هوش مصنوعی

نیم‌سال اول ۰۱-۰۰

مدرس: دکتر محمدحسین رهبان

دانشکده مهندسی کامپیوتر

تمرین هفتم ـ بخش دوم

محمدجواد هزاره

شماره دانشجویی: ۹۸۱۰۱۰۷۴

۱ سوال ۱

آ) استفاده از Q-Valueها مناسب‌تر خواهد بود. با دانستن Q-Valueها بهبود دادن سیاست داده شده بدون نیاز به دانستن T و R شدنی خواهد بود. چرا که اگر π سیاست داده شده باشد، داریم:

$$\hat{\pi}(s) = \arg\max_a Q^{\pi}(s, a)$$

بنابراین برای سنجش یک سیاست بهتر است از Q-Value استفاده کنیم.

ب) استفاده از policy iteration باعث می‌شود نتوانیم کل فضا را explore کنیم. در این روش با استفاده از یک policy اولیه شروع به سنجش آن و آپدیت آن می‌کنیم و این باعث می‌شود اگر policy اولیه به یک سو گرایش داشته باشد، الگوریتم بهترین سیاستی را پیدا کند که به بهترین ارزش‌های آن سو بگراید. در حالی که ممکن است در گوشه‌ی دیگری از فضای حالت، حالت‌های با پاداش بهتر وجود داشته باشد که با استفاده از این روش عامل هرگز به آن حالت‌ها نخواهد رفت. به عبارتی دیگر عامل گمان می‌کند که به بهترین سیاست با دریافت بهترین پاداش رسیده در حالی که اصلا تمام حالات مسئله را بررسی نکرده است. برای رفع این مشکل، از روش ϵ-greedy استفاده می‌شود که با انجام حرکت‌های تصادفی با احتمال ϵ، به عامل این امکان را می‌دهد که در بعضی از مواقع به حالت‌های جدید رفته و بتواند تمام فضای حالت مسئله را کشف کند و به بهترین پاداش ممکن دست یابد.

ج) می‌خواهیم اثبات کنیم که اگر از s شروع کرده و سیاست π' را دنبال کنیم، امیدریاضی مطلوبیت یا همان $V^{\pi'}(s)$، بیش‌تر از حالتی خواهد شد که سیاست π را دنبال کنیم. برای اثبات، از استقرا استفاده می‌کنیم. تعریف می‌کنیم:

$$V_n(s) := \begin{pmatrix} \text{امیدریاضی مطلوبیت، اگر از } s \text{ شروع کرده و } n \text{ قدم اول را از سیاست } \pi' \\ \text{و در ادامه از سیاست } \pi \text{ پیروی کنیم.} \end{pmatrix}$$

پایه: پایه‌ی استقرا برای V_1 همان فرضی است که در سوال داده شده. می‌دانیم:

$$V_1 = \mathbb{E}_{a\sim\pi'}[Q^{\pi}(s, a)] \geq \mathbb{E}_{a\sim\pi}[Q^{\pi}(s, a)] = V^{\pi}(s)$$

گام استقرا: فرض کنید بدانیم V_n خواسته‌ی مسئله را برآورده می‌کند یا به عبارتی $V_n(s) \geq V^\pi(s)$. اثبات می‌کنیم $V_{n+1}(s)$ نیز حداقل به اندازه‌ی $V^\pi(s)$ است. برای این منظور، اگر از s شروع کرده و با دنبال کردن π' برای n قدم، به ترتیب پاداش‌های R_1 تا R_n را دریافت کرده و حالت‌های s_1 تا s_n را دیده باشیم، (مشخصا R_iها و s_iها متغیرهایی تصادفی هستند که از توزیع T و R و فرض این‌که از سیاست π' استفاده کرده‌ایم پیروی می‌کنند) آن‌گاه برای V_n داریم:

$$V_n(s) = \mathbb{E}_{\pi'}\left[R_1 + \gamma R_2 + \cdots + \gamma^{n-1} R_n + \gamma^n Q^\pi(s_n, \pi(s_n))\right] \geq V^\pi(s) \qquad (1)$$

هم‌چنین برای V_{n+1} نیز داریم:

$$V_{n+1}(s) = \mathbb{E}_{\pi'}\left[R_1 + \gamma R_2 + \cdots + \gamma^{n-1} R_n + \gamma^n \mathbb{E}_{a\sim\pi'}[Q^\pi(s_n, a)]\right] \qquad (2)$$

از طرفی چون سیاست‌های π در روش ϵ-greedy خود سیاست‌هایی تصادفی هستند، دایم:

$$Q^\pi(s, \pi(s)) = \mathbb{E}_{a\sim\pi}[Q^\pi(s, a)] \qquad (*)$$

هم‌چنین با توجه به فرض مسئله، داریم:

$$\forall s : E_{a\sim\pi'}[Q^\pi(s, a)] \geq E_{a\sim\pi}[Q^\pi(s, a)]$$
$$\implies E_{a\sim\pi'}[Q^\pi(s_n, a)] \geq E_{a\sim\pi}[Q^\pi(s_n, a)] \qquad (**)$$

در نتیجه با جایگذاری رابطه‌ی $(*)$ در (1) و استفاده از $(**)$ در مقایسه‌ی روابط (1) و (2) به راحتی می‌توان نتیجه گرفت که: $V_{n+1}(s) \geq V_n(s) \geq V^\pi(s)$.

بنابراین مراحل استقرا تکمیل شده و حکم اثبات می‌شود. از آن‌جایی که به ازای هر n داریم $V_n(s) \geq V^\pi(s)$، اگر تمام قدم‌ها را نیز از π' پیروی کنیم باز هم نابرابری برقرار و $V^{\pi'}(s) \geq V^\pi(s)$ خواهد بود.

The END!