



تمرین سری پنجم: فرآیندهای مارکوفی و یادگیری تقویتی

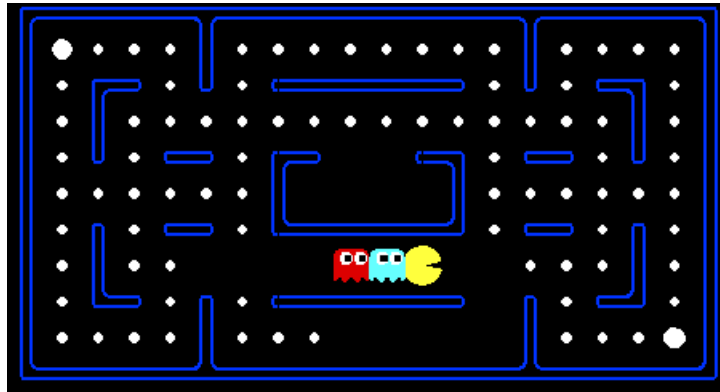
لطفاً به نکات زیر توجه کنید:

- مهلت ارسال این تمرین تا ۱۳ خرداد است.
- در صورتی که به اطلاعات بیشتری نیاز دارید می‌توانید به صفحه‌ی تمرین در وب‌سایت درس مراجعه کنید.
- این تمرین شامل سوال‌های برنامه‌نویسی می‌باشد، بنابراین توجه کنید که حتماً موارد خواسته‌شده در سوال را رعایت کنید. در صورتی که به هر دلیلی سامانه‌ی داوری نتواند آن را اجرا کند مسئولیت آن تنها به عهده‌ی شماست.
- ما همواره هم‌فکری و هم‌کاری را برای حل تمرین‌ها به دانشجویان توصیه می‌کنیم. اما هر فرد باید تمامی سوالات را به تنهایی تمام کند و پاسخ ارسالی حتماً باید توسط خود دانش‌جو نوشته‌شده باشد. لطفاً اگر با کسی هم‌فکری کردید نام او را ذکر کنید. در صورتی که سامانه‌ی تطبیق، قلبی را تشخیص دهد متأسفانه هیچ مسئولیتی بر عهده‌ی گروه تمرین نخواهد بود.
- لطفاً برای ارسال پاسخ‌های خود از راهنمای موجود در صفحه‌ی تمرین استفاده کنید.
- هر سوالی درباره‌ی این تمرین را می‌توانید از دستیاران حل تمرین بپرسید.

- آدرس گروه درس: <https://groups.google.com/forum/#!forum/ai972>

- صفحه تمرین: <https://quera.ir/course/assignments/9648/problems>

سوال های عملی



مقدمه :

شما در این پروژه باید به روش Q-learning پکمن خود را آموزش دهید. سورس کد از [اینجا](#) قابل دانلود است.

سوال اول (۳۰ نمره)

شما باید با استفاده از Q-learning یک عامل بسازید که ساختار ساده‌ای داشته ولی از طریق ارتباط برقرار کردن با دنیای بیرون و آزمون و خطا یادگیری را انجام دهد. این کار را با تابع زیر انجام می‌شود.

```
update(state, action, nextState, reward)
```

یک نمونه اولیه برای Q-Learning در QLearningAgent در فایل `qlearningAgents.py` تعریف شده است و شما می‌توانید آن را با آپشن `'-a q'` انتخاب کنید. برای این سوال باید توابع `update`, `computeValueFromQValues`, `getQValue` و `computeActionFromQValues` را پیاده سازی کنید.

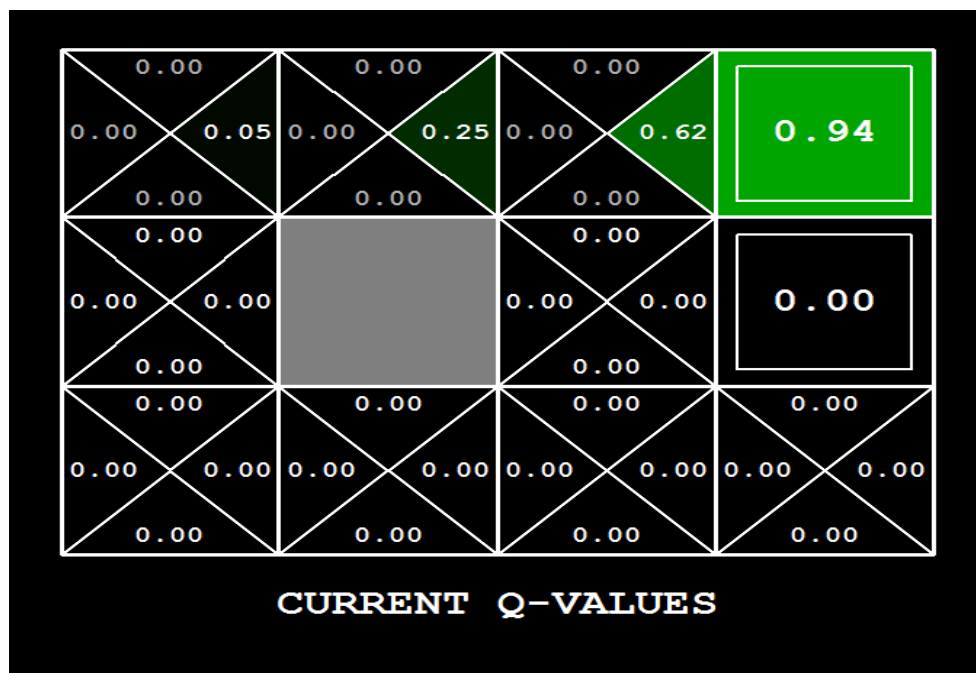
یادداشت: برای `computeActionFromQValues` باید برای رفتار بهتر مدل، به صورت رندوم عمل کنید. تابع `random.choice` کمک کننده خواهد بود. در مورد بعضی حالات، عملی (اکشنی) که عامل شما تا به حال انجام نداده است هم یک Q-value دارد که صفر است. پس زمانی که مابقی اکشن‌ها مقادیر منفی بگیرند، این اکشن بهترین به حساب می‌آید.

مهم: مطمئن شوید در توابع `computeValueFromQValues` و `computeActionFromQValues`، شما فقط با `getQValue` دسترسی به Q-values دارید.

به وسیله ی *update* در لحظه ی Q-learning، شما می توانید فرایند یادگیری Q-learning را به صورت کنترل دستی ببینید

```
python gridworld.py -a q -k 5 -m
```

یادآوری می‌کنیم که k -تعداد اپیزودهایی است که عامل شما نیاز دارد تا یادگیری‌اش را کامل کند. راهنمایی: برای راحتی debugging، می‌توانید با پارامتر `--noise 0.0` نویز را کم کنید. اگر به صورت دستی پکمن را شمال و شرق ببرید در ۴ اپیزود، Q -values شما مانند زیر خواهد شد.



سوال دوم (۲۵ نمره)

با اضافه کردن سیستم انتخاب اکشن epsilon-greedy در `getAction`، Q-learning خود را کامل کنید. یعنی گاهی اوقات ایجنت یک حرکت رندوم را انتخاب می کند و در بقیه مواقع از شیوهی قبلی انتخاب بهترین Q-values اکشن را انتخاب می کند.

```
python gridworld.py -a q -k 100
```

ممکن است نتیجهی کلی بعد از مرحلهی آموزش در این سوال نسبت به سوال قبلی، به دلیل وجود انتخاب تصادفی افت داشته باشد. اما در نهایت جدول حاصله بسیار دقیق تر میشود.

شما می توانید یک المان را از لیست با توزیع احتمال نرمال به کمک تابع `random.choice` انتخاب کنید. همچنین می توانید یک متغیر باینری را با احتمال موفقیت p به کمک تابع `util.flipCoin(p)` شبیه سازی کنید که اگر احتمال p بیاید خروجی آن `True` میشود و اگر $1-p$ بیاید خروجی `False` است.

پیاده سازی خود را به روش زیر تست کنید:

```
python autograder.py -q q5
```

بدون هیچ تغییری در کد، شما می توانید روبات کرالر Q-learning را به این صورت اجرا کنید:

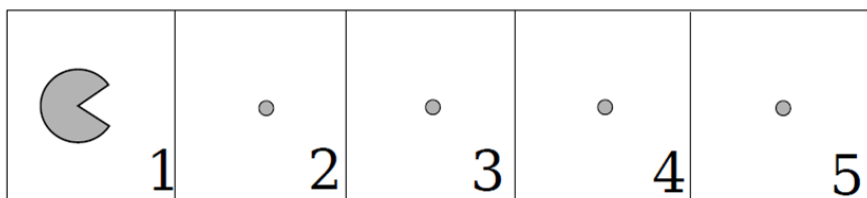
```
python crawler.py
```

اگر دستور بالا کار نکرد، احتمالاً بخشی از پیاده سازی شما محدود به `GridWorld` است. شما باید آن را طوری تغییر دهید که برای تمام MDP ها عمومیت داشته باشد.

در نظر داشته باشید که `delay` پارامتر شبیه سازی است. ولی `learning rate` و `epsilon` پارامترهای الگوریتم یادگیری شما هستند و `discount factor` ویژگی محیط است.

سوال های تئوری

سوال اول (۱۵ نمره)



پکمن در یک مستطیل ۱x۵ همانند شکل است. در خانه های ۱ تا ۴ عملیات ممکن برای عامل رفتن به سمت راست R یا پرواز F است. عمل R پکمن را به خانه سمت راست خانه ای که در آن است برده و یک نقطه را می خورد و F آن را به خانه پایانی برده و بازی را به اتمام می رساند. تنها عمل ممکن برای پکمن در خانه ۵ پرواز می باشد. خوردن هر نقطه ۱۰+ امتیاز و امتیاز پرواز کردن ۲۰+ است.

Policy های زیر را در نظر بگیرید:

$$\pi_0(s)=F \text{ for all } s$$

$$\pi_1(s)=R \text{ if } s < 3, \text{ else } F$$

$$\pi_2(s)=R \text{ if } s < 5, \text{ else } F$$

الف) با در نظر گرفتن discount=1 مقادیر زیر را محاسبه کنید:

- I. $V^{\pi_0}(1)$
- II. $V^{\pi_1}(2)$
- III. $V^{\pi_2}(1)$
- IV. $V^*(1)$
- V. $V^*(4)$

ب) به ازای چه مقادیری از discount π_0 از π_1 و π_2 بهتر است؟

ج) به ازای چه مقادیری از discount π_1 از π_0 و π_2 بهتر است؟

د) به ازای چه مقادیری از discount π_2 از π_0 و π_1 بهتر است؟


سوال دوم (۱۵ نمره)

فرض کنید $MDP(S, A, T, R, \gamma, S_0)$ به شما داده شده است و قرار است راهبرد بهینه را برای این مسئله پیدا کنید. اما به جای اینکه بتوانید Action های خود را آزادانه انتخاب کنید، در هر مرحله باید یک سکه بیاندازید. اگر سکه شیر آمد می‌توانید Action خود را آزادانه انتخاب کنید، اگر خط آمد یک Action به صورت تصادفی از بین Action های موجود برای شما انتخاب می‌شود. یک مسئله $MDP(S', A', T', R', \gamma', S'_0)$ جدید با محدودیت جدید تعریف کنید که به راهبرد بهینه دست یابید. (راهنمایی: برای تعریف یک مسئله MDP جدید، لازم است پارامترهای جدید را بر حسب پارامترهای قبلی مسئله بنویسید)

سوال سوم (۱۵ نمره)

شکل زیر را در نظر بگیرید. pacman تلاش می‌کند تا policy بهینه را یاد بگیرد. اگر وارد یکی از خانه‌های رنگ شده شود بازی به اتمام می‌رسد. حرکت در ۴ جهت بالا پایین چپ و راست می‌باشد. پکمن از خانه (1,3) شروع میکند.

با فرض اینکه distance factor = 0.5 و Q-Learning rate = 0.5 باشد، به سوالات پاسخ دهید:

3		-80	+100
2			
1	+25	-100	+80
	1	2	3

الف) مقدار V^* را برای خانه های زیر پیدا کنید:

$V^*(3,2)$

$V^*(2,2)$

$V^*(1,2)$

جدول زیر حرکت های پکمن را در فضای بالا نشان می دهد هر خط دارای tuple شامل (s, a, s', r) است.

Episode 1	Episode 2	Episode 3
$(1,3), S, (1,2), 0$	$(1,3), S, (1,2), 0$	$(1,3), S, (1,2), 0$
$(1,2), E, (2,2), 0$	$(1,2), E, (2,2), 0$	$(1,2), E, (2,2), 0$
$(2,2), S, (2,1), -100$	$(2,2), E, (3,2), 0$	$(2,2), E, (3,2), 0$
	$(3,2), N, (3,3), +100$	$(3,2), S, (3,1), +80$

ب) با استفاده از Q-Learning مقادیر Q-Value زیر را بدست آورید

$Q((2, 2), E)$

$Q((1,2), S)$

$Q((3,2), N)$