دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی
پاییز ۱۴۰۰
استاد: محمدحسین رهبان
گردآورندگان: محمد محدی، حمیدرضا کامکاری
بررسی و بازبینی: محمدرضا یزدانی‌فر

پاسخ تمرین هفتم سری اول — Markov Decision Processes — مهلت ارسال: ۹ دی

سوالات (۱۰۰ نمره)

۱. (۱۰۰ نمره)

با توجه به تقارن خیلی از حالت‌ها می‌توانیم تنها سه حالت متناظر با تمام حالت‌ها مدل کنیم. حالت‌های واقعی در یکی از این سه‌دسته با rotation یا قرینه‌کردن قرار می‌گیرد.

در شکل

می‌توانید یکسری یال‌های تعمیم یافته (Hyperedge) ببینید که هر کدام بیانگر یکسری action است. برای سادگی بسیاری از action ها که نتایج مشابهی دارند به صورت یکجا نوشته شده‌اند. حالت بالا_چپ را cross-state و حالت بالا راست را adjacent-state می‌نامیم و حالت پایین را dead می‌نامیم.

- اگر در حالت cross-state باشیم و به سمت دیوار حرکت کنیم، به احتمال یک به حالت adjacent می‌رسیم چرا که خودمان چه کامپیوتر به مشکل بخورد و چه نخورد سر جایمان می‌مانیم ولی روح به هر طریقی یک واحد نزدیک می‌شود.
- اگر در حالت cross-state باشیم و یکی از دو جهت سمت روح را بزنیم به احتمال 0.1 سر جای خود می‌مانیم و به حالت adjacent-state می‌رویم. به احتمال 0.9 نیز در یکی از دو جهت حرکت می‌کنیم که دو احتمال وجود دارد، یا روح به احتمال $\frac{1}{2}$ به همان خانه packman می‌رود که به حالت dead می‌رسیم و یا اینکه به احتمال $\frac{1}{2}$ به خانه مخالف می‌رویم که عملا یک rotation از همین حالت cross-state است.
- اگر در حالت adjacent-state به سمت دیوار یا به سمت روح حرکت کنیم در هر صورت روح packman را می‌خورد.
- اگر در حالت adjacent-state باشیم و به سمت فرار از روح حرکت کنیم، به احتمال 0.1 سر جای خود می‌مانیم و خورده می‌شویم و در غیر اینصورت روح به جای قبلی ما می‌آید و ما به جهت مخالف می‌رویم که با یک دوران می‌توان دید همان adjacent-state است.

اگر به هر حالتی به‌جز حالت dead برویم ۱ امتیاز دریافت می‌کنیم.

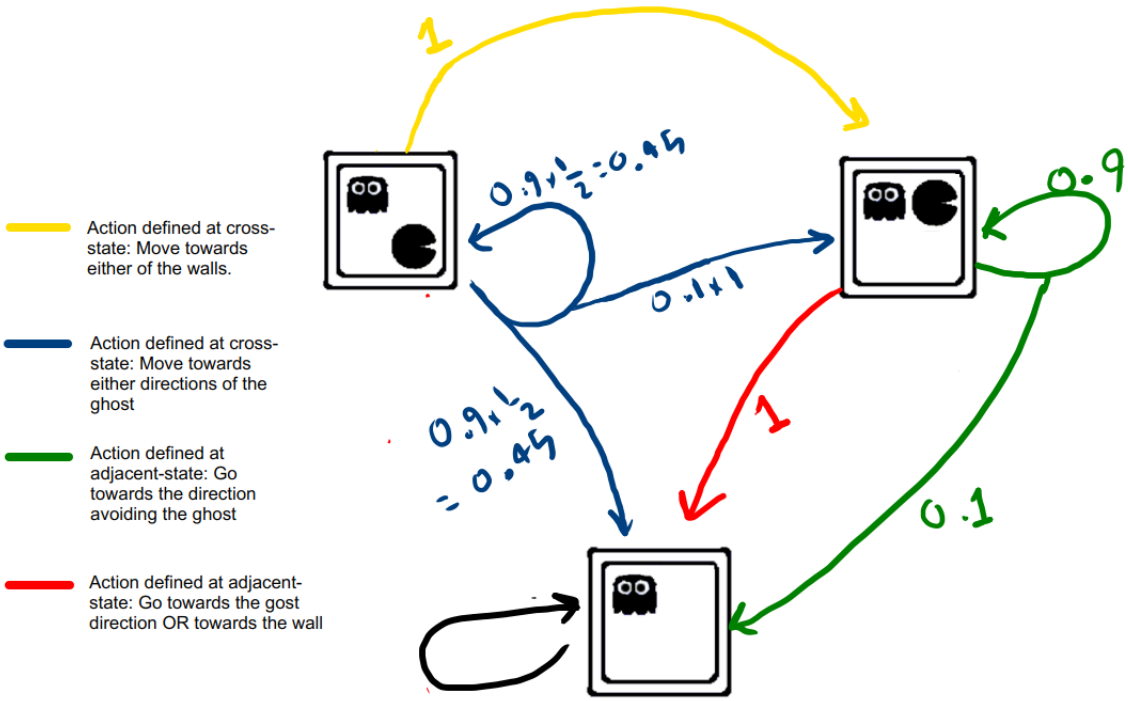
برای دو بخش بعدی سؤال به اینصورت عمل می‌کنیم:

ابتدا فرض کنید که یک سیاست مانند π داشته باشیم. اگر یک state مانند s و یک action مانند a وجود داشته باشد که

$$q^{\pi}(s,a) > v^{\pi}(s)$$

می‌توان این سیاست را بهبود داد. برای این کار، می‌توان سیاست π' را در نظر گرفت که همان π است با این تفاوت که در s به صورت قطعی a را انجام دهیم. در این صورت داریم:

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



شکل ۱ : شکل MDP.

اثبات:

$$\begin{aligned}
v^{\pi}(s_1) &= \mathbb{E}_{a_1\sim\pi}[q^{\pi}(s_1,a_1)] \\
&\leq \mathbb{E}_{a_1\sim\pi'}[q^{\pi}(s_1,a_1)] \\
&= \mathbb{E}_{a_1\sim\pi',a_2\sim\pi}[R(s_1,a_1)+\gamma q^{\pi}(s_2,a_2)] \\
&\leq \mathbb{E}_{a_1,a_2\sim\pi'}[R(s_1,a_1)+\gamma q^{\pi}(s_2,a_2)] \\
&= \mathbb{E}_{a_1,a_2\sim\pi',a_3\sim\pi}[R(s_1,a_1)+\gamma R(s_2,a_2)+\gamma q^{\pi}(s_3,a_3)] \\
&\leq \mathbb{E}_{a_1,a_2,\ldots\sim\pi'}[R(s_1,a_1)+\gamma R(s_2,a_2)+\gamma^2 R(s_3,a_3)+\ldots] = v^{\pi'}(s_1)
\end{aligned}$$

پس می‌دانیم که تا زمانی که s و a پیدا بشوند، می‌توانیم سیاست را بهتر کنیم(دقت کنید که در استیت s سیاست جدید حتما مقدار بیشتری می‌گیرد). پس در نهایت به سیاستی می‌رسیم که به شکل زیر است:

$$v^{\pi}(s) = max_a q^{\pi}(s,a) = max_a R(s,a) + \gamma sum_{s'} P(s,a,s')v^{\pi}(s')$$

حال ادعا می‌کنیم که اگر دو سیاست بودند که ویژگی بالا را داشته باشند، مقادیر بدست آمده توسط آن‌ها یکسان است. به بیان دیگر:

$$\forall s : v^{\pi}(s) = v^{\pi'}(s)$$

اثبات: ابتدا مقادیر v را به عنوان یک بردار در نظر می‌گیریم که درایه‌ی i ام آن، $v(s_i)$ است و بردار را V می‌نامیم. سپس یک operator مانند T تعریف می‌کنیم که به شکل زیر است:

$$TV^{\pi}(s) = max_a \mathbb{E}[R(s,a) + \gamma v^{\pi}(s')]$$

حال ادعا می‌کنیم که

$$||TV^{\pi} - TV^{\pi'}||_{\infty} = \gamma ||V^{\pi} - V^{\pi'}||_{\infty}$$

اثبات:

$$\begin{aligned}||TV^{\pi} - TV^{\pi'}||_{\infty} &= max_s|TV^{\pi}(s) - TV^{\pi'}(s)| \\ &= max_s|max_a|R(s,a) + v^{\pi}(s)| - max_a|R(s,a) + v^{\pi'}(s)|| \\ &\leq max_s|max_a|R(s,a) + v^{\pi}(s) - R(s,a) - v^{\pi'}(s)|| \\ &= max_s max_a|\gamma v^{\pi}(s) - \gamma v^{\pi'}(s)|| \\ &= \gamma max_s|v^{\pi}(s) - v^{\pi'}(s)| = \gamma||V^{\pi} - V^{\pi'}||_{\infty}\end{aligned}$$

حال می‌دانیم که اگر دو سیاست دیگر قابل بهبود نباشند، داریم:

$$||V^{\pi} - V^{\pi'}||_{\infty} = ||TV^{\pi} - TV^{\pi'}||_{\infty} = \gamma||V^{\pi} - V^{\pi'}||_{\infty}$$

که چون γ عددی مثبت و کمتر از یک است، داریم:

$$||V^{\pi} - V^{\pi'}||_{\infty} = 0$$

حال می‌توان گفت هر سیاست غیر قابل بهبودی، از همه‌ی سیاست‌ها بزرگتر مساوی است. زیرا اگر یک سیاست کوچکتر نباشد، آن را غیر قابل بهبود می‌کنیم و در آخر برابر با سیاست غیر قابل بهبود اولیه می‌شود. پس در ابتدای کار، حتما باید کوچکتر یا مساوی باشد.

ضمنا برای این که ثابت کنیم یک سیاست بهینه قطعی وجود دارد، کافی است در state هایی که بین چند action یکی را انتخاب می‌کنیم، تنها آن action ای را انجام دهیم که $Q(s,a)$ آن بیشینه است. طبق لمی که در ابتدا ثابت کردیم، سیاست بدست آمده بهتر یا مساوی سیاست قبلی است. بنابراین می‌توان در هر state به صورت قطعی یک action را انجام داد.