دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی

پاییز ۱۴۰۰
استاد: محمدحسین رهبان
گردآورندگان: محمد محمدی، آتوسا چگینی، حمیدرضا کامکاری
بررسی و بازبینی: محمدرضا یزدانی‌فر

مینی‌پروژه پنجم | MDP, RL | مهلت ارسال: ۱۷ بهمن

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- امکان ارسال با تاخیر پاسخ این مینی‌پروژه تا سقف ۱ 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 نظری (۳۰ نمره)

۱. (۱۵ نمره) در این مثال می‌خواهیم از یک ضریب تخفیف (discount factor) با مقدار کمتر از یک توجه کنیم ($\gamma < 1$). می‌دانیم که عملگر بلمن (Bellman operator) یک عملگر دارای انقباض می‌باشد. بنابراین اگر عملگر بلمن را به صورت زیر تعریف کنیم:

$$B_k v(s) = \max_a \left[R(s,a) + \gamma_k \sum_{s' \in S} p(s'|s,a) v(s') \right]$$

اگر $\gamma_k = \gamma$ مقداری ثابت باشد در نتیجه $B_k = B$ و داریم:

$$\forall v_1, v_2 : ||Bv_1 - Bv_2||_\infty \leq \gamma ||v_1 - v_2||_\infty$$

این در واقع تئوری پشت این موضوع است که الگوریتم value iteration پایان پذیر است چرا که

$$||B_K ... B_1 v_1 - B_K ... B_1 v_2||_\infty \leq \gamma_1 \gamma_2 ... \gamma_k ||v_1 - v_2||_\infty = \gamma^K ||v_1 - v_2||_\infty$$

و به ازای $\gamma_k = \gamma < 1$ به صفر میل می‌کند.
حال فرض کنید γ_k را ثابت در نظر نمی‌گیریم. بلکه به صورت متغیر و به صورت زیر در نظر می‌گیریم:

$$\gamma_k = 1 - \frac{1}{k+1}$$

بنابراین اگر قرار باشد K بار پیمایش انجام بدهیم داریم:

$$v_{k+1} = B_k v_k$$

و مقدار k از K شروع شده و به ۱ می‌رسد. به صورت شهودی، مقدار ضریب تخفیف در ابتدا کم است و هر چه زمان جلو تر می‌رود به ۱ نزدیک‌تر می‌شود. حالا بر این اساس به سؤالات زیر پاسخ بدهید.

الف) ثابت کنید با ضریب تخفیف متغیر نیز، هر B_k یک عملگر انقباضی است یا به عبارت دیگر:

$$\forall v_1, v_2 : ||B_k v_1 - B_k v_2||_\infty \leq \gamma_k ||v_1 - v_2||_\infty$$

(راهنمایی: برای اثبات مشابه حالت ثابت عمل کنید و از این لینک http://people.eecs.berkeley.edu/ pabbeel/cs287-fa09/lecture-notes/lecture5-2pp.pdf استفاده کنید.)

ب) ثابت کنید

$$\prod_{k=1}^{K} \gamma_k \leq \frac{1}{K+1}$$

ج) ثابت کنید الگوریتم value iteration با استفاده از این ضریب تخفیف نیز به مقداری خاص همگرا می‌شود.

د) متد جدید پیشنهاد شده را از نظر سرعت همگرایی با متد قبلی مقایسه کنید. آیا این ضریب تخفیف بهتر است یا حالت ثابت؟ توضیح دهید.

۲. (۱۵ نمره) تصور کنید گراف G وزن‌دار و جهت‌دار داریم و آن‌را با ماتریس W مدل کرده‌ایم که در آن

$$W_{i,j} = \begin{cases} weight(e_{i\to j}) & \text{if } (i \to j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

فرض کنید در هر رأسی که هستیم می‌توانیم به یکی از رئوس مجاور برویم. اگر در رأسی مثل v باشیم و قصد کنیم به رأس مجاوری برویم به احتمال $\eta \leq 1$ به همان رأس هدف می‌رویم و به احتمال $1-\eta$ ممکن است به هر یک از رئوس مجار به احتمال مساوی برویم. به صورت دقیق‌تر، اگر در رأس v باشیم و عملیات u را انتخاب کنیم داریم:

$$P(w|v, action = u) = \begin{cases} \eta + \frac{1-\eta}{d_v} & \text{if } w = u \\ \frac{1-\eta}{d_v} & \text{otherwise} \end{cases}$$

که در آن d_v درجه خروجی رأس v می‌باشد. فرض کنید از روی هر یالی که می‌گذریم به اندازه وزن آن یال جایزه (reward) می‌گیریم. حال فرض کنید policy به صورت تصادفی و پیوسته است و آن را با ماتریس Π نشان می‌دهیم که به صورت زیر است:

$$\Pi_{i,j} = \begin{cases} \text{the probability of taking action } j \text{ while in } i & \text{if } (i \to j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

با توجه به این توضیحات

الف) فرض کنید $v_\Pi(s)$ تابع value به ازای هر رأس باشد اگر از تابع policy متناظر با Π استفاده کنیم و $v^*(s)$ تابع value باشد اگر از بهترین استراتژی استفاده کنیم. . با توجه به توضیحات سؤال رابطه بلمن مربوط به v_Π و v^* را بنویسید.

ب) اگر فرض کنیم $\eta = 1$ یعنی تصمیماتمان نتایج قطعی‌ای دارند، در اینصورت ثابت کنید می‌توانیم مقدار value ها را مستقیما با حل معادله زیر به‌دست بیاوریم:

$$[I - \gamma\Pi]\, v_\Pi = [\Pi \odot W]\vec{1}$$

که در آن $\vec{1}$ برداری با اندازه تعداد رئوس گراف است که تمام اعضای آن برابر یک هستند و $\odot$ عملگر ضرب هادامارد دو ماتریس است. به عبارت دقیقتر:

$$[A \odot B]_{i,j} = A_{i,j} B_{i,j}$$

ج) هدف ما در این سؤال پیدا کردن ماتریس Π ای است که مقدار v_Π ها در آن بیشینه شود. حال نشان دهید پیدا کردن بهترین Policy معادل حل کردن مسئله بهینه‌سازی درجه دو زیر است:

$$\begin{aligned}
&\max_{\Pi} || \left[I - \gamma\Pi\right]^{-1} \left[\Pi \odot W\right]\vec{1}||_2^2 \\
&\forall (i \to j) \notin E(G) : \Pi_{i,j} = 0 \\
&\forall i, j \in V(G) : \Pi_{i,j} \geq 0,\ \vec{1} = \Pi\vec{1} \\
&\left[I - \gamma\Pi\right]^{-1} \left[\Pi \odot W\right]\vec{1} \succeq 0
\end{aligned}$$

سوالات عملی (۹۰ + ۳۰ نمره)

برای سوالات عملی به فایل jupyter notebook داخل آرشیو مراجعه کنید.