دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری عمیق

نیم‌سال اول ۰۱-۰۲

استاد: حمید بیگی
گردآورندگان:علی سلطانی، علی قاری زاده، حسین خلیلی
بررسی و بازبینی: حسن حمیدی

تمرین سری ششم | یادگیری تقویتی | مهلت ارسال: ۱۵ بهمن

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- امکان ارسال تمرین با **تاخیر مجاز وجود ندارد.** ارسال بعد از ۱۵ بهمن بسته خواهد شد.
- هم‌کاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.
- پاسخ تمامی سوالات (تئوری و عملی) را در یک فایل فشرده به صورت $DL\_HW6\_[firstName]\_[lastName]\_[StudentId]$ نام‌گذاری کرده و ارسال کنید.

سوالات نظری (۶۰ نمره)

۱. Markov Decision Process (نمره ۱۵)

(آ) جمله زیر را اثبات یا رد کنید:
دو MDP مشابه با فاکتورهای تخفیف متفاوت حتماً آپتیمال پالیسی یکسانی دارند.

(ب) وقتی می‌خواهیم یک مسئله را با استفاده از MDP مدل کنیم، می‌توانیم به صورت‌های مختلفی عامل و محیط را تعریف کنیم. یک مسئله مطرح کنید و برای آن دو مدل مختلف طراحی کنید. کدامیک از این دو مدل برای مسئله‌تان مدل بهتری است؟ با استفاده از چه معیارهایی می‌توان تصمیم گرفت که در یک مسئله کدام مدل‌سازی بهتر است؟

(ج) تفاوت الگوریتم policy iteration با الگوریتم value iteration چیست؟ هر کدام از این الگوریتم‌ها چه محدودیت‌هایی دارند؟

۲. Multi-armed Bandits (نمره ۱۵)

فرض کنید یک مسئله‌ی bandit با ۵ بازو داریم و الگوریتم $\epsilon - greedy$ را روی آن اعمال کرده‌ایم. همچنین داریم:

$$Q_n = \frac{R_1 + R_2 + ... + R_{n-1}}{n-1}$$

همچنین تخمین‌های اولیه برابر ۰ هستند. اگر عمل‌ها و پاداش‌ها به ترتیب به شکل زیر باشند:

$$A_1 = 1, R_1 = -2$$
$$A_2 = 2, R_2 = 3$$

$$A_3 = 3, R_3 = -1$$
$$A_4 = 2, R_4 = 2$$
$$A_5 = 3, R_5 = 0$$
$$A_6 = 4, R_6 = 5$$

در برخی از مراحل حالت ϵ رخ داده و باعث انتخاب یک عمل تصادفی شده‌است. در کدام مراحل حتماً این حالت رخ داده است؟ در کدام مراحل ممکن است این حالت رخ داده باشد؟

۳. محاسبه دستی Value Iteration (نمره ۱۵)

یک فرآیند مارکوف را در نظر بگیرید که دارای سه حالت s_1, s_2, s_3 و دارای دو عمل a_1, a_2 است. احتمال انتقال و پاداش مورد انتظار برای هر حالت بصورت زیر می باشد:

$$s_1 : \{$$
$$a_1 : (\{s_1 : 0.2, s_2 : 0.6, s_3 : 0.2\}, 8.0)$$
$$a_2 : (\{s_1 : 0.1, s_2 : 0.2, s_3 : 0.7\}, 10.0)\}$$

$$s_2 : \{$$
$$a_1 : (\{s_1 : 0.3, s_2 : 0.3, s_3 : 0.4\}, 1.0)$$
$$a_2 : (\{s_1 : 0.5, s_2 : 0.3, s_3 : 0.2\}, -1.0)\}$$

$$s_3 : \{$$
$$a_1 : (\{s_3 : 1\}, 0.0)$$
$$a_2 : (\{s_3 : 1\}, 0.0)\}$$

همچنین $\gamma = 1$ است. در این سوال هدف بدست آوردن Optimal Deterministic Policy است. مقادیر خواسته شده را با استفاده از روش Value Iteraion تا دو Iteration بصورت دستی محاسبه نمایید.

(آ) مقادیر مربوط به Value Function هر حالت را در حالت ماکزیمم مقداردهی اولیه کنید. سپس مقادیر مربوط به $q_k(.,.)$ و $v_k(.)$ را با استفاده از v_{k-1} بدست آورید. در نهایت $\pi_k(.)$ را از $q_k(.,.)$ برای $k = 1$ و $k = 2$ محاسبه کنید.

(ب) ثابت کنید که $\pi_k(.)$ به ازای $k > 2$ برابر است با $\pi_2(.)$.

۴. مقاله (نمره ۱۵)

مقاله Learning Reinforcement Deep for Agents Imagination-Augmented مطالعه کرده و معماری I۲A را توضیح داده و مزایا و معایب آن را بررسی نمایید.

سوالات عملی (۴۰ نمره)

---

۵. در این تمرین قصد داریم به پیاده سازی یک Deep Q-Network بپردازیم و این شبکه را بر روی مدل فرود سفینه ارزیابی کنیم. (نمره ۲۰)

(آ) همانطور که در فایل Jupiter ارائه شده توضیح داده شده است، دقت شود که برای آموزش و تست مدل حتما از گوگل کولب استفاده کنید و قبل از اجرای کدها، حتما پکیج های مورد نیاز را نصب کنید.

(ب) فایل model.py را بگونه ای کامل کنید که بتوان از آن در شبکه ی Q استفاده کرد.

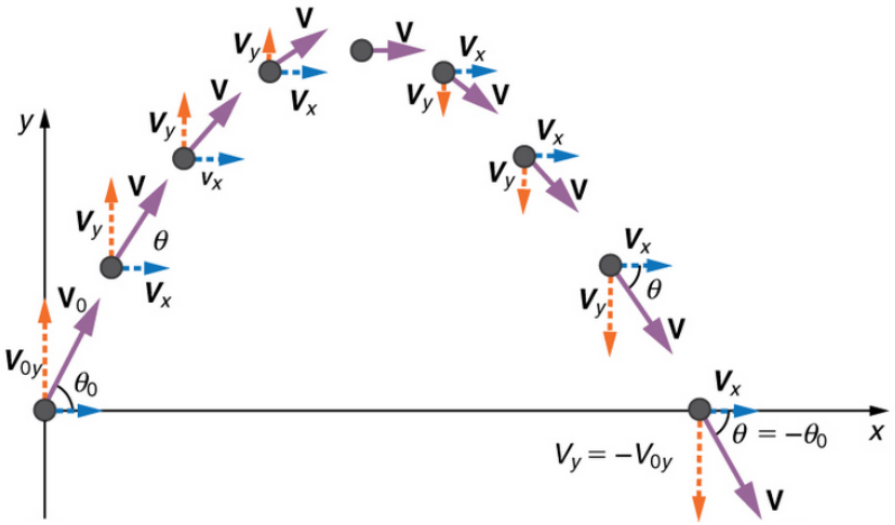
(ج) در قسمت خواسته شده در فایل Deep-Q-N.py کد را به گونه ای تغییر دهید که مقدار loss قابل محاسبه باشد.

(د) مقدار نهایی score را به ازای مقادیر مختلف Learning Rate رسم کنید.

(ه) آیا دفعات آپدیت کردن شبکه به ازای یک مقدار ثابت Learning Rate در مقدار نهایی score تاثیر گذار است؟ (نمودار مربوطه رسم شود.)

۶. هدف این سوال ایجاد یک Environment Custom در قالب کتاب خانه Gym است. در این مسئله یک توپ که تنها می تواند با زاویه θ نسبت به زمین پرتاب شود را قرار است بررسی نماییم. سرعت پرتاب توپ همیشه ثابت و برابر با v است که موقع ساخت محیط داده می شود. و هدف این است با کمترین تعداد قدم توپ را به نقطه L برسانیم فاصله L ازنقطه صفر نیز موقع ساخت محیط داده می شود.

(نمره ۲۰)



برای سادگی فرض کنید به محض برخورد پرتابه به زمین سرعت آن صفر می شود و نیاز است مجددا پرتاب شود. زاویه می تواند از صفر تا ۱۸۰ تغییر کند. برای معادلات پرتابه از این لینک کمک بگیرید.