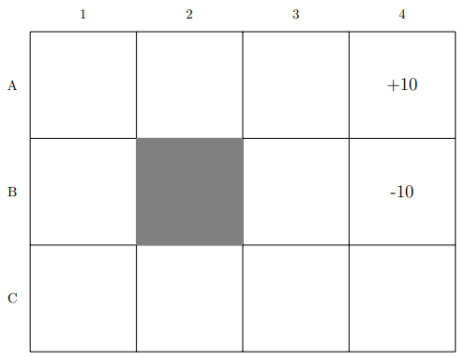
مسئله ۱

با توجه به شکل زیر و حالت شروع و هدف آن (پاداش هدف برابر با ۱۰۰ است)، $V^*(s)$، $Q^*(s,a)$ و اکشن های سیاست بهینه را بیابید. ($\gamma = 0.9$)



مسئله ۲

مسئله زیر را در نظر بگیرید. مقادیر اولیه ارزش حالت عمل را صفر در نظر بگیرید ($\gamma = 0.8$).



الف) فرض کنید عامل طی سه episode توالی زیر را جهت حرکت انتخاب می کند. اگر بروزرسانی ارزش حالت عمل بر اساس الگوریتم Qlearning انجام گیرد، در پایان اخرین توالی مقادیر ارزش عمل حالت ها چه خواهد بود؟

- B1 → A1 → A2 → A3 → B3 → B4
- A2 → A3 → A4
- C1 → C2 → C3 → B3 → A3 → A4

ب) فرض کنید عامل اکنون از سیاستی استفاده می‌کند که همیشه عملی را انجام می‌دهد که بیشترین مقدار Q را داشته باشد. آیا این سیاست بهینه است؟ چرا؟ (عامل ممکن است از هر حالت تصادفی ای شروع به حرکت کند)

مسئله ۳

یک شرکت ارایه دهنده خدمات برای هر خدمت خود n+1 قیمت دارد که nقیمت سودآور و یک قیمت بدون سود است .مردم به این خدمت نیاز دارند و تمایل هر مشتری به خرید خدمت وابسته به قیمت پیشنهادی و شهرت شرکت است. این بدان معنی است که اگر قیمت مناسب بود خدمت را خریده و در غیر این صورت انصراف میدهد. همچنین، برخی از مشتریان نارضایتی (چه خدمت را خریده و یا نخریده باشند) و یا رضایت خود از قیمت را از طریق کامنت در شبکههای اجتماعی نشر میدهند و این امر باعث تغییر شهرت شرکت میشود. تعداد مشتریان بالقوه زیاد است و قیمت مد نظر هر یک و نظر آنان نسبت به شرکت مشخص نیست. هدف این شرکت حداکثر کردن سود خود صرفا از طریق تنظیم قیمت است. یافتن قیمت مناسب را با یک مسئله RL مدل کرده و شبه کد آن را نیز ارائه دهید.

مسئله ۴

در مسئله ۱ پیاده سازی فرض کنید علاوه بر Mr. Nobody در آن لحظه، ۳ بازیکن دیگر نیز در محیط حضور دارند. در هر دور بازی Mr. Nobody آخرین فردی است که بازی میکند. هر کدام از این ٤ نفر به ترتیب یک بار اقدام به انتخاب یک دکمه برای بازی میکنند. Mr. Nobody نحوه انتخاب دکمه بازیکنان دیگران را نگاه میکند ولی متاسفانه نمی تواند پاداش دریافتی و عکس العمل دیگر بازیکنان را ببیند. حال به سوالات زیر پاسخ دهید (نیاز به استفاده از اعداد مسئله ۱ پیاده سازی نیست).

الف) با توجه به اینکه Mr. Nobody اطلاعاتی نسبت به سیاست دیگران ندارد، شبه کدی ارائه دهید که او چگونه میتواند از رفتار مشاهده شده از آنها برای دریافت پاداش بیشتر و برنده شدن استفاده کند؟

ب) به نظر شما بهتر است Mr. Nobody از همان ابتدا تقلب کند و یا اجازه دهد مدت زمانی بگذرد؟ پاسخ خود را تحلیل کنید

مسئله ۵

یک مسئله multi agent multi armed bandit را در نظر بگیرید. شما میتوانید رفتار سایر عامل ها را ببینید ولی درکی از پاداش دریافتی آنها ندارید. اما در این مسئله آنها به شما واریانس پاداش هر کدام از بازوها را میگویند. با این اطلاعات چگونه مسئله را حل میکنید؟

پیاده سازی

مسئله ۱

در یک مرکز تفریحی یک ماشین شانس 4 دکمه ای، قرار داده اند.

دکمه اول از توزیع N(a,1) دکمه دوم از توزیع N(b, 2،) دکمه سوم از توزیع N(c, 1) و دکمه چهارم با احتمال 0.7 از توزیع N(d,2) و با احتمال 0.3 از توزیع U(-d, 1) پاداش به بازیکنان میدهد (دقت کنید که a رقم یکان شماره دانشجویی شما باشد، c=b-1، b=a-2 و d رقم دهگان شماره دانشجویی شما است)

الف) الگوریتم epsilon-greedy با اپسیلون ثابت ۰٫۲ را پیاده سازی کنید و نمودار پاداش دریافتی به ازای هر تکرار را رسم کنید.

ب) الگوریتم epsilon-greedy را این بار با اپسیلون متغیر با استراتژی اپسیلون متغیر اجرا و نمودار پاداش را رسم کنید.

ب) فرض کنید که Mr. Nobody در بخش الف و ب پاداش خود را بر حسب utility function با آلفا برابر ۱ بتا برابر ۱ دریافت کرد. حال برای دو مقدار مختلف از هر پارامتر utility بررسی کنید که تاثیر دریافت پاداش با utility های مختلف بر نتیجه ی بخش ب چگونه خواهد بود؟

$$utility(r) = \begin{cases} \alpha r & r \geq 0 \\ \beta r & r < 0 \end{cases}$$

مسئله ۲

از میان مسائل موجود در لینک زیر Cart Pole را با روش q learning (بدون شبکه های عصبی) حل و نمودار میانگین پاداش دریافتی به ازای هر اجرا (episode) رسم کنید (هر episode شامل چندین تکرار است).

https://www.gymlibrary.dev/environments/classic_control/

ساخت محیط با استفاده از راهکار های موجود در لینک بالا باشد. کاملا مجاز هستید که از پیاده سازی های موجود کمک بگیرید. تا حد امکان پیاده سازی شما ساده و مطابق با مفاهیم و علامت گذاری هایی که در کلاس فرا گرفته اید باشد. گزارش این مسئله کاملا گویا و شامل تمام مراحل پیاده سازی باشد. گزارش را با توصیف محیط مانند اکشن های موجود، هدف نهایی و ... شروع کنید. نیاز به پرداختن جزئیات کد در گزارش نیست.

در نهایت پس از آموزش کامل مدل خود، آن را روی مسئله اجرا و یک ویدیو کوتاه تهیه کنید که مدل توانسته مسئله را حل کند(نمونه ویدیو در گروه ارسال می شود).

مطالعاتی

مقاله ی پیوست شده را مطالعه کنید و برداشت خود از آن را شرح دهید (حداقل ۲ صفحه).

موفق باشید 😊