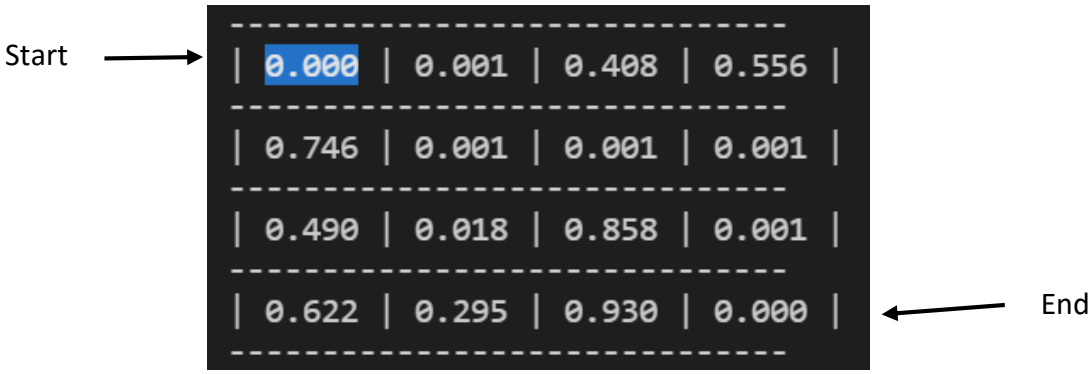
در فصل زمستان در یک دریاچه یخی مسیر های مختلفی برای پیمایش بین نقطه شروع و پایان وجود دارد. عامل هوشمندی قرار است طی روز های مختلف این مسیر را بپیماید. نکته مهم برای این پیمایش این است که در هر منطقه از این دریاچه یک احتمال برای شکستن یخ وجود دارد. نقطه‌ی شروع و پایان مسیر در شکل ۱ نشان داده شده است.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



شکل ۱ - تصویر بالا نشان دهنده ی نقشه این دریاچه است که دور تا دور آن را دیوار است.

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 ۴ حرکت را به صورت تصادفی انجام می‌دهد. کد این محیط، در فایل ضمیمه آورده شده است، توجه داشته باشید که کلاس محیط به عنوان تنها ورودی اجباری شماره دانشجویی هر فرد را دریافت می‌کند. هم چنین یک فایل راهنمای کلی نیز در ضمیمه این فایل وجود دارد.

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

۱. **بررسی روش‌های model-based** :

ابتدا فرض می‌کنیم که عامل مدل محیط را می‌داند. با استفاده از دستور زیر چهار خروجی لازم برای اجرای الگوریتم value iteration را دریافت می‌کنیم. (چهار خروجی شامل حالت های ممکن لحظه بعد، احتمال گذار به آن حالت‌ها و احتمال سقوط و خاتمه یافتن در هر حالت را گزارش می‌کند) :

```
states, probs, fail_probs, dones = self.possible_consequences(action, state_now)
```

حال با کمک الگوریتم value iteration و با در نظر گرفتن مقدار discount factor =0.9 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 روش های 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. (با استفاده از پنجره‌ی متحرک مناسب) و همگرایی به سیاست بهینه را بررسی نمایید.(در صورت امکان با انجام محاسبات لازم)

اگر رقم آخر شماره دانشجویی شما زوج است:

۲,۱,۱. 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 و ۰,۰۵ انجام دهید و نتایج بدست آمده را از حیث میزان حسرت در افق ۴۰۰۰ اپیزود (سرعت همگرایی و مقدار همگراشده) با یکدیگر مقایسه کنید.

اگر رقم آخر شماره دانشجویی شما فرد است:

۲,۱,۲. با استفاده از روش off-Policy MC خواسته‌های مسئله را پاسخ دهید و موارد خواسته شده را یکبار برای اپسیلون کاهشی و بار دیگر برای اپسیلون ۰,۱ انجام دهید و نتایج بدست آمده را از حیث میزان حسرت در افق ۴۰۰۰ اپیزود (سرعت همگرایی و مقدار همگراشده) با یکدیگر مقایسه کنید.(توجه: سیاست رفتاری را یک سیاست epsilon-gready در نظر گرفته و در هر مرحله آن را بر اساس آخرین مقدار Q-value ها بروز کنید.)

۲,۲. الگوریتم q-learning را به ازای نرخ یادگیری ۰,۱ و کاهشی پیاده سازی نمایید و نتایج بدست آمده را از حیث میزان حسرت در افق ۲۰۰۰ اپیزود (سرعت همگرایی و مقدار همگراشده) با یکدیگر مقایسه کنید.

۲,۳. الگوریتم Sarsa و 2-Step Tree Backup را پیاده سازی کنید و نتایج بدست آمده را حیث میزان حسرت در افق ۲۰۰۰ اپیزود (سرعت همگرایی و مقدار همگراشده) با یکدیگر مقایسه کنید.

۳. ترکیب روش‌های Model-Based و Model-free :

در این بخش فرض ثابت بودن احتمال شکستن خانه را کنار گذاشته ایم. بنابراین احتمال شکستن هر خانه(سقوط) با توجه به تغییرات آب و هوا تغییر می‌کند، بنابراین محیط حالت ناپایا (Non-Stationary) دارد. (هنگام تعریف محیط، متغییر nonStationary را True کنید.)

در این بخش می‌خواهیم با ترکیب q-learning و مقادیر بدست آمده از value iteration این مسئله‌ی را حل نماییم. سیاست مورد استفاده برای عامل را epsilon-greedy در نظر بگیرید. مقدار اپسیلون را برابر ۰,۱ در نظر بگیرید. در این الگوریتم در هر گام ، عامل با توجه به رابطه‌ی زیر تصمیم‌گیری می‌کند:

$$Q[s,a] = w * Q\_Model\_Based[s,a] + (1-w) * Q\_Model\_Free[s,a]$$

با اجرای الگوریتم فوق به ازای $w = 0, 0.1, 0.2, \ldots, 1$ ، میانگین مجموع پاداش های دریافتی در طول اپیزودها را برای تکرارهای متفاوت محاسبه کنید و نتیجه را در یک نمودار نمایش دهید. در این مسئله نیز مانند بخش قبل مسئله را برای ۱۰،۰۰۰ اپیزود برای ۲۰ بار تکرار انجام دهید.

امتیازی: ترکیب اطلاعات ناقص محیط با Model-Free (Model-Based Learning)

در این بخش فرض کنید که عامل ما می‌خواهد به طور کلی یادگیری بر روی دریاچه‌های یخ زده را یاد بگیرید نه صرفا در این دریاچه خاص. بنابراین دیگر احتمالات شکستن هر منطقه از دریاچه را در اختیار ندارد و تنها احتمال انتقال از یک حالت به حالت دیگر (State Transition) را می‌داند. یعنی می‌داند که با این کفش‌ها بر روی یخ چگونه جابجا می‌شود. حال می‌خواهیم با فرض دانستن این اطلاعات یادگیری را انجام دهیم. شبه کد روش پیشنهادی خود را بیان کرده و آن را پیاده‌سازی نمایید. (در صورت استفاده از مقالات یا منابع جانبی به آن منبع ارجاع دهید.)

نکات تکمیلی:

- سعی کنید از پاسخ های روشن در گزارش خود استفاده کنید و اگر پیش فرضی در حل سوال در ذهن خود دارید، حتما در گزارش خود آن را ذکر نمایید.
- حجم گزارش شما به هیچ وجه معیار نمره دهی نیست، پس لطفا در حد نیاز توضیح دهید.
- برای پیاده‌سازی الگوریتم های بخش های ۲ و ۳ باید از کلاس agent کتابخانه‌ی amalearn استفاده کنید در غیر این صورت مقدار قابل توجهی از نمره را دریافت نخواهید کرد.
- از نمودارهای واضح در گزارش خود استفاده کنید، نمودارهایتان حتما دارای لیبل واضح روی هر محور و توضیح مناسب باشد.
- لطفا در گزارش و کدهای خود از تمرین دیگران استفاده نکنید .مشورت و همفکری در مورد سوال‌ها اشکالی ندارد اما اگر شباهت بیش از اندازه در تمرین ها دیده شود منجر به صفر شدن نمره خواهد شد.
- تمام فایل ها را در قالب یک فایل zip در سایت درس بارگذاری کنید.
- حتما فرمت گزارش که در سایت درس قرار داده شده است را رعایت نمایید.
- در صورت وجود هر نوع سوال در رابطه با این سری تمرین می‌تونید از طریق بخش پرسش و پاسخ سایت ایلرن و هم چنین ایمیل‌های زیر با دستیاران آموزشی درارتباط باشید. از آنجایی که معمولا سوالات بوجود آمده برای شما برای سایر دوستان‌تان نیز وجود دارد توصیه می‌شود تا حد امکان سوالات خود را در فروم مطرح کنید.

- عرفان میرزایی erfunmirzaei@gmail.com (سوال ۲)
- رضا کرباسی arzkarbasi@gmail.com (سوال ۱ و ۳)

شاد و سلامت باشید):