نیم‌سال دوم سال ۹۷-۹۸

جواب تمرین سری پنجم: فرآیندهای مارکوفی و یادگیری تقویتی

لطفاً به نکات زیر توجه کنید:

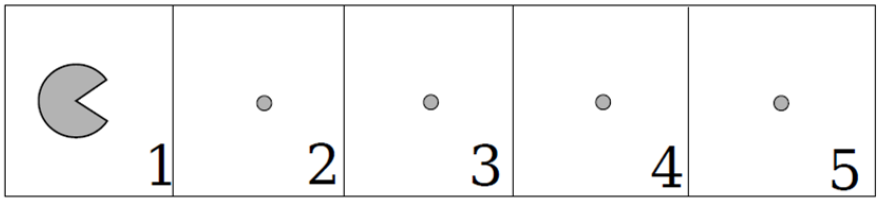
- برای برخی از سوالات، جواب‌های دیگری را نیز می‌توان متصور بود و این جواب‌ها تنها جواب‌های درست مسائل نیستند.

- آدرس گروه درس: https://groups.google.com/forum/#!forum/ai972
- صفحه تمرین: https://quera.ir/course/assignments/9648/problems

موفق باشید

سوال‌های تئوری

سوال اول (۱۵ نمره)



پکمن در یک مستطیل ۱x۵ همانند شکل است. در خانه‌های ۱ تا ۴ عملیات ممکن برای عامل رفتن به سمت راست R یا پرواز F است. عمل R پکمن را به خانه سمت راست خانه‌ای که در آن است برده و یک نقطه را می‌خورد 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 for all s

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

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

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

I. $V^{\pi 0}(1) = 20$

II. $V^{\pi 1}(2) = 30$

III. $V^{\pi 2}(1) = 60$

IV. $V^{*}(1) = 60$

V. $V^{*}(4) = 30$

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

$$V^{\pi 0}(1) > V^{\pi 2}(1) \rightarrow 20 > 10 + 10(discount + dicount^2 + discount^3 + 2discount^4)$$

واضح است که باید discount < 0.5 باشد. از آنجا که discount عددی بین صفر و یک است، پس:

$$0 < discount < 0.5$$

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

هیچ زمانی این اتفاق نمی‌افتد.

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

با استفاده از قسمت ۲ و یک محاسبه ساده می‌توان متوجه شد که همواره این شرایط برقرا است.

سوال دوم (۱۵ نمره)

فرض کنید MDP(S, A, T, R, γ,S_0) به شما داده‌شده 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' , γ', 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)

$$S' = S$$

$$A' = A$$

$$R' = R$$

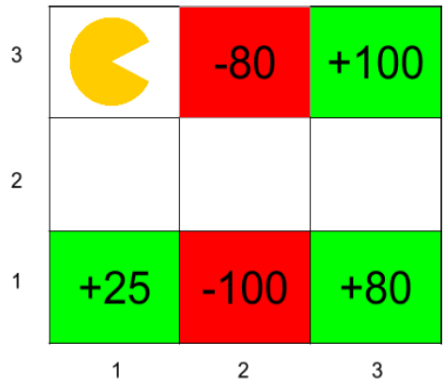
$$\gamma' = \gamma$$

$$T' = \forall s \in S, s' \in S, a \in A, T'(s,a,s') = P(heads)T(s,a,s') + P(tails)\sum \frac{1}{|A|} T(s,a,s')$$

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:



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

$V^*(3,2) = 100$ $V^*(2,2) = 50$ $V^*(1,2) = 25$

برای مثال برای $V^*(2,2)$ رفتن به خانه (3,3) بیشترین امتیاز را نصیب ما می‌کند که برابر است با

0 + discount * 100

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((3,2),N) = 50 Q((1,2),S) = 0 Q((2, 2),E) = 12.5

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(R(s,a,s') + max_{a'}Q(s,a))$$