هوش مصنوعی - تمرین تئوری سری 6

محمدامین نظامی - 98102346

1

(آ) درست؛ چرا ضریب تخفیف کوچک موجب کوچک شدن rewards می‌شود که این مورد در هر قدم از بازی موجب رفتار بهینه یابی محلی (رویکرد greedy) می‌شود

(ب) درست؛ پاداش منفی زندگی موجب در نظر گرفتن جریمه برای انجام تعداد حرکت بیشتر می‌شود که بزرگی آن معادل این است که reward ها و جریمه‌های مراحل بعدی در تصمیم گیری بی اثر شده و حرکاتی که موفقیت فعلی را به پاداش می‌رساند انتخاب شود

(ج) نادرست؛ تاثیر این دو پارامتر به شکل متفاوتی اعمال می‌شود بدین ضریب تخفیف منفی رفتار ضرری روی پاداش خواهد داشت اما پارامتر منفی زندگی تاثیر جمعی روی پاداش خواهد داشت

(د) نادرست؛ طبق توضیحات بخش قبلی رفتار این دو پارامتر به دو شکل متفاوت اعمال می‌شود

2

(آ) فرض می‌کنیم که بازی به شکل episodic باشد که در هر مرحله i مقدار پاداش دریافتی برابر R_i باشد. برای پاداش کلی داریم

$$R_{total} = G_t + R_{t+1} + R_{t+2} \cdots \xrightarrow[\text{factor}]{\text{discount}} G_t + R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

پس مقدار ارزش هر State با توجه به policy، π به شکل $E[Q_t]$ تعریف می‌شود

$$V^{\pi}(S) = \mathbb{E}_{\pi}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+2} + \cdots \mid S_t = S\right]$$

طبق این رابطه مقدار پاداش تا جایی که بازی ادامه داشته باشد، روی مقدار پاداش هر مرحله به شرط آنکه S شروع کننده بازی باشد، امید ریاضی گرفته می‌شود و محاسبه می‌شود در صورت تغییر وضعیت بین است که چه مقدار پاداش بدست خواهد آمد

$$V^{\pi}(S) = \mathbb{E}_{\pi}\left[R_{t+1} + \gamma_x\left(V^{\pi}(S_{t+1})\right) \mid S_t = S\right]$$

(ب)

$$\Rightarrow V^{\pi}(S) = \sum_{\alpha \in A} \pi(S,\alpha)\left(\tilde{R}(S,a) + \gamma \sum_{S' \in S} T(S,\alpha,S')\, V^{\pi}(S')\right)$$

که در این رابطه $\tilde{R}(S,\alpha) = \sum_{S' \in S} T(S,\alpha,S') \times R(S,\alpha,S')$ و $\pi(S,\alpha)$ احتمال انتخاب حرکت α

در اینجا S است (چرا که به دلیل وجود حالت C، استراتژی deterministic نخواهیم داشت) و استراتژی ما mix خواهد بود

(ج) در policy evaluation داریم:

$$V_{k+1}^{\pi}(S) = \sum_{S' \in S} T(S, \pi(S), S') \left[R(S, \pi(S), S') + \gamma V_k^{\pi}(S') \right]$$

که با توجه به انتخاب استراتژی mix باید با احتمال هر S وزن دهی شود $\left(E[V_{k+1}^{\pi}(S)] \right)$

$$V_2^{\pi}(A) = -8 + 0.5 \times V_1(B) = -7 \quad \& \quad V_2^{\pi}(B) = 0.5\left(2 + 0.5(V_1(A))\right) + 0.5 \times (-2 + 0.5 V_1(C)) = 1$$

$$V_2^{\pi}(C) = 0.5\left(8 + 0.5\, V_1(B)\right) + 0.5\left(4 + 0.5 \times \left(\tfrac{1}{4} V_1(A) + \tfrac{3}{4} V_1(C)\right)\right) = 7$$

(د) برای هر State باید مقداری که Q را بیشینه می‌کند را بدست آوریم

یک انتخاب اولیه

$$\pi_2(A) = ab$$

$$Q_2(B, ba) = -2 + 0.5\, V_2(A) = -5.5 \quad \& \quad Q_2(B, bc) = -2 + 0.5\, V_2(C) = 1.5$$

$$\Rightarrow \pi_2(B) = bc$$

$$Q_2(C, ca) = 4 + 0.5\left(\tfrac{1}{4} V_2(A) + \tfrac{3}{4}(V_2(B))\right) = -5.75 \quad \& \quad Q_2(C, cb) = 8 + 0.5 V_2(B) = 8.5$$

$$\Rightarrow \pi_2(C) = cb$$

(ه) اگر π' به صورت greedy رفتار کند یعنی $\pi'(S) = \underset{\alpha \in A}{\operatorname{argmax}} Q^{\pi}(S, \alpha)$:

$$Q^{\pi}(S, \pi(S)) = \max_{\alpha \in A} Q^{\pi}(S, \alpha) \geq Q^{\pi}(S, \pi(S)) = V^{\pi}(S) \Rightarrow \text{increase reward}$$

$$V^{\pi}(S) \leq Q^{\pi}(S, \pi'(S)) = E_{\pi'}\left[R_{t+1} + \gamma V^{\pi}(S_{t+1}) \mid S_t = S\right] \leq E_{\pi'}\left[R_{t+1} + \gamma Q^{\pi}(S_{t+1}, \pi'(S_{t+1})) \mid S_t = S\right] =$$

$$\leq \cdots \leq E_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \cdots \mid S_t = S\right] \cdot V^{\pi'}(S)$$

اگر در عبارت فوق $Q^{\pi}(S, \pi(S)) = V^{\pi}(S) \leq Q^{\pi}(S, \pi'(S)) = \max_{\alpha \in A} Q^{\pi}(S, \alpha)$ شود (توقف): $V^{\pi}(S)$ شرایط Bellman را ارضا می‌کند، π' و π هردو optimal خواهند بود

3

(آ)

$$Q_{k+1}(s,\alpha) = \sum_{s'} T(s,\alpha,s')\left[R(s,\alpha,s') + \gamma \max_{\beta}\left(Q_k(s',\beta)\right)\right]$$

$$\Rightarrow Q(3,left) = \sum_{s'=\{2\}} T(3,left,s')\left[R(3,left,s') + 0.9 \max_{\beta=\{U,R,L\}}\left\{Q(s',\beta)\right\}\right]$$

$$\Rightarrow Q(3,left) = -1 + 0.9\ max(8,3,6) = 6.2$$

(ب)

در استراتژی حریصانه در هر State بین حالت‌های شناخته شده آنکه بیشترین پاداش را می‌دهد انتخاب خواهد شد اما در اینجا که شناخت از محیط وجود ندارد، ممکن است action ای باشد که State بعدی آن و reward آن ناشناخته (اما بسیار زیادتر از بقیه پاداش‌ها باشد) باشد و در الگوریتم greedy انتخاب نشود که باعث ناهوشمندی این الگوریتم در محیط ناشناخته خواهد بود

برای برقراری تعادل باید بین exploration و exploitation تعادل برقرار کرد و گاهی با عمل غیر حریصانه باید exploration را افزایش داد تا به سیاستی (راهی) با reward بیشتر و global تری نزدیک شد.

(ج) چرا که استفاده از Q-Value بدون نیاز به مدل محیط می‌شویم و می‌توانیم با $\pi(S) = \underset{\alpha}{argmax}\ (Q(S,\alpha))$ حرکت بهینه را پیدا کنیم اما اگر از V-Value استفاده کنیم محاسبات پیچیده‌تر خواهد بود و نمی‌توانیم سیاست بهینه را نیز بدست آوریم

(د) با این کار همه action ها احتمال انتخاب می‌گیرند که باعث می‌شود که هرکشی احتمال انتخاب داشته باشد که معادل برآورده کردن exploration است که در حالت greedy آن را نداشتیم.

$$S=1: \pi(1,up) = \frac{e^4}{e^3+e^4} \ \&\ \pi(1,right) = \frac{e^3}{e^3+e^4}$$

$$S=2: \pi(2,right) = \frac{e^8}{e^8+e^3+e^6} \ \&\ \pi(2,left) = \frac{e^3}{e^3+e^8+e^6} \ \&\ \pi(2,up) = \frac{e^6}{e^3+e^8+e^6}$$

$$S=3: \pi(3,left) = \frac{e^7}{e^7+e^9} \ \&\ \pi(3,up) = \frac{e^9}{e^7+e^9}$$

$$S=4: \pi(4,right) = \frac{e^5}{e^5+e^2} \ \&\ \pi(4,down) = \frac{e^2}{e^5+e^2}$$

$S = 5$: $\pi(5, right) = \frac{e^8}{e^8+e^5+e^6}$ & $\pi(5, left) = \frac{e^5}{e^8+e^5+e^6}$ & $\pi(5, down) = \frac{e^6}{e^8+e^5+e^6}$

(د)

$Q(2, up) \leftarrow Q(2, up) + 0.2\left[-1 + 0.8\,Q(5, right) - Q(2, up)\right] = 4.8$

$Q(5, right) \leftarrow Q(5, right) + 0.2\left[+10 + 0.8\,Q(6) - Q(5, right)\right] = 8.4$

(24

(آ) ابتدا مقدار Q ها را 0 در نظر می‌گیریم و با هر نمونه آن را بروزرسانی می‌کنیم: فرض: $\alpha = 0.1$ & $\gamma = 0.9$

$Q(A,1) \leftarrow Q(A,1) + 0.1\left[-3 + 0.9 \max_{\beta} Q(B,\beta) - Q(A,1)\right] = 0.3$

$Q(B,1) \leftarrow Q(B,1) + 0.1\left[4 + 0.9 \max_{\beta} (Q(A,\beta)) - Q(B,1)\right] = 0.427$

$Q(A,2) \leftarrow Q(A,2) + 0.1\left[-4 + 0.9 \max_{\beta} (Q(A,\beta)) - Q(A,2)\right] = -0.373$

$Q(A,1) \leftarrow Q(A,1) + 0.1\left[-3 + 0.9 \max_{\beta} (Q(B,\beta) - Q(A,1))\right] = 0.00843$

$Q(A,2) \leftarrow Q(A,2) + 0.1\left[9 + 0.9 \times 1 - Q(A,2)\right] = -0.23$

(ب) این سیاست می‌تواند آن باشد که در $S=A$، $a=2$ انتخاب شود چرا که در $S=B$ نیز $a=1$، $reward=-3$ است اما در حالت $a=1$ $reward = \frac{-4+1}{2} = -1.5$

انتخاب کنیم تا از انتخاب رندوم R بیشتری دریافت کنیم

(ج) از آنجایی که Q-learning، off-policy است، انتظار داریم با استفاده از هریک به مقدار یکسان برسیم چرا که Value مستقل update policy است. سیاست random هرگز دچار exploitation نخواهد شد و محیط را explore خواهد کرد بدون توجه به مقدار reward. در مقابل سیاست greedy (حریصانه)، exploit می‌کند و state های ناشناخته (بدون تجربه) را explore نخواهد کرد که ممکن است با یادگیری آن‌ها reward بیشتری را تجربه کنیم

5 اگر بخواهیم یک الگوریتم برای کاهش این نرخ ε استفاده کنیم، می توانیم با هر بار ε را انجام action، ε را به شکل $\exp(-\frac{n}{10})$ کاهش دهیم $\frac{\alpha}{n}$ اما این نسبت به تغیر استراتژی حریف امن نخواهد برای اینکه از این مورد نیز در امان باشیم می توانیم در هر مرحله α را با ضریب بسیار کوچکی از تفاوت estimated R با پاداش دریافتی از بازی افزایش دهیم تا در صورتی که سیاست حریف تغیر کرد (که معادل عدم همخوانی پاداش دریافتی با پاداش تخمینی است) مقدار α افزایش یابد