هدف این تمرین آشنایی با الگوریتم‌هایی برای حل مسئله‌ی MDP با فرض ناشناخته بودن محیط می‌باشد. از این روش‌ها در ادبیات به عنوان روش‌های بدون مدل (Model-Free) یاد می‌شود. در این تمرین دو سوال تحلیلی و یک مسئله‌ی پیاده‌سازی که شامل بخش‌های مختلف می‌شود در نظر گرفته شد که طی آن با الگوریتم‌های SARSA, Expected SARSA, Q-learning, Tree Backup n-step آشنا می‌شوید.

سوالات تحلیلی

۱. الگوریتم‌های Sarsa و Expected-Sarsa در یک محیط گسسته از منظر میزان پشیمانی،

۱,۱. در دریافتن سیاست e-optimal چه تفاوتی دارند؟

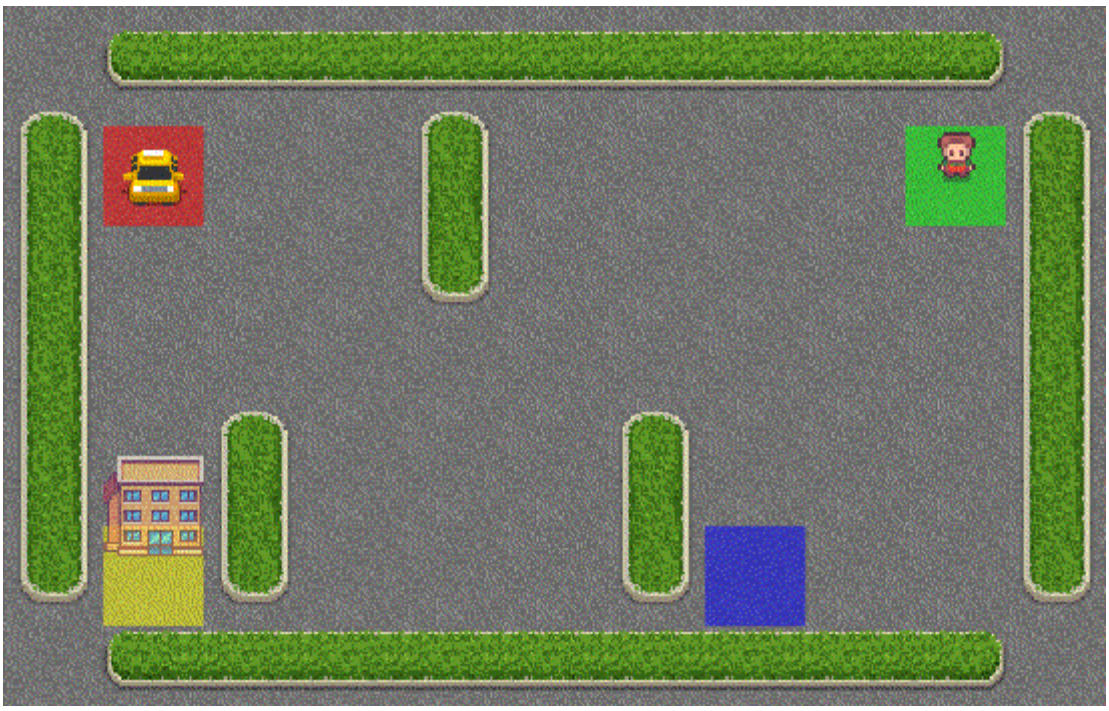
۱,۲. در دریافتن سیاست بهینه چه تفاوتی دارند؟

۱,۳. فرض کنید برای یک MDP مدل‌های کاهش بهینه نرخ یادگیری و کاهش بهینه epsilon برای یادگیری Sarsa و Expected-Sarsa با ماشین تصمیم‌گیری epsilon-greedy به دست آمده است. سرعت کاهش نرخ یادگیری کدامیک سریع‌تر است؟ همین سوال را در خصوص epsilon نیز پاسخ دهید.

۲. یک عامل یادگیر در یک مسئله‌ی MDP از n-step-return برای یادگیری استفاده می‌کند. این عامل یادگیر همواره از حالت s_0 شروع می‌کند. حداکثر خطای تخمین یک سیاست مشخص $V^{\pi}(s)$ برای این عامل به ازای یک مقدار مشخص n چه قدر است؟

توضیح مسئله

علی که برخلاف شما درس یادگیری تعاملی را نداشته است، قصد دارد تاکسی اینترنتی‌ای راه‌اندازی کند که نیازی به راننده نداشته باشد و عاملی هوشمند به جای راننده مسافرین را به مقصدشان برساند. بدین منظور ابتدا می‌خواهد که امکان این مسئله را در یک محیط فرضی بسنجد. محیط شهر یک جدول ۵ در ۵ است که دور تا دور و بخش‌هایی از درون آن دیوارهایی وجود دارد. برای آشنایی دقیق‌تر با این محیط می‌توانید از این لینک استفاده کنید.



شکل ۱- تصویر بالا نشان دهنده‌ی نقشه این شهر است که دور تا دور آن را دیوار کشیده شده است.

عامل در هر زمان می‌تواند از بین چهار حرکت پایین(۰)، بالا(۱)، راست(۲)، چپ(۳)، سوار کردن مسافر(۴) و پیاده کردن مسافر(۵) یکی را انتخاب کند. البته توجه کنید که در حالت‌های مرزی در صورت انتخاب حرکت غیرمجاز عامل در سر جای خود باقی می‌ماند. از طرفی، بسته به شماره‌ی دانشجویی شما، مبدا مسافر یکی از خانه‌های رنگی و مقصد وی خانه‌ی دیگری می‌باشد. عامل به ازای رساندن مسافر پاداش ۲۰+، به ازای سوار کردن یا پیاده کردن غیر مجاز پاداش ۱۰- و در غیر این صورت به دلیل زمان از دست رفته پاداش ۱- دریافت می‌کند.

نحوه‌ی استفاده از محیط

برای این تمرین قصد داریم با Gym که یک رابط کاربردی برای یادگیری تعاملی و مجموعه‌ای از محیط‌های مختلف است، آشنا شویم. در این لینک توضیح ساده‌ای از نحوه‌ی استفاده از محیط‌های آن داده شده است. هم‌چنین کد این محیط، در فایل ضمیمه آورده شده است، توجه داشته باشید که شما

حتماً باید در زمان reset محیط، seed محیط را برابر با سه رقم آخر شماره دانشجویی خود کنید. برای مثال اگر شماره‌ی دانشجویی شما ۸۱۰۱۹۷۱۲۳ باشد، قطعه کد به صورت زیر خواهد بود.

```
import gym
env = gym.make('Taxi-v3')
env.seed(seed=123)
Initial_state = env.reset()
```

هر حالت در این محیط با یک عدد نمایش داده می‌شود. می‌توان از دستور زیر برای پیدا کردن حالت مسئله استفاده کرد. مثلا اگر در حالت ۱۸۹ باشیم، اطلاعات حالت مذکور را به شکل زیر می‌توان پیدا کرد.

```
taxi_row, taxi_col, pass_idx, dest_idx = env.decode(189)
```

نکات پیاده‌سازی

برای پیاده‌سازی سوالات آینده به این نکات توجه بفرمایید:

- 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 ۰/۹ در نظر بگیرید. هم‌چنین مقدار نرخ یادگیری را برابر ۰/۱ در نظر بگیرید.
- برای تمامی روش‌های زیر مسئله را حداقل ۲۰ بار تکرار به اندازه‌ی حداقل ۲۰۰۰ اپیزود انجام دهید و متوسط پاداش دریافتی در طول یادگیری را رسم نمایید (با استفاده از پنجره-ی متحرک مناسب) و همگرایی به سیاست بهینه را بررسی نمایید. (در صورت امکان با انجام محاسبات)
- هم‌چنین در پایان هر یک از سوال‌های این بخش یکی از عامل‌ها(بهترین یا میانگین عامل‌ها) را پس از آموزش به اندازه‌ی ۲۰ اپیزود تست کنید و هم‌چنین با استفاده از دستور render رفتار آن را نمایش دهید

تذکر ۱: 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 قید نشده باشد شما می‌توانید این پارامترها را تغییر دهید.

سوالات پیاده‌سازی

۱. الگوریتم q-learning را یکبار به ازای نرخ یادگیری ۰.۱ و بار دیگر به ازای نرخ یادگیری کاهشی پیاده سازی نمایید و نتایج بدست آمده را از حیث میزان حسرت (سرعت همگرایی و مقدار همگراشده) با یکدیگر مقایسه کنید. روش انتخابی خود برای کاهش مقدار اپسیلون در طی فرآیند یادگیری را توضیح دهید.

۲. تعدادی از حالت‌ها در محیط معرفی شده قابل دستیابی نیستند. پس از توصیف ویژگی این حالت‌ها راهکاری الگوریتمی برای بدست آوردن شماره‌ی این حالت ها ارائه کنید و آن راهکار را با استفاده از حل سوال پیشین امتحان کنید.

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

۴. با توجه به شماره‌ی دانشجویی خود به سوال زیر پاسخ دهید.

اگر رقم آخر شماره دانشجویی شما زوج است:

۴,۱.با استفاده از روش on-Policy MC مسئله را حل کنید و موارد خواسته شده را یکبار برای اپسیلون کاهشی و هم چنین برای اپسیلون ۰/۱ انجام دهید و نتایج بدست آمده را از حیث میزان حسرت (سرعت همگرایی و مقدار همگراشده) با یکدیگر مقایسه کنید.

اگر رقم آخر شماره دانشجویی شما فرد است:

۴,۲. با استفاده از روش off-Policy MC خواسته‌های مسئله را پاسخ دهید و موارد خواسته شده را یکبار برای اپسیلون کاهشی و هم چنین برای اپسیلون ۰/۱ انجام دهید و نتایج بدست آمده را از حیث میزان حسرت (سرعت همگرایی و مقدار همگراشده) با یکدیگر مقایسه کنید. (توجه: سیاست رفتاری را یک سیاست epsilon-greedy در نظر گرفته و در هر مرحله آن را بر اساس آخرین مقدار Q-value ها بروز کنید.)

۵. سرعت یادگیری در سوال آخر نسبت به سوال‌های پیشین تفاوت محسوسی می‌کند؟ در صورت جواب مثبت علت این مسئله را توضیح دهید.
(امتیازی): راهکاری برای افزایش سرعت ارائه کرده و نتیجه‌ی آن را مقایسه و تحلیل کنید.

نکات پیاده‌سازی و تحویل

- مهلت ارسال این تمرین تا پایان روز **یکشنبه ۱۱ دی ماه** خواهد بود.
- در رسم نمودارها حتماً باید title، axis label و grid داشته باشد و مقادیر به‌صورت گویا نمایش داده شود.
- پیاده‌سازی تنها با پایتون قابل‌قبول است.
- حجم گزارش شما هیچ‌گونه تأثیری در نمره نخواهد داشت و تحلیل و نمودارهای شما بیشترین ارزش را دارد.
- گزارش خود را در قالب آپلود شده در سامانه نوشته و ارسال کنید.
- انجام این تمرین به‌صورت یک‌نفره می‌باشد.
- سعی کنید از پاسخ‌های روشن در گزارش خود استفاده کنید و اگر پیش‌فرضی در حل سوال در ذهن خود دارید، حتما در گزارش خود آن را ذکر نمایید.
- لطفاً گزارش، فایل کدها و سایر ضمائم موردنیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمایید.

HW1_[Lastname]_[StudentNumber].zip

- در صورت وجود سؤال و یا ابهام می‌توانید تنها از طریق رایانامه زیر با دستیار آموزشی در ارتباط باشید:

- علی نقدی alinaghdi8@gmail.com (سوالات تحلیلی)

- عرفان میرزایی erfunmirzaei@gmail.com (سوالات پیاده‌سازی)

- علیرضا توکلی alirezata3akoli@gmail.com (سوالات پیاده‌سازی)

امیدواریم که سلامت باشید :)