هدف از این تمرین آشنایی با مسائل MDP، مدلسازی و حل آنهاست. در این تمرین دو الگوریتم مهم value iteration و policy iteration را با شرایط مختلفی پیاده سازی و تحلیل می‌کنیم.

مسئله ۱- آشنایی با MDP

برای مسائل زیر یک MDP ارائه دهید. لازم به ذکر است برای توصیف مدل لازم است که استیت‌ها، اکشن‌ها، پاداش و انتقال بین استیت‌ها را تعیین کنید.

سوال ۱

یک اپ فیتنس موجود است که متناسب با شرایط هر کسی باید برنامه ورزشی و رژیم مناسبی را ارائه دهد. برای این هدف از reinforcement learning based recommender systems در این اپ استفاده شده است. مدل MDP ارائه دهید و بررسی کنید این برنامه نسبت به یک توصیه‌گر معمولی چه مزایا و معایبی دارد.

سوال ۲ (امتیازی)

در خانه‌ی کدخدا مراسمی برگزار خواهد شد و همه‌ی اهالی دهکده می‌توانند در این مراسم شرکت کنند و با احتمالی برنده یک جایزه‌ی نقدی شوند. اما مدتی است که یک بیماری خطرناک در دهکده شیوع پیدا کرده و اهالی از اینکه حامل این بیماری هستند یا نه، اطلاعی ندارند (علائم بیماری پنهان است)؛ فرد مبتلا ممکن است که پس از مدتی فلج شود و یا حتی بمیرد. احتمال سرایت از یک فرد بیمار به یک فرد سالم با مقدار زیر متناسب است.

$$(\text{تعداد افراد بیمار})^{a} \times (\text{تعداد افراد سالم}), \qquad a > 1$$

توجه: این مسئله multi-agent است که در آن رفتار عامل‌های مختلف بر یکدیگر اثر می‌گذارد.

مسئله ۲- پیاده سازی دستی

نقشه‌ی ۳ در۴ زیر را در نظر بگیرید.

		Hell	
Obstacle			
			Goal

همان‌طور که مشاهده می‌شود یک خانه‌ی هدف داریم که ورود به آن برای ما ۱۰ امتیاز مثبت دارد. هم‌چنین ورود به خانه‌ی قرمز، ۱۰ امتیاز منفی خواهد داشت. یک مانع هم داریم که باعث می‌شود نتوانیم وارد آن خانه مشخص بشویم. در این زمین بازی احتمال انتقالات بدین صورت است که اگر در خانه ای باشیم و تصمیم بگیریم به سمت پایین حرکت کنیم و مانعی سر راهمان قرار نداشته باشد، دقیقا با احتمال ۱ وارد خانه پایین خواهیم شد (زمین بازی لغزنده نیست). از طرفی وقتی در خانه های کناری قرار داریم، با حرکت به سمت دیوار، با احتمال ۱ در خانه‌ی خودمان می‌مانیم.

با توجه به توضیحات بالا به بخش های زیر پاسخ دهید.

سوال ۱

ابتدا با در نظر گرفتن discount factor = 0 مسئله‌ی بالا را با شروع از یک مقدار اولیه و سیاست رندم، با استفاده از الگوریتم policy iteration تا ۵ تکرار به صورت دستی حل کنید. در هر مرحله تغییرات مقدار ارزش هر خانه را مشاهده کنید و با استفاده از ماهیت الگوریتم های iterative، تحلیل خود را از چگونگی روند تغییر ارزش خانه‌ها و همچنین سیاست بهینه بیان کنید.

سوال ۲

در این قسمت مانند سری قبل عمل کرده با این تفاوت که discount factor را برابر 0.9 قرار دهید. علاوه بر تحلیل های خواسته شده‌ی پیشین، تاثیر تغییر مقدارdiscount factor را هم در این دو قسمت روی روند رسیدن به سیاست و ارزش بهینه بررسی کنید.

سوال ۳

در این بند، مسئله را با discount factor = 0.9 این‌بار با الگوریتم value iteration انجام دهید. در این قسمت هم مقدار بدست آمده برای سیاست و همچنین ارزش هر استیت را در هر گام بررسی کنید.

آیا مفهوم خاصی برای ارزش بدست آمده در هر مرحله وجود دارد؟ آیا لزوما در هر مرحله‌ی این الگوریتم، سیاستی متناظر با ارزش های بدست آمده وجود دارد؟ دلیل خود را بیان کنید.

سوال ۴

فرض کنید به جای استفاده از مقدار اولیه‌ی رندم برای ارزش استیت ها، از یک ارزشی استفاده کنیم که سیاست متناظر با آن سیاست بهتری نسبت به سیاست متناظر با ارزش های با مقدار رندم باشد. آیا این کار ممکن است در iteration های کمتری به جواب بهینه رسید؟ آیا این عمل تاثیری روی **نرخ همگرایی** هم دارد؟ دلایل خود را کامل توضیح دهید.

مسئله ۳- شبیه‌سازی

در این بخش قرار است مسئله روی دریاچه‌ی یخی را حل کنیم. در ابتدا دریاچه به این صورت است که تنها یک مسیر امن وجود دارد که با وارد نمودن شماره‌ی دانشجویی برای هر فرد متفاوت است. احتمال شکستن برای هر خانه این مسیر امن یک مقدار بسیار کوچک 0.0001 است و احتمال شکستن باقی خانه‌ها 1.0 است. یعنی ورود به باقی خانه ها برابر با شکستن و از بین رفتن است. این دریاچه‌یخی 6در6 است و هر چهار طرف آن فنس کشیده شده که خارج شدن از محیط دریاچه را غیرممکن می‌کند. خانه‌ی آخر یعنی (5,5) خانه‌ی هدف است و از نقطه‌ی (0,0) شروع می‌کنیم (احتمال شکستن هر دوی این خانه‌ها صفر است).

Start →

0.0	1.0	1.0	1.0	1.0	1.0
0.0001	1.0	1.0	1.0	1.0	1.0
0.0001	0.0001	1.0	1.0	1.0	1.0
1.0	0.0001	0.0001	1.0	1.0	1.0
1.0	1.0	0.0001	0.0001	0.0001	1.0
1.0	1.0	1.0	1.0	0.0001	0.0

شکل ۱: نقشه دریاچه یخی (خانه‌های خاکستری مسیر امن برای رسیدن به خانه هدف را نشان می‌دهند).

در فایل env.py ضمیمه شده نقشه‌ی ابتدایی دریاچه قرار داده شده‌است. در کلاس FrozenLake موجود در این فایل، توابع مورد نیاز را کامل و هر تابع دیگری که نیاز است، تعریف کنید. هم‌چنین شما می‌بایست مدل را در هر قسمت، طبق موارد خواسته شده تغییر دهید. فایل sample.py نیز به‌عنوان یک نمونه کد ضمیمه شده‌است.

بخش اول

در ابتدا محیط را به صورت زیر در نظر بگیرید.

پاداش رسیدن به خانه هدف 100، برای هر حرکت 1- و برای افتادن در خانه‌های شکسته‌ی دریاچه 10- امتیاز را در نظر بگیرید.

از آن‌جایی که در این حالت لغزندگی بسیار کم است، احتمال انتقال‌ها به این صورت است که با احتمال 0.94 در جهت همان اکشنی که انجام داده، و به طور مساوی و به حالت رندم به یکی از جایگاه های ممکن دیگر می‌رود. در صورت برخورد با فنس، عامل سر جای خود می‌ماند.

در این قسمت، با استفاده از الگوریتم‌های value iteration و policy iteration، مقادیر ارزش استیت‌ها، ارزش استیت-اکشن‌ها و سیاست بهینه را بدست آوردید (مقادیر ارزش استیت ها و سیاست بهینه را روی نقشه‌ی دریاچه نمایش دهید). مقدار discount factor را برابر 0.9 قرار دهید.

بخش دوم

در این قسمت قرار است تغییراتی در نقشه‌ی دریاچه اعمال کنیم. احتمال شکستن خانه ها تغییر خواهد کرد. احتمال شکستن هر خانه در مسیر امن همچنان مقدار کمتری است (0.001) و باقی مسیر دارای یک مقدار رندم در بازه‌ی 0 و 1 است.

از طرفی به دلیل لغزندگی، با احتمال 0.7 در جهت اکشن انجام شده و به طور مساوی و به حالت رندم به یکی از جایگاه های ممکن دیگر می‌رود.

مقادیر پاداش را در این حالت مانند سری قبل در نظر بگیرید.

در این قسمت نیز با استفاده از الگوریتم‌های value iteration و policy iteration، مقادیر ارزش استیت‌ها، ارزش استیت-اکشن‌ها و سیاست بهینه را بدست آوردید (مقادیر ارزش استیت ها و سیاست بهینه را روی نقشه‌ی دریاچه نمایش دهید). مقدار discount factor را برابر 0.9 قرار دهید.

مقادیر ارزش‌ها و سیاست بهینه‌ی بدست آمده را با قسمت قبل مقایسه کرده و تحلیل کنید. هم‌چنین، سرعت همگرایی الگوریتم‌ها را با هم مقایسه کنید.

بخش سوم

در این قسمت می‌خواهیم تاثیر تابع پاداش و مقدار ترشولد θ در الگوریتم‌ها را بررسی کنیم. نقشه‌ی محیط را تغییر داده و ابعاد دریاچه را به‌صورت 15در15 در نظر بگیرید. شرایط قسمت دوم را برای این نقشه‌ی جدید اعمال کنید.

الف) آیا تغییر مقدار ترشولد در الگوریتم value iteration، در یافتن سیاست بهینه تاثیری دارد؟ مقدار ترشولد باید چگونه باشد تا یک سیاست بهینه حاصل شود؟ پاسخ خود را با تغییر θ (از مقادیر بزرگتر مثل 1 تا مقادیر نزدیک به صفر مثل 0.000001) و بدست آوردن نتایج توضیح دهید.

ب) آیا با سیاست بهینه بدست آمده در قسمت الف، عامل به خانه‌ی هدف می‌رسد؟ در غیر اینصورت، مقدار پاداش خانه‌ی هدف را به نحوی تغییر دهید که مسیری که از خانه‌ی اول شروع می‌شود، با اتخاذ سیاست بهینه به خانه‌ی هدف برسد. مشاهدات خود در این قسمت را تحلیل و توجیه کنید.

نکات پیاده‌سازی و تحویل

- مهلت ارسال این تمرین تا پایان روز **سه‌شنبه ۱۵ آذر** ماه خواهد بود.
- پیاده‌سازی تنها با پایتون قابل‌قبول است.
- حجم گزارش شما هیچ‌گونه تأثیری در نمره نخواهد داشت و تحلیل‌های شما بیشترین ارزش را دارد.
- گزارش خود را در قالب آپلود شده در سامانه نوشته و ارسال کنید.
- انجام این تمرین به‌صورت یک‌نفره می‌باشد.
- لطفاً گزارش، فایل کدها و سایر ضمائم موردنیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمایید.

HW3_[Lastname]_[StudentNumber].zip

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

kimia.alavi@ut.ac.ir

Elahe.bvakili97@gmail.com