دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی
پاییز ۱۳۹۹
استاد: محمدحسین رهبان

تمرین هفتم، بخش دوم | **یادگیری تقویتی** | مهلت ارسال: ۲۳ دی (بخش نظری)

- مهلت ارسال بخش نظری و عملی به‌ترتیب ۲۳ دی و ۱۰ بهمن، ساعت ۲۳:۵۹ خواهد بود.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین (به استثنای هفته‌ی امتحان میانترم) تا سقف پنج روز و در مجموع ۱۵ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهندبود.
- 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 منبع خارج از کتاب و اسلایدهای درس، نام هم‌فکران و آدرس منابع مورد استفاده‌برای حل سوال مورد نظر را ذکرکنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.

سوالات نظری (۴۱ نمره)

۱. (۲۱ نمره) درستی یا نادرستی عبارات زیر را مشخص کرده و به طور کامل توضیح دهید.

(آ) expectimax در حل مسائل MDP کمک‌کننده خواهد بود.

(ب) فرض کنید S_1 تا S_n نشان‌دهنده یک پروسه در MDP باشد. وقوع S_n به S_{n-2} وابسته نیست و همچنین بالعکس.

(ج) در یادگیری تقویتی، regret عمدتاً به معنی اختلاف میانگین reward کسب‌شده با میانگین optimal reward می‌باشد.

۲. (۲۰ نمره) پرواز هواپیمای بدون سرنشین [1] یکی از مسائل مورد مطالعه در حوزه RL می‌باشد که برای دانستن موقعیت آن به طور معمول دوربینی در جلوی آن نصب می‌شود. حال فرض کنید که در مسئله ماز(دیوار ها بلند هستند) یک UAV قرار داده شده است و به دنبال این هستیم که با کمترین میزان برخورد به دیوارها، از ماز خارج شود. همچنین در محیط مربوطه تنها ۲ نوع گوشه ۹۰ و ۴۵ درجه وجود دارد. در این محیط ، S, A و Reward Function را چگونه توصیف می‌کنید؟ به طور کامل توضیح دهید.

سوالات عملی (۳۰ + ۲۰ نمره)

۱. (۳۰ نمره)
کتابخانه OpenAI Gym شامل مجموعه‌ای از محیط‌های یادگیری تقویتی در زبان پایتون است. در این سوال محیط Mountain Car از این مجموعه را مورد بررسی قرار می‌دهیم. Mountain Car مسئله‌ای از یادگیری تقویتی است که هدف آن یادگیری سیاستی برای صعود ماشین از تپه‌ای شیب‌دار و رسیدن به هدف مشخص شده با پرچم است. همچنین موتور ماشین به اندازه کافی قدرتمند نیست تا بتواند مستقیما از تپه سمت راست صعود کند بنابراین باید با صعود از تپه سمت چپ شتاب کافی را کسب کند.
در این مسئله، حالت ماشین با آرایه ای از جایگاه و سرعت آن مشخص می شود. محدوده جایگاه و سرعت ماشین را در جدول زیر مشاهده می کنید.

[1] Unmanned Aerial Vehicle

Num	Observation	Min	Max
0	position	-1.2	0.6
1	velocity	-0.07	0.07

عامل هوشمند در هر مرحله، مجاز به انجام سه حرکت است: no push, push right و push left . حرکت عامل به محیط داده شده و محیط حالت بعد را به همراه پاداش حرکت برمیگرداند. برای هر گامی که ماشین به هدف نمی رسد، هزینه ۱ − درنظر گرفته شده است.

اکنون شما باید با استفاده از Q-learning سیاست بهینه را در هر حالت بیابید. برای انجام اینکار بایستی ۴ تابع به نام‌های discretize_state (برای گسسته کردن فضای پیوسته)، get_action (گرفتن اکشن)، update_q (به‌روزرسانی q_valueها با استفاده از اکشن انجام‌شده) و q_learning (که فرآیند training را تشکیل می‌دهد) را پیاده‌سازی کنید. برای پیاده‌سازی می‌توانید از کد موجود در این لینک استفاده کنید. (هرچند که برای بخش امتیازی، قضاوت براساس تابع score موجود در لینک می‌باشد)

۲. (۲۰ نمره) (امتیازی)
در این بخش انتظار می‌رود که بتوانید با بهینه‌تر کردن پیاده‌سازی بخش اول و یا استفاده از متدهای یادگیری عمیق(مانند Deep-Q Network که می‌توانید از این لینک برای مطالعه بیشتر استفاده کنید) نتایج خود را بهبود بخشیده و در نهایت خروجی تابع score در این لینک حداقل ۱۳۵ − شود.