به نام خدا

یادگیری تقویتی

نیم‌سال بهار ۴۰۲ـ۱۴۰۱

اساتید: دکتر رهبان، آقای حسنی

دانشکده‌ی مهندسی کامپیوتر

مدت: ۱۲۰ دقیقه | امتحان پایانترم | بارم: ۱۲۰ نمره

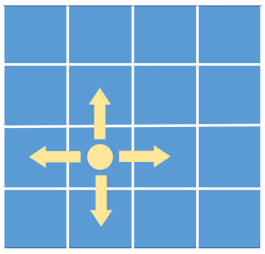
نام و نام خانوادگی : | شماره دانشجویی:

سوال ۱: Bayesian RL (۲۵ نمره)

(آ) تفاوت‌های belief MDP با MDP چیست؟ استدلال کنید چرا در belief MDP دیگر مدل ناشناخته نیست و می‌توان بر مبنای آن به برنامه‌ریزی پرداخت؟

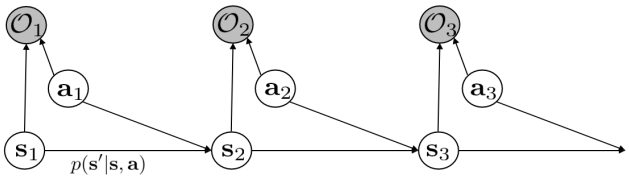
(ب) مثال gridworld ارائه شده در درس را در نظر بگیرید که در آن یک ربات انسان‌نما به حرکت در محیط برای رسیدن به مقصدی مشخص می‌پردازد. برای این ربات انتخاب کنش در چهار جهت اصلی شمال، جنوب، شرق و غرب در نظر گرفته شده است. برای حرکت به سمت هر کدام از این جهت‌ها، ربات از ترکیب دو حرکت چرخش درجا و سپس یک گام رو به جلو استفاده می‌کند. متسفانه به دلیل استهلاک مفاصل یکی از پاهای ربات، به هنگام حرکت رو به جلو، شاهد مقداری شیفت به سمت طرفین (در جهت عمود بر راستای حرکت) هستیم. برای سادگی این شیفت را به صورت نویز یک‌بعدی جمع‌شونده گاوسی با میانگین نامشخص و انحراف معیار ثابت $\mathcal{N}(\mu, 1)$، در نظر می‌گیریم. هر چند تصحیح این عیب بدون تعمیر مکانیکی قطعات گران‌قیمت ممکن نیست، ولی مهندسان خلاق دانشگاه شریف به رفع این مشکل به صورت نرم‌افزاری روی آورده‌اند. بر روی ربات یک سنسور نصب کرده‌اند که میزان شیفت (x) بعد از هر گام را اندازه می‌گیرید. حال برای متغیر مجهول μ باور با توزیع پیشین $\mathcal{N}(\mu_0, \sigma_0)$ در نظر می‌گیرند که امید دارند با استفاده از بروزرسانی بیزی و با ثبت مشاهدات سنسور، به مرور باور حول متغیر اصلی μ نوک‌تیز (پیکی) شود.

فرض کنید این سنسور در اولین ثبت خود به میزان x_1 واحد شیفت ثبت کرده است. با استفاده از این اطلاعات، باور (توزیع پیشین) را بروزرسانی کرده و متغیرهای توزیع پسین $\mathcal{N}(\mu_1, \sigma_1)$ را محاسبه کنید. با در نظر گرفتن این باور جدید، توزیع پسین predictive $p(x|x_1, \mu_1, \sigma_1)$ را محاسبه کنید.



سوال ۲: RL with Soft Optimality (۲۲ نمره)

در این سوال، بجای MDP استاندارد، مدل گرافی زیر را در نظر می‌گیریم که در آن منظور از $\mathcal{O}_t$ متغیر بهینگی در لحظه t است که به صورت یک متغیر تصادفی برنولی با احتمال یک بودن به صورت $p(\mathcal{O}|s_t, a_t) = exp(r(s_t, a_t))$ تعریف می‌شود.



(آ) با در نظر گرفتن دنباله $\tau = \{s_{1:T}, a_{1:T}\}$ توزیع $p(\tau|\mathcal{O}_{1:T})$ را محاسبه کنید و تفاوت آن با $p(\tau)$ در یک MDP معمولی را اشاره کرده و توضیح دهید چه ویژگی مثبتی به آن اضافه شده است.

(ب) حال به محاسبه احتمال backward $\beta_t(s_t, a_t) = p(\mathcal{O}_{t:T}|s_t, a_t)$ پرداخته و یک رابطه‌ی بازگشتی برای محاسبه آن از لحظه T تا t ارائه دهید. برای نوشتن روابط از متغیر کمکی $\beta_t(s_t) = \mathbb{E}_{a_t \sim p(a_t|s_t)}[\beta_t(s_t, a_t)]$ بهره بگیرید.

(ج) حال با استفاده از تعاریف $\beta_t(s_t, a_t)$ و $\beta_t(s_t)$، به محاسبه سیاست بر حسب آن‌هابپردازید.

$$\pi(a_t|s_t) = p(a_t|s_t, \mathcal{O}_{t:T})$$

(د) فرض کنید یک رابطه بازگشتی هم برای محاسبه احتمالات forward $\alpha_t(s_t) = p(s_t|\mathcal{O}_{1:t-1})$ در دست داریم. حال هدف ما محاسبه توزیع حاشیه‌ای $p(s_t|\mathcal{O}_{1:T})$ بر حسب احتمالات backward و forward است. این توزیع را محاسبه کرده و توضیح دهید چگونه soft optimality مشاهده شده در رفتار عامل‌های هوشمند زیستی (مانند انسان) را توجیه می‌کند.

سوال ۳: Exploration in RL (۲۰ نمره)

(آ) توضیح دهید چرا الگوریتم Bootstrapped DQN در اکتشاف در وظیفه‌های پیچیده بهتر از $\epsilon - \text{greedy}$ عمل می‌کند؟

(ب) برای هر نمونه انتخاب شده در این الگوریتم از یک ماسک استفاده می‌شود. اولا چیستی این ماسک را توضیح دهید و ثانیا بگویید چرا از آن استفاده می‌شود.

(ج) به صورت دقیق و قدم به قدم توضیح دهید پاداش کمکی در روش VIME به چه صورت محاسبه می‌شود؟

سوال ۴: Inverse RL (۱۳ نمره)

در محاسبه گرادیان تابع هزینه در الگوریتم Guided Cost Learning، از متوسط گرادیان تابع پاداش روی نمونه‌های خبره منهای همین کمیت که روی نمونه‌های سیاست فعلی محاسبه می‌شود. با این حال در جمله دوم (میانگین گرادیان روی نمونه‌های گرفته شده از سیاست فعلی) به هر عبارت در میانگین وزن داده می‌شود. این وزن دقیقا چیست (فرآیند بدست آمدن آن را بنویسید) و چرا استفاده می‌شود؟

$$\nabla_\psi \mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \nabla_\psi r_\psi(\tau_i) - \frac{1}{\sum_j \omega_j}\sum_{j=1}^{M} \omega_j \nabla_\psi r_\psi(\tau_j)$$

سوال ۵: Offline RL (۲۰ نمره)

(آ) با در نظر گرفتن رابطه آپدیت بلمن، به صورت مستدل توضیح دهید چرا در چارچوب برون‌خطی (offline)، تخمین ارزش دچار بیش‌برآورد می‌شود.

$$Q(s,a) \leftarrow r(s,a) + \mathbb{E}_{a' \sim \pi_{new}[Q(s',a')]}$$

(ب) یکی از راهکارهای رفع مشکل بیش‌برآورد در چارچوب برون‌خطی، یادگیری ارزش به صورت محتاطانه است که تابع هدف آن برای محاسبه ارزش در ادامه داده شده است. نقش هر کدام از چهار جزء موجود در این تابع هدف را بیان کنید.

$$\hat{Q}^\pi = arg\min_Q \max_\mu \alpha \mathbb{E}_{s\sim D, a\sim\mu(a|s)}[Q(s,a)] - \alpha \mathbb{E}_{(s,a)\sim D}[Q(s,a)] - \mathbb{E}_{s\sim D}[\mathcal{H}(\mu(\cdot|s))] + \mathbb{E}_{(s,a,s')\sim D}\left[(Q(s,a) - (r(s,a) + \mathbb{E}[Q(s',a')]))^2\right]$$

(ج) اگر بخواهیم از این روش محتاطانه در چارچوب یادگیری تقویتی مبتنی بر مدل استفاده کنیم، چه تغییراتی در تابع هدف لازم است؟ تابع هدف جدید را بازنویسی کنید.

(د) فرض کنید پس از یادگیری برون‌خطی، به عامل امکان تعامل محدود با محیط داده شده باشد. آیا کماکان استفاده از این یادگیری محتاطانه را پیشنهاد می‌کنید؟ چرا؟

سوال ۶: Unsupervised RL (۲۰ نمره)

در روش یادگیری مهارت DADS، برای تخمین $p(s'|s,a)$ از یک مدل پارامتری استفاده می‌شود. جهت یادگیری پارامترهای این مدل، از vaira-tional lower bound روی $I(s';z|s)$ استفاده می‌شود. یک راهکار معقول در مراحل بهینه‌سازی این مدل پارامتری، تنگ (tight) کردن این کران است.

(آ) با نوشتن و ساده‌سازی کران پایین نشان دهید از چه گرادیانی می‌توان برای به روز رسانی پارامترهای مدل دینامیک استفاده کرد.

(ب) در صورتی که بخواهیم اطلاعات متقابل مشروط بین s' و z را ماکزیمم کنیم، از چه الگوریتمی باید استفاده کنیم. جزئیات این الگوریتم را شرح دهید.