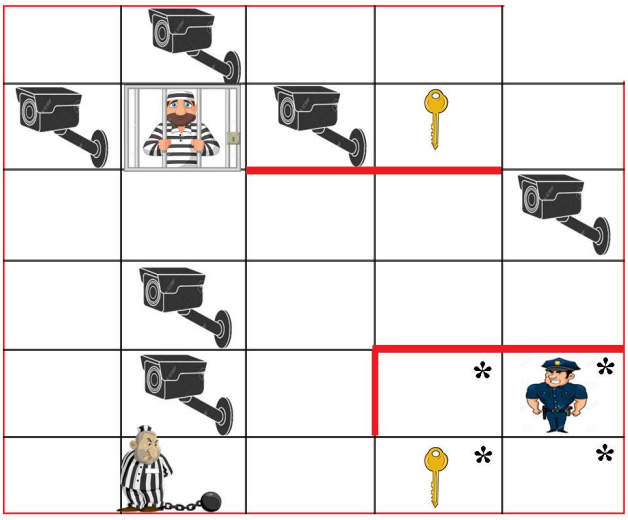
به نام او، برای او

تمرین چهارم درس یادگیری ماشین
پاییز ۱۳۹۷

سوال ۱. رهایی از انفرادی

چنگیز در زندان به رهایی‌دهنده از انفرادی معروف است! او به ازای هر زندانی که از انفرادی آزاد می‌کند مبلغ کلانی دریافت می‌کند. در این مسئله می‌خواهیم شریک جرم او شویم و مسیر بهینه‌ای که می‌تواند چنگیز را به زندانی محبوس در انفرادی برساند را یاد بگیریم!

نقشه‌ی زندان به صورت زیر است:



چنگیز در هر مرحله تنها می‌تواند یک خانه به بالا، پایین، چپ و یا راست حرکت کند. البته در صورتی‌که به دیوار نخورد! (خطوط قرمز دیوار هستند و عبور از آنها غیر ممکن است.) وزنه‌ی سنگینی که به پای چنگیز زنجیرشده نه تنها امکان بالا رفتن از دیوار را از او گرفته، که حتی عبور میان خانه‌های نقشه را هم برایش دردآور کرده‌است.

چنگیز ابتدا بایستی خود را به کلید برساند. همان‌طور که در نقشه می‌بینید در زندان دو کلید وجود دارد که هر دو درب سلول انفرادی را باز می‌کنند. چنگیز کافی است یکی از این کلیدها را بردارد و سپس سراغ زندانی محبوس در انفرادی برود.

در برخی از خانه‌های مسیر دوربین نصب شده است، درصورتی‌که چنگیز در این خانه‌ها قرار بگیرد، تصویرش ضبط شده و به جرم پرسه زدن در زندان شکنجه می شود.

همانگونه که در نقشه می بینید، زندانبانی در حال نگهبانی دادن است. درصورتی‌که چنگیز وارد خانه‌ای از جدول که زندانبان در آن قرار دارد شود، دستگیر می شود و تا آخر عمرش به زندانی دورافتاده تبعید می شود و دیگر فرصت این شغل شریف را نخواهد داشت! زندانبان در هر لحظه با احتمال مساوی در یکی از چهار خانه ایست که با علامت * نشان داده شده اند.

فضای حالت و تابع پاداش را تعریف کنید و سپس نحوه یادگیری مسیر بهینه برای چنگیز را با روش n-step TD(lambda) برای لامبدای صفر و همینطور یک لامبدای غیرصفر که جواب مناسب می دهد پیاده‌سازی کنید. Performance و سرعت یادگیری چنگیز به ازای دو حالت مختلف را بررسی کنید.

نقطه شروع حرکت همواره همان خانه‌ایست که چنگیز در نقشه نشان داده شده است. مسیر بهینه مسیری است که کمترین جابجایی و همچنین کمترین احتمال و میزان جریمه و شکنجه شدن را برای چنگیز به همراه داشته باشد.

سوال ۲. محاسبه‌ی خطای تخمین

یک عامل یادگیر در یک مسأله MDP از $n-step-return$ برای یادگیری استفاده می‌کند. این عامل یادگیری همواره از حالت S_0 آغاز می کند. حداکثر خطای تخمین یک سیاست مشخص $\left(V^{\pi}(S)\right)$ برای این عامل به ازای یک مقدار مشخص n چقدر است؟

سوال ۳. پاداش آخر کار

در یک محیط گسسته عامل یادگیر می‌داند که متوسط پاداش آنی رفتن به حالت نهایی، غیرصفر و در دیگر موارد صفر است. $\left(R^{a}_{ss=\tau} \neq 0 \text{ و } R^{a}_{ss\neq\tau} = 0\right)$ عامل متوسط پاداش دریافتی در حالت‌های نهایی را نمی‌داند. در صورتی که عامل $P^{a}_{ss'}$ را بداند، روشی برای یادگیری تعاملی سیاست بهینه ارائه دهید. نقطه شروع هر اپیزود یادگیری تصادفی تعیین می‌شود.

لطفا به نکات زیر توجه کنید:

- ✓ حجم گزارش شما به هیچ‌وجه معیار نمره‌دهی نیست، پس لطفا در حد نیاز توضیح دهید.
- ✓ تایپ کردن تمرین‌ها اجباری نیست ولی در صورتی که روی کاغذ می‌نویسید علاوه بر آپلود اسکن در صفحه‌ی درس، برگه‌ی خود را در اولین کلاس درس پس از ددلاین به استاد تحویل دهید.
- ✓ سعی کنید از پاسخ‌های روشن در گزارش خود استفاده کنید و اگر پیش‌فرضی در حل سوال در ذهن خود دارید، حتما در گزارش خود آن را ذکر کنید.
- ✓ از نمودارهای واضح در گزارش خود استفاده کنید، نمودارهایی که دارای لیبل‌گذاری روشن روی هر محور و همینطور توضیح مناسب باشد.
- ✓ کدهایی که به همراه گزارش تحویل می‌دهید باید قابل اجرا باشد. همچنین توجه کنید که به تمرین بدون گزارش نمره‌ای تعلق نمی‌گیرد.
- ✓ لطفا در گزارش و کدهای خود از تمرین دیگران استفاده نکنید، مشورت و همفکری در مورد سوال‌ها اشکالی ندارد اما اگر شباهت بیش از اندازه در تمرین‌ها دیده شود منجر به صفر شدن نمره خواهد شد.
- ✓ تمام فایل‌ها را در قالب یک فایل zip یا rar در سایت درس بارگذاری کنید.
- ✓ برای پیاده‌سازی تمرین فقط از زبان‌های MATLAB و یا Python می‌توانید استفاده کنید.

موفق و سلامت باشید. :)