بسم الله الرحمن الرحیم

دانشگاه علم و صنعت ایران

بهار ۱۴۰۰

مبانی یادگیری عمیق | تمرین سری دوازدهم | تحویل: دوشنبه ۲۴ خرداد

---

۱. به سوالات زیر در رابطه با یادگیری تقویتی پاسخ بدهید.

الف) تفاوت میان Exploration و Exploitation چیست؟ آیا عاملی که فقط Exploit و یا فقط Explore می‌کند، می‌تواند موفق عمل کند؟ توضیح دهید.

ب) تفاوت میان ارزش وضعیت و پاداش را توضیح دهید.

پ) فرض کنید در یک مسئله *episodic* پاداش‌های زیر دریافت شود و این قسمت در $T = 5$ به پایان برسد.

$$R_1 = 2, R_2 = 0, R_3 = -1, R_4 = 2, R_5 = 8$$

مقادیر بازده (return) برای تمام گام‌ها را به ازای $\gamma = 0.5$ و $\gamma = 1.0$ محاسبه کنید.

۲. همانطور که می‌دانیم در روش‌های یادگیری Temporal Difference و Monte Carlo نیازی به داشتن مدلی از محیط نداریم، یعنی از قبل مشخص نیست با انجام یک عمل، با چه احتمالی به کدام حالت می‌رویم و چه پاداشی می‌گیریم، به همین دلیل به این روش‌ها model-free می‌گویند. در مورد این دو روش به سوالات زیر پاسخ دهید.

الف) به نظر شما کدام یک از این دو روش برای مسائل اپیزودیک و کدام یک برای مسائل ادامه‌دار (continuous) مناسب است؟

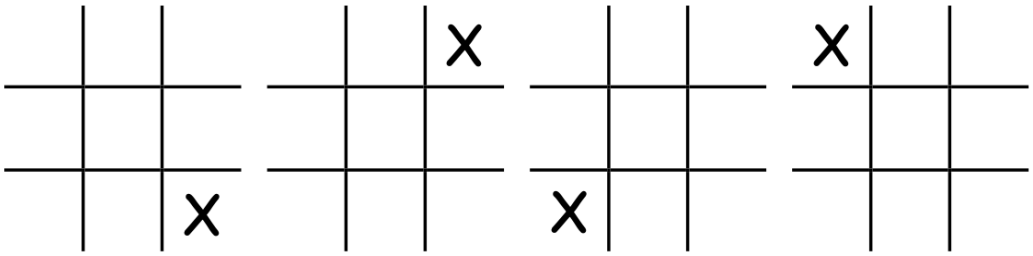
ب) تعداد دفعات به روز رسانی ارزش حالت‌ها در کدام روش بیشتر است؟

پ) فرض کنید نتایج حاصل از ۳ اپیزود در یک مساله اپیزودیک به صورت زیر بوده است و دنباله حالات و پاداش‌های زیر تاکنون بدست آمده: (حروف نشان دهنده حالت‌ها هستند و پس از آنها پاداش بدست آمده به صورت یک عدد نوشته شده است).

Episode1: A, 0, B, 0, C, 0, D, 1, T
Episode2: B, 1, C, 1, T
Episode3: D, 0, T

ت) فرض کنید ارزش اولیه همه‌ی حالات 0 و $\gamma = 0.9, \alpha = 0.2$ باشند. ارزش حالات A, B, C, D را پس از این اپیزودها با دو روش TD(0) و Monte Carlo بدست آورید.

۳. در بازی tic tac toe حدود هشت هزار حالت مختلف ممکن است اتفاق بیافتد. از آنجایی که در این بازی، قرار گرفتن افقی، عمودی یا قطری مهره‌ها اهمیت دارد، برخی از حالت‌ها معادل با یکدیگر هستند و ارزش یکسانی دارند. به طور نمونه، هر ۴ حالت زیر دارای یک ارزش هستند و با یک احتمال مساوی منجر به پیروزی می‌شوند.



بنابراین، تعداد حالت‌های متمایز بسیار کمتر است. حال می‌توان روش را طوری تغییر داد که با کم شدن تعداد وضعیت‌ها، عامل سریع‌تر یاد بگیرد (نیاز نباشد هر کدام از ۴ حالت بالا بارها تجربه شوند و مستقل در نظر گرفته شوند). در پوشه تمرین دو فایل پایتون وجود دارد:

- tictactoe_self_play: دو عامل x و o همزمان در حال یادگیری هستند. این کد برای آموزش دو عامل است.
- tictactoe_human_ai: در این کد عامل آموزش دیده o با عامل انسانی x بازی می‌کند و برای آزمون عملکرد عامل آموزش دیده قابل استفاده است.

این دو فایل را مطالعه کنید و سپس فایل tictactoe_self_play را به صورت زیر تغییر دهید:

الف) در هر iteration یک پاداش دریافت می‌شود. با جمع‌آوری این پاداش‌ها نمودار پاداش را در انتهای آموزش رسم کنید.

ب) از ویژگی تقارن در tic tac toe استفاده کنید و تعداد حالت‌های موجود را کاهش دهید.

پ) نمودار پاداش را برای قسمت (**ب**) رسم کنید و با قسمت (**الف**) مقایسه کنید.

نکات تکمیلی

۱) لطفاً پاسخ سوالات (تئوری و توضیحات پیاده‌سازی) را به طور گویا و به زبان فارسی و در صورت امکان تایپ همراه با سورس کدهای نوشته شده، در یک فایل فشرده شده به شکلHW12_YourStudentID.zip قرار داده و بارگذاری نمایید.

۲) منابع استفاده شده را به طور دقیق ذکر کنید.

۳) برای سهولت در پیاده‌سازی‌ها و منابع بیشتر، زبان پایتون پیشنهاد می‌شود. لطفا کدهای مربوطه را به طور جداگانه در فرمت .ipynb ارسال نمایید.

۴) ارزیابی تمرین‌ها براساس صحیح بودن راه حل‌ها، گزارش مناسب، بهینه بودن کدها و کپی نبودن می‌باشد.

۵) در مجموع تمام تمرین‌ها، تنها ۷۲ ساعت تاخیر در ارسال پاسخ‌ها مجاز است اما پس از آن به صورت خطی از نمره شما کسر خواهد شد (معادل با روزی ۵۰ درصد).

۶) تمرین‌ها باید به صورت انفرادی انجام شوند و حل گروهی تمرین مجاز نیست.

۷) در رابطه با پرسش و پاسخ در رابطه با تمرین‌ها می‌توانید در گروه مربوطه مطرح کنید.

موفق باشید.