



دانشکده فنی دانشگاه تهران دانشکده برق و کامپیوتر

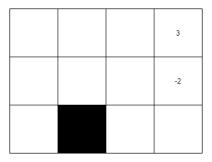
تمرین ۵ سیستمهای هوشمند

رایانامه sharifi.sina1377@gmail.com hesam.77s@gmail.com

طراحان: سینا شریفی امیرحسام سلیم نیا

- دانشجویان عزیز، قبل از پاسخ گوئی به سوالات به نکات زیر توجه کنید:
- ۱. کدها و گزارش باید با الگو **IS_HW5_StudentNumber.zip** در محل تعیین شده آپلود شوند.
- ۲. گزارش کار شما نیر از معیار های ارزیابی خواهد بود، در نتیجه زمان کافی برای تکمیل آن اختصاص دهید.
- ۳. در صورتی که از Jupyter Notebook نیاز به ارسال جداگانه کدها و گزارش نیست و هردو را میتوانید در یک فایل Notebook خود را نیز همراه فایل Notebook ارسال Notebook ارائه دهید. حتما خروجی html یا pdf فایل Notebook خود را نیز همراه فایل نمایید.
- ۴. شما ميتوانيد سوالات خود را از طريق ايميل hesam.77s@gmail.com و hesam.77sسgmail.com بيرسيد.

- ۱. به سوالات تشریحی زیر پاسخ دهید.
- (آ) مفهوم پاداش لحظه ای و تاخیری را در یادگیری تقویتی توضیح دهید.
- (ب) در چه مواردی ترجیح میدهیم از روش های مستقل از مدل بجای روش های مبتنی بر مدل استفاده کنیم؟ توضیح دهید.
- (ج) سه مرحله از الگوریتم $policy\ iteration$ را روی جدول زیر اجرا کنید. فرض کنید در انتخاب جهت، به علت وجود نامعینی، به احتمال 0.6 به سمت جهت دلخواه و 0.2 به دو جهت مجاور میرویم. علاوه براین مقدار 0.2 و برای حرکت، جایزه ی منفی در نظر نمیگیریم.



شکل ۱: جدول سوال یک

مى باشد. Q-Learning مى باشد. Q-Learning مى باشد.

در این سوال، یک عامل داریم که در یک فضای n*n قرار گرفته، این عامل از یک نقطه مشخص شروع کرده و هدف این است که به نقطه مقصد برسد. این عامل میتواند m*p جهت (U,D,L,R) را انتخاب کند ولی با انتخاب هر جهت، ممکن است به احتمال P=0,05 به مقصد دلخواه نرسد و به طور تصادفی به یکی از m*p خانه دیگر برود.

غیر از خانه نهایی که دارای جایزه R1=100 است، تعدادی خانه دیگر نیز در فضا وحود دارد که عبور از آن ها دارای جایزه c3=-100 و c2=-20 و c3=-0.00 میباشد. نقشهی میباشد. تعدادی از خانه ها نیز دارای هزینه ی ENV.map قرار داده شده است که خانه ها با توجه به جدول زیر مقدار دهی شده اند. c3=-100 و c3=-100 قرار داده شده است که خانه ها با توجه به جدول زیر مقدار دهی شده اند. c3=-100 و c3=-100

0	state with cost c1
1	starting point
2	state with cost c2
3	state with cost R1
4	state with cost R2
5	state with cost R3

شکل ۲: جدول سوال دو

eta=1 پس از حداقل Softmax دور یادگیری با روشهای Q-Learning و سیاست تصمیم گیری Softmax با ثابت Q-Learning یک بار به ازای R2=30 و بار دیگر به ازای R2=30 موارد زیر را بدست آورده و گزارش کنید:

- (آ) نمودار مجموع پاداش دریافتی در طول مسیر را بر حسب epoch محاسبه کرده و رسم نمایید.
- (ب) سیاست بهینه در شرایط حریصانه (مسیر بهینه) را برای هر دو مقدار R2=30 و R2=30 محسابه کنید.
 - جیان کنید. $\epsilon-greedy$ مزیت و تفاوت سیاست Softmax در مقایسه با روش

موفق باشيد