دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی

پاییز ۱۴۰۰

استاد: محمدحسین رهبان
گردآورندگان: محمد محدی، حمیدرضا کامکاری
بررسی و بازبینی: محمدرضا یزدانی‌فر

تمرین هفتم سری دوم — Reinforcement Learning — مهلت ارسال: ۱۶ دی

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین تا سقف سه روز و در مجموع ۲۵ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۰ درصد از نمره تمرین به صورت ساعتی کسر 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.

سوالات (۱۰۰ نمره)

۱. (۱۰۰ نمره) روش policy iteration یک روش برای یافتن سیاست بهینه در یک MDP است. این روش به صورت کلی، از دو مرحله‌ی زیر تشکیل شده است:

(آ) سنجش (evaluation) یک سیاست مانند π که از پیش بدست آمده است. برای مثال، می‌توان مقادیر V-value یا Q-value را پیدا کرد.

(ب) پس از سنجش سیاست پیشین، آن را بهبود دهیم و یک سیاست بهتر بدست آوریم.

لازم به ذکر است که دو مرحله‌ی بالا، ممکن است چندین بار تکرار شوند تا در نهایت، دقت و عملکرد مناسبی بدست بیاید. با توجه به این روش، به سوالات زیر پاسخ دهید.

(آ) فرض کنید که در یک محیط ناشناخته، می‌خواهیم یک سیاست بهینه بیابیم. منظور از محیط ناشناخته این است که مقادیر R و P شناخته شده نیستند. برای بدست آوردن سیاست، قصد داریم از روشی مشابه policy iteration استفاده کنیم. بنابراین، در هر گام، ابتدا سیاست موجود را می‌سنجیم و سپس آن را بهبود می‌دهیم. در این روش، بهتر است که در سنجش از $V-value$ استفاده کنیم یا $Q-value$؟ دلیل خود را توضیح دهید.

(ب) زمانی که MDP ناشناخته باشد، از روشی به اسم ϵ-greedy استفاده می‌شود. توضیح دهید چرا نیاز است که از این روش استفاده کنیم و اگر از روش عادی مانند policy iteration استفاده کنیم، چه مشکلاتی ممکن است پیش بیاید.

(ج) فرض کنید که سیاست از پیش بدست آمده π باشد و در صورتی که از روش ϵ-greedy استفاده کنیم، سیاست به π' تبدیل شود. اگر بدانیم

$$\forall s : \mathbb{E}_{a\sim\pi'}[Q^{\pi}(s,a)] \geq \mathbb{E}_{a\sim\pi}[Q^{\pi}(s,a)]$$

ثابت کنید

$$\forall s : V^{\pi'}(s) \geq V^{\pi}(s)$$