به نام خدا

دانشکده مهندسی برق و کامپیوتر دانشگاه تهران

امتحان پایان‌ترم درس یادگیری تعاملی

۴ بهمن ۱۴۰۱

پیش‌نوشت ۱: استفاده از کتاب و جزوه مجاز است.

پیش‌نوشت ۲: لطفا توجه فرمایید که علاوه بر حل مسئله ، روش استدلال و فکر کردن شما بر روی مسئله و نزدیک شدن به راه حل نیز اهمیت زیادی برای ما دارد. در صورت نیاز به فرض اضافه برای حل مسئله، به صورت مستدل آن فرضیات را مشخص و استفاده نمایید.

پیش‌نوشت ۳: لطفا پاسخ را خوانا و دقیق، و با نگاه ریاضی و محاسباتی، و در حد امکان کوتاه بنویسید و از توان داستانسرایی خود در این امتحان استفاده نکنید 😊.

سوال ۱-) در یک مسئله 100-armed-bandit به عامل یادگیر سه گزاره ناقص (همه بازوهای زرد رنگ.....؛ اکثر بازوهای با شماره مضرب ۴ .....؛ نیمی از بازوهای مشکی رنگ....) در خصوص ویژگی‌های متوسط پاداش بازوها داده شده است. بازوها رنگی است و شماره نیز دارد. "....." نشانگر نقص در گزاره است که عامل از آن مطلع نیست و عامل فقط معنی قسمت قابل خواندن گزاره‌ها را درک می‌کند. روشی برای استفاده از این گزاره‌های ناقص (شامل مدل یادگیری مناسب و روش بکارگیری این سه گزاره ناقص) برای کاهش حسرت عامل ارایه دهید. علاوه برتوضیح روش و دلایل آن، لازم است پشتوانه ریاضی روش پیشنهادی ارایه و شبه کد روش نیز ارایه گردد. (۲۵ نمره)

سوال۲-) در یک MDP هدف آن است که سیاستی که expected discounted utility یک عامل انسانی را حداکثر میکند محاسبه شود. مدل utility و ضرایب مربوط به subjective probability این عامل بر اساس مدل Prospect Theory داده شده است. (۲۵ نمره)

الف-) معادله بلمن مربوطه را استخراج نمایید.

ب-) تعداد حالت و عمل این مسئله بسیار زیاد است و لذا به جای حل مستقیم معادله بلمن، لازم است که یک شبیه‌ساز از مسئله ایجاد و یک عامل RL با استفاده از روش SARSA سیاست بهینه را در تعامل با شبیه‌ساز بیابد. تغییرات لازم در مدل یادگیری و یا در پیاده‌سازی شبیه ساز را برای یافتن سیاست بهینه به همراه ریاضیات مربوطه ارایه دهید.

سوال ۳-) در روش Gradient Bandit و در روش Policy Gradient در محیط پیوسته، استفاده از متوسط پاداش به عنوان خط-مبنا، عموما میزان حسرت را کاهش میدهد. (۲۰ نمره)

الف-) دلیل این امر را به صورت ریاضی در یکی از دو روش ارایه دهید.

ب-) گفته شده است که به جای استفاده از خط-مبنا در Policy Gradient متوسط گرادیان بر روی چند نمونه جهت یادگیری استفاده شود. به نظر شما آیا این راه حال جایگزین مناسبی برای استفاده از خط-مبنا است؟

سوال ۴-) Reward Shaping در یک MDP(S,A,R,P) به دسته روشی گفته می‌شود که در آن علاوه بر پاداش R، یک پاداش R' نیز در تعداد بسیار معدودی از حالت-عمل‌ها برای افزایش سرعت یادگیری به عامل داده شود. عامل تفاوتی بین R و R' در یادگیری قائل نمی‌شود. در یک MDP گسسته اپیزودیک با ضریب تخفیف کوچکتر از یک، و روش on-policy MC learning، R' چه شرایط ریاضی را باید ارضا کند تا اضافه شدن آن به مسئله باعث تغییر سیاست بهینه نشود و در ضمن حسرت را کاهش دهد؟ (۲۰ نمره)

سوال ۵-) در یک محیط حالت پیوسته و عمل گسسته، برای تنظیم Exploration-Exploitation Balance (EEB) روشهای مختلفی وجود دارد. یکی از روشها استفاده از چند شبکه عمیق به جای یک شبکه برای تخمین ارزش حالت-عمل در روش Q-learning است. شبه کد یادگیری برای ایجاد EEB در این روش را به صورت کامل ارایه دهید و استدلال نمایید که چرا این روش منجر به EEB مناسب می‌شود. در صورتی که این روش در مقایسه با روش Q-learning با استفاده از یک شبکه عمیق مزایا (و معایب) دیگری نیز دارد آنها را مستدل ارایه نمایید. (۲۰ نمره)

توجه: نمره از ۱۱۰ محاسبه خواهد شد.

موفق باشید.

مجید نیلی