تمرین سری دوم درس یادگیری تعاملی

پاییز ۱۴۰۰

بخش اول مدلسازی MDP

برای مسائل زیر مدلی برای حل ارائه دهید. لازم به ذکر است برای مدل نیاز دارید استیت ها، اکشن ها انتقال بین استیت ها به صورت احتمالی و پاداش را تعیین کنید. (توجه کنید که جواب یکسان برای مسائل وجود ندارد و همچنین به پاسخ های خلاقانه نمره امتیازی تعلق خواهد گرفت)

۱. در یک منطقه ی کشاورزی در ابتدای هر فصل می خواهیم تصمیم بگیریم با توجه به وضعیت آب و هوا و خاک چه میزان محصول کاشت کنیم. لازم به ذکر است که کاشت محصول هزینه دارد و سود از فروش محصولات است و وضعیت آب و هوا بر روی میزان محصول و کیفیت آن ها و وضعیت خاک اگر محصول بیشتر از حد کاشت شود نیاز به هزینه برای بهبود دارد.

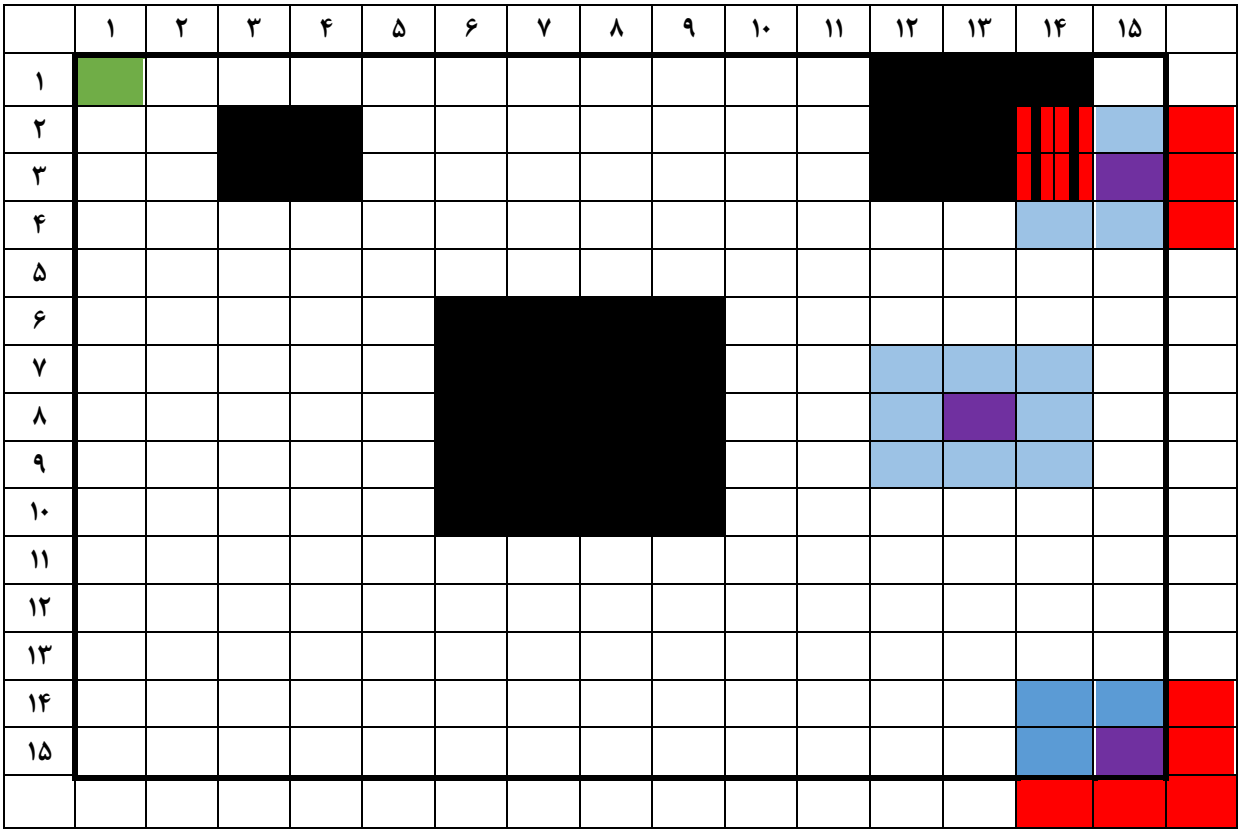
۲. در یک هتل هر روزه تعدادی رندوم مسافر خواهیم داشت و تعدادی رندوم قصد تخیلیه در فردای آن روز را میکنند. در هر روز مدیر هتل مشخص میکند که فردای آن روز چند مسافر را قبول کنند و می خواهد این عدد را بیشینه کند اما هر مشتری که با نبود اتاق خالی مواجه شود کامنتی منفی برای هتل قرار میدهد که از اعتبار هتل میکاهد.

۳. در یک کارخانه قصد داریم برای هر دستگاه با توجه به سن و وضعیت تصمیم بگیریم که نیاز به بررسی، تعمیر و یا تعویض دارند یا خیر. توجه داشته باشید که خراب شدن دستگاه ضرر به محصولات خواهد زد و بهتر است قبل از خرابی کامل تعمیر و یا تعویض انجام شود اما بررسی و تعمیر و یا تعویض هر دستگاه نیازمند هزینه است که بهتر است برای دستگاه های خوب انجام نگیرد.

بخش دوم پیاده سازی

مسئله مسیریابی برای رباتها از مسائل مهم و اساسی برای رباتهای امروزی میباشد. مواردی مانند عدم برخورد با موانع موجود و پیدا کردن بهترین مسیر در محیط داده شده از این قبیل مسائل میباشند. این مسائل را میتوان در ترکیب با مسائل MDP حل کرد. محیط زیر را طبق توضیحات داده شده در نظر گرفته، و پس از پیاده سازی موارد خواسته شده، به سوالات هر بخش پاسخ داده و تحلیل خود را ارائه کنید.

یک محیط grid طبق شکل۱ با ابعاد ۱۵ در ۱۵ را در نظر بگیرید. ربات ما در ابتدا در نقطه (۱۵،۱۵) قرار گرفته است. هدف ربات این است که به خانه (۱،۱) برود. ربات مورد نظر ما در هر استیت قادر به انجام ۹ عمل مختلف میباشد. ۸ عمل برای جابجا در جهت های ۸ گانه (حرکت های مورب مجاز است) و یک عمل برای باقی ماندن در نقطه فعلی. برای ربات دو سری مجموعه استیت داریم. مجموعه استیت‌های قابل دسترسی و مجموعه استیت‌های ممکن. همسایه‌ی ممکن همسایه‌ای است که خارج از محدوده محیط نباشد و مانع نباشد. همسایه ی در دسترس همسایه ای است که با یکی از اکشن های ممکن بتوان به آن رسید. برای مثال در موقعیت ابتدایی ربات نقاط قرمز و آبی نشان داده شده نقاط قابل دسترسی هستند. و نقاط آبی نشان داده شده نقاط قابل دسترس و ممکن‌اند. دقت کنید که نقاط قرمز با حاشور مشکی مانع اند (۱۴و۲) و (۱۴و۳) همچنین تمامی استیت های موجود در محیط که مانع نیستند نیز نقاط ممکن می‌باشند.



شکل ۱ – ربات در ابتدا در خانه بنفش – هدف رسیدن به خانه سبز – نقاط مشکی مانع

در انجام هر اکشن ربات با احتمال p به جهت انتخابی میرود و در غیر این صورت به یکی از همسایگان "ممکن و در دسترس" لیز میخورد. برای انجام هر حرکت بخاطر وجود اصطکاک انرژی مصرف شده و در نتیجه پاداش منفی ای درنظر گرفته شده است. همچنین هنگام برخورد با مانع هزینه‌ی برخورد با مانع نیز

برای ربات در نظر گرفته شده است. همچنین در هنگام رسیدن به هدف ربات پاداش دریافت میکند. دقت کنید که اگر ربات قصد به رفتن به مانع را داشته باشد ریوارد منفی را میگیرد ولی استیت آن با احتمال p تغییر نمی‌کند یا با احتمال دیگر به یکی از همسایه های ممکن و در دسترس کنونی لیز میخورد.
حالت های زیر حالت‌های ممکن در محیط هستند:

- حالت پایه:
  - احتمال انجام اکشن و رفتن به استیت بعدی برابر ۰,۸
  - احتمال لیز خوردن ربات و رفتن به یک خانه "ممکن و در دسترس" یا ماندن در استیت فعلی برابر ۰,۲ تقسیم بر تعداد آن هاست
  - هزینه برخورد با مانع ۱-
  - هزینه اصطکاک برابر با ۰,۰۱ -
  - پاداش رسیدن به خانه هدف برابر با ۵
- حالت بدون اصطکاک : همانند حالت پایه می باشد با این فرق که هزینه اصطکاک برابر با ۰ در نظر گرفته شود. در صورت برخورد با مانع پاداش منفی ۰,۰۱ و پاداش رسیدن به خانه هدف را برابر با ۵ برای ربات در نظر گرفته شود.
- حالت با اصطکاک زیاد: همانند حالت پایه میباشد با این تفاوت که هزینه اصطکاک برابر با ۱- (برای همه اکشن‌ها به جز اکشن ماندن در خانه)، هزینه برخورد با مانع برابر با ۱۰- و پاداش رسیدن به خانه هدف را برابر با ۵ در نظر بگیرید.

با توجه به MDP تعریف شده، توابع مشخص شده در فایل تمرین را کامل نموده تا مراحل زیر را پیاده سازی کرده و به سوالات مربوطه پاسخ دهید. لازم به ذکر است که اگر فرمت ارائه شده در توابعی که دارای doc string اند را رعایت ننمایید نصف نمره از شما کاسته خواهد شد.

۱- سیاست بهینه را برای حالت پایه با استفاده از روش policy iteration به دست آورید. مقدار discount factor برابر با ۰,۹ در نظر گرفته شود.
۲- سیاست بهینه را با روش policy iteration را برای حالت بدون اصطکاک به دست آورده و با نتایج مرحله دوم مقایسه کنید. در مقایسه طول مسیر طی شده توسط ربات را در نظر داشته باشید. مقدار discount factor برابر با ۰,۹ در نظر گرفته شود.
۳- حال حالت با اصطکاک زیاد را در نظر گرفته و سیاست بهینه را با استفاده از روش policy iteration به دست آورده و با دو حالت قبل مقایسه کنید. مقدار discount factor برابر با ۰,۹ در نظر گرفته شود.
۴- با توجه به مراحل ۲ و ۳ بهترین حالت برای ریوارد محیط را در نظر گرفته و نقش تفاوت مقدارهای مختلف برای discount factor را برای ۴ مقدار مختلف در مسئله بررسی کنید. تحلیل خود از نتایج به دست امده و همچنین آینده نگری ربات با توجه به discount factor تعیین شده را بررسی کنید.

۵- الگوریتم value iteration را برای محیط داده شده اجرا کرده و نتایج به دست آمده را با بهترین نتیجه قسمت ۵ مقایسه کنید.

۶- (امتیازی) دلیل تفاوت بخش ۱و ۲و ۳ را بررسی کنید و راه حلی برای آن ارائه 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.
- 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 سوال ها 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 ایمیلهای اعلام شده با دستیاران آموزشی درارتباط باشید.

بنفشه کریمیان – banafshehkarimian@ut.ac.ir

امیرحسین مصباح – amir.mesbah@ut.ac.ir

شاد و سلامت باشید :)