به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس سیستم‌های هوشمند

تمرین شماره 6

دی ماه 1401

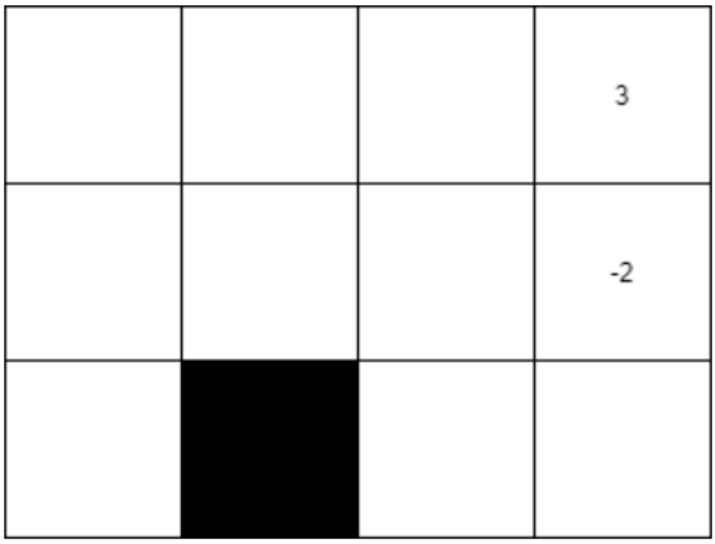
فهرست سوالات

سوال 1: یادگیری تقویتی مبتنی بر مدل (تحلیلی) :

دراین سوال قصد داریم با الگوریتم تصمیم مبتنی بر تکرار[1] از روش های یادگیری تقویتی مبتنی بر مدل[2] درقالب یک سوال تحلیلی آشنا شویم و منطق الگوریتم را بصورت دقیق تر بررسی نماییم.

سه مرحله از الگوریتم *Policy-Iteration* را روی جدول زیر اجرا کنید.

فرض نمایید به دلیل وجود نامعینی، به احتمال 0.6a به سمت جهت دلخواه و با احتمال 0.2b به دو جهت مجاور خواهیم رفت. مقدار α را در الگوریتم 0.2 فرض نمایید و همچنین اینکه عامل[3] برای حرکت کردن مجازات نخواهد شد.



شکل 1: جدول سوال یک

* توجه نمایید که a رقم یکان و b رقم دهگان شماره دانشجویی شما است.
* به عنوان نمونه اگر شماره دانشجویی شما به 54 ختم میشود عامل با احتمال 0.64 به سمت جهت دلخواه و با احتمال 0.25 به دو جهت مجاور خواهد رفت.

[1] Policy Iteration
[2] Model-Based Reinforcement Learning
[3] Agent

سوال 2: یادگیری تقویتی مبتنی بر مدل (پیاده سازی) :

دراین سوال قصد داریم که در قالب حل مساله زیر با الگوریتم *Value-Iteration* آشنا شویم.

تعریف مساله:

فردی تصمیم دارد که در یک فستیوال بخت آزمایی شرکت نماید. شرکت دراین بخت آزمایی باین گونه است که وی میبایست بر روی توالی پرتاب یک سکه بگونه ای شرط بندی کند که اگر سکه شیر بیاید به همان اندازه ای که روی آن سمت سکه شرط بسته هست پول برنده شود و درصورتی که سکه خط بیاید به همان اندازه پول از دست بدهد. این روند مادامی ادامه می یابد که یا او به هدف خود که برنده شدن $100\$$ است دست پیدا کند یا اینکه تمام پول خود را ببازد.

در مدل سازی مساله بالا لطفا به نکات زیر توجه نمایید:

- دراین مساله احتمال شیرآمدن سکه را با p_h نمایش خواهیم داد.
- شرکت کننده درهر مرحله پرتاب تصمیم میگیرد که چه مقدار از سرمایه خود را شرط ببندد و این مقدار پول، یک مقدار صحیح است.
- درمدلسازی مساله، پاداش برای هر گذاری که منجر به پیروزی شود برابر $+1$ خواهد بود و در بقیه موارد برابر 0 خواهد بود.
- ازآنجا که دراین مسئله، با یک زنجیره مارکوفی بدون تخفیف[1] روبه رو هستیم، مقدار گاما 1 لحاظ خواهد شد.

عاملی طراحی نمایید که با الگوریتم *Value-Iteration*، سیاست بهینه ای را برای شیوه شرط بندی شرکت کننده درهر مرحله ارایه دهد. (لطفا این پیاده سازی و ترسیم نمودارها را برای دو مقدار $p_h = 0.25$ و $p_h = 0.55$ انجام دهید.)

همچنین اینکه دو منحنی سیاست بهینه برحسب سرمایه و همچنین منحنی مقدار ارزش های گذار[2] برحسب سرمایه را نیز بصورت گرافیکی ترسیم نمایید.

* برای شهود بهتر در تفسیر نمودارها توصیه میشود منحنی سیاست بهینه برحسب سرمایه را با دستور `plt.bar()` ترسیم نمایید. (سعی نمایید که الگوی سیاست بهینه بدست آمده را بازای دومقدار p_h مختصرا تفسیر نمایید.)

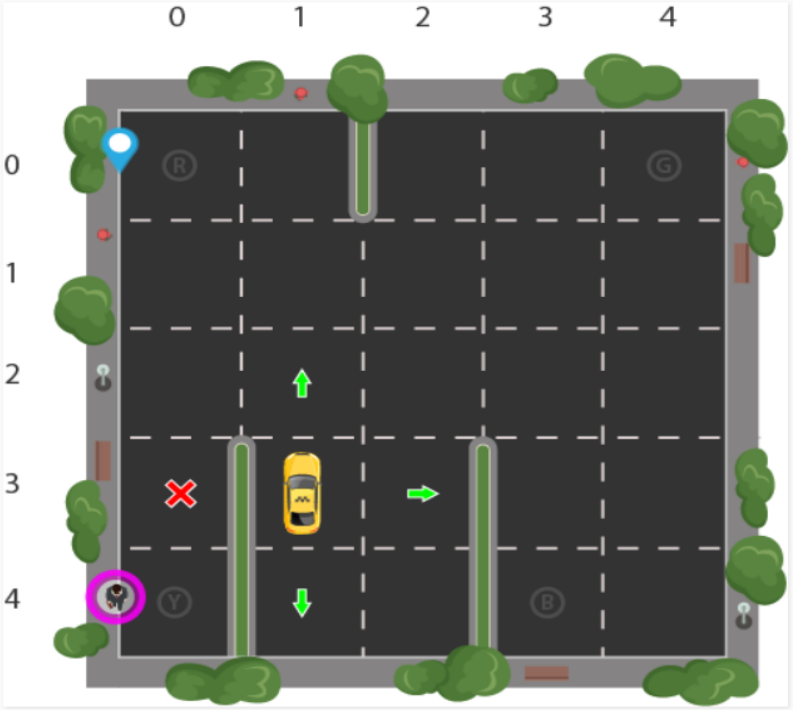
* همچنین لطفا توجه نمایید که درک مساله و مدلسازی درست فضای حالت[3] و فضای اقدام[4] دراین مساله بسیار حائز اهمیت میباشد، بنابراین درگزارشکار خود بطور شفاف و دقیق جزییات مدلسازی و پیاده سازی الگوریتم را توضیح دهید.

[1] Undiscounted MDF
[2] Value Estimates
[3] State Space
[4] Action Space

سوال 3: یادگیری تقویتی غیرمبتنی بر مدل(پیاده سازی) :

دراین قسمت قصدداریم در قالب محیط یک بازی به بررسی الگوریتم های یادگیری تعاملی بپردازیم و بررسی کنیم با حرکت هوشمندانه در محیط نسبت به حرکت رندوم تا چه اندازه میتوانیم بصورت بهینه تر به مقصد که دراین مثال رساندن مسافر به مقصد مورد نظر میباشد دست یابیم.

محیط بازی مورد نظر بصورت زیر میباشد . هدف دراین محیط این است که تاکسی زرد مسافر را از یک لوکیشن سوار کند و در یک لویکیشن دیگر پیاده نماید.



شکل 2: محیط بازی

تاکسی زرد درحرکت خود تابع یکسری قوانین است که میبایست رعایت شوند: 1) مورد اول اینکه راننده درمورد وقت مسافران اهمیت بخرج دهد و درکمترین زمان ممکن آنها را در مقصد موردنظر پیاده کند. 2) مسافران را در مقصد درست پیاده نماید.

راننده تاکسی زرد به بازای هر مسافری که آن را با موفقیت به مقصد مورد نظرش برساند یک پاداش مثبت برابر 20+ دریافت خواهد کرد و درصورتی که مسافر را به مقصد اشتباه برساند جریمه خواهد شد و پاداشی منفی برابر 10- خواهد گرفت و درنهایت نیز از آنجایی که سوخت تاکسی محدود است و طولانی شدن مسیر منجر به تمام شدن بنزین خواهد شد لذا هر حرکت ای که منجر به رسیدن به مقصد نشود نیز راننده را متحمل جریمه خواهد کرد و درطی آن، عامل پاداشی منفی برابر 1- خواهد گرفت. همچنین اینکه اگر دقت کنید راننده درمسیر خود دیوارهایی را میبیند که درصورتی که به آنها برخورد کند جریمه خواهد شد و پاداشی منفی برابر 1- دریافت خواهد کرد و در مکان خود باقی خواهد ماند.

همچنین لطفا توجه نمایید که راننده تاکسی در مجموع شش حرکت {حرکت به بالا، حرکت به چپ، حرکت به راست، حرکت به پایین، سوار کردن مسافر و پیاده نمودن مسافر} را میتواند انجام دهد.

برای استفاده از محیط پیاده سازی شده، کافی است فایل قرار داده شده در پوشه تمرین را به کدهای خود اضافه نمایید. به علاوه فایلی به عنوان راهنما هم قرار داده شده که نشان میدهد چگونه میتوانید به فضای حالت و عمل دسترسی بیابید.

درطول کار با کتابخانه OpenAi Gym برای این محیط بازی به نکات زیر توجه نمایید:

- R,G,Y,B چهار لوکیشنی هستند که هم مسافران درآنجا سوار میشوند و هم میتوانند مقصد نهایی مسافران باشند.
- دراین محیط بصورت پیشفرض لوکیشن با رنگ آبی نشان دهنده مکان مسافر فعلی و لوکیشن با رنگ بنفش نشان دهنده مقصد مسافر فعلی است.
- همچنین اینکه مادامی که تاکسی خالی از مسافر باشد با رنگ زرد درنقشه به نمایش درخواهد آمد و مادامی که حاوی مسافر باشد با رنگ سبز به نمایش درخواهد آمد.

الف) حل محیط بازی بدون استفاده از روش Q-Learning و مبتنی بر پیمایش رندوم

درابتدا قصد داریم تا با استفاده از یک loop بینهایت کاملا بصورت رندوم در محیط گام برداریم تا منجر به رسیدن یک مسافر به مقصد درست شود. این پیاده سازی را بر مبنای پیماش کاملا رندوم انجام دهید و تعداد گام ها و مقدار کل جریمه را درطول این پیاده سازی گزارش کنید.

در پیاده سازی این قسمت برای گام نهادن درمحیط میتوانید از دستور `env.action_space.sample()` استفاده نمایید.

* درطول این پیاده سازی برای نمایش گرافیکی حرکت تاکسی میبایست درطول loop اطلاعات هربار *render* شدن محیط را در یک دیکشنری ذخیره کنید تا بتوانید پس از پایان حلقه بصورت گرافیکی حرکت تاکسی را به تصویر بکشید.

ب) حل محیط بازی با استفاده از روش Q-Learning و مبتنی بر پیمایش هوشمندانه

برای محیط بیان شده، الگوریتم یادگیری بر اساس معیار Q را شبیه سازی کنید. (نتایج را با حالتی که از الگوریتم Q استفاده نمیکردیم مقایسه نمایید.)

* پارامترهایی مانند میانگین پاداش درهرجابه جایی[1]، میانگین تعداد گام ها در هر سفر[2] و میانگین کل مقدارجریمه ها درطول هر اپیزود[3] میتواند درمقایسه نتایج مفید باشد.

با استفاده از پیاده سازی بالا باین سوال پاسخ دهید. دقت کنید که برای پاسخ خود دلیل ارائه کنید. آیا میتوانید با تغییردادن پاداشها به مقادیر معقول به همگرایی سریعتری برسید؟ منطق خود را برای این تغییرات ذکر کرده و آن را شبیه سازی کنید.

نکات سوال سوم:

- برای کدهایی که ضمیمه میکنید، حتما گزارشکار دقیق با ذکر جزییات پیاده سازی بنویسید. (این گزارش ها معیار تفاوت کد شما با مدل های مشابه موجود در اینترنت خواهد بود.)
- از پاسخ های روشن در گزارشکار خود استفاده نمایید و تمام فرضیات خود را به طور شفاف بیان نمایید.
- همچنین توجه نمایید که ایده های خلاقانه در *tune* کردن هایپرپارامترها و بکارگیری بهینه تر الگوریتم Q-Learning دراین سوال میتواند نمره امتیازی در بر داشته باشد و یا اینکه کاستی نمرات شما دراین بخش را جبران نماید.
- بطور ویژه دراین تمرین درک و مدلسازی درست مسئله ها اهمیت بیشتری نسبت به پیاده سازی صرف خواهد داشت که درپی آن گزارشکار دقیقتری را نیز میطلبد پس لطفا سعی کنید درگزارشکار خود بصورت شفاف جزییات مدلسازی و پیاده سازی خود را ذکر نمایید.

[1] Average Rewards per move
[2] Average number of timesteps per trip
[3] Average number of penalties per episode

نکات تحویل:

- **مهلت تحویل این تمرین 10 بهمن میباشد.**
- **انجام این تمرین به صورت یک نفره است.**
- **برای انجام این تمرین تنها مجاز به استفاده از زبان برنامه نویسی پایتون هستید.**
- **در صورت وجود تقلب نمره تمامی افراد شرکت کننده در آن -100 لحاظ میشود.**
- **لطفا پاسخ تمرین خود را (به همراه کد/گزارش سوال کامپیوتری) به صورت زیر در صفحه درس آپلود نمایید:**

HW [HW number] _ [Last name] _ [Student number].zip

- **در صورت وجود هر گونه ابهام یا مشکل لطفا به مسئولان حل تمرین ایمیل بزنید.**

Question #1 : Erfan Hajihashemi + Oveys Delafrooz
Question #2 : Oveys Delafrooz + Mohammad Heydari
Question #3 : Mohammad Heydari + Erfan Hajihashemi