راه حل تکلیف شماره‌ی ۴

فصل ۲۰ و ۲۱ یادگیری (یادگیری مدل‌های احتمالاتی / یادگیری تقویتی)

۱) رویکرد بیزی معرف کردن هر دو دارو است. رویکرد ماکزیمم درست‌نمایی معرف داروی anti-B است. در این مورد که دو نسخه از B وجود دارد، رویکرد بیزی هنوز معرف هر دو دارو را پیشنهاد می‌دهد در حالی که رویکرد ماکزیمم درست‌نمایی، معرف داروی anti-A است، چرا که دارای 40% شانس درست بودن دارد در برابر شانس 30% برای هر یک از موارد B. البته این یک کاریکاتور است و شما ممکن است برای یافتن پزشک تحت فشار باشید، حتی یک طرفدار ماکزیمم درست‌نمایی سرسخت که ممکن است مشابه این تجویز کند!

۲) داریم:

$$L = -m(\log\sigma + \log\sqrt{2\pi}) - \sum_j \frac{(y_j - (\theta_1 x_j + \theta_2))^2}{2\sigma^2}$$

با مشتق‌گیری نسبت به پارامترها:

$$\frac{\partial L}{\partial \theta_1} = -\sum_j \frac{x_j(y_j - (\theta_1 x_j + \theta_2))}{\sigma^2} = 0$$

$$\frac{\partial L}{\partial \theta_2} = -\sum_j \frac{(y_j - (\theta_1 x_j + \theta_2))}{\sigma^2} = 0$$

$$\frac{\partial L}{\partial \sigma} = -\frac{m}{\sigma} + \sum_j \frac{(y_j - (\theta_1 x_j + \theta_2))^2}{\sigma^3} = 0$$

که راه‌حل دستگاه فوق می‌شود:

$$\theta_1 = \frac{m\left(\sum_j x_j y_j\right) - \left(\sum_j y_j\right)\left(\sum_j x_j\right)}{m\left(\sum_j x_j^2\right) - \left(\sum_j x_j\right)^2}$$

$$\theta_2 = \frac{1}{m}\sum_j (y_j - \theta_1 x_j)$$

$$\sigma^2 = \frac{1}{m}\sum_j (y_j - (\theta_1 x_j + \theta_2))^2$$

۳) الف) با انتگرال‌گیری بر روی بازه‌ی $[0,1]$، ثابت نرمال‌سازی برای توزیع $beta[a,b]$ به‌صورت
$\alpha = \Gamma(a+b)/\Gamma(a)\Gamma(b)$

داده می‌شود که در آن $\Gamma(x)$ تابع گاما است.

برای اعداد صحیح a و b با استقرا روی a داریم: فرض می‌کنیم $\alpha(a,b)$ ثابت نرمال‌سازی باشد.

برای حالت پایه داریم:

$$\alpha(1,b) = 1/\int_0^1 \theta^0 (1-\theta)^{b-1}\, d\theta = -1/\left[\tfrac{1}{b}(1-\theta)^b\right]_0^1 = b$$

و

$$\frac{\Gamma(1+b)}{\Gamma(1)\Gamma(b)} = \frac{b\cdot\Gamma(b)}{1\cdot\Gamma(b)} = b.$$

برای گام استقراء فرض می‌کنیم برای هر b داریم:

$$\alpha(a-1, b+1) = \frac{\Gamma(a+b)}{\Gamma(a-1)\,\Gamma(b+1)} = \frac{a-1}{b}\;\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

حال با انتگرال‌گیری جزء به جزء مقدار $\alpha(a,b)$ را ارزیابی می‌کنیم:

$$\begin{aligned}
1/\alpha(a,b) &= \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\, d\theta \\
&= \left[\theta^{a-1}\cdot\frac{1}{b}(1-\theta)^b\right]_1^0 + \frac{a-1}{b}\int_0^1 \theta^{a-2}(1-\theta)^b\, d\theta \\
&= 0 + \frac{a-1}{b}\;\frac{1}{\alpha(a-1,b+1)}
\end{aligned}$$

پس داریم:

$$\alpha(a,b) = \frac{b}{a-1}\,\alpha(a-1,b+1) = \frac{b}{a-1}\;\frac{a-1}{b}\;\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

ب) میانگین با انتگرال زیر داده می‌شود:

$$\begin{aligned}
\mu(a,b) &= \alpha(a,b)\int_0^1 \theta\cdot\theta^{a-1}(1-\theta)^{b-1}\, d\theta = \alpha(a,b)\int_0^1 \theta^a (1-\theta)^{b-1}\, d\theta \\
&= \alpha(a,b)/\alpha(a+1,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\;\frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b+1)} = \frac{a}{a+b}
\end{aligned}$$

ج) یافتن مُد با مشتق‌گیری:

$$d\, beta[a,b](\theta)/d\theta = 0$$

$$\frac{d}{d\theta}\left(\alpha(a,b)\,\theta^{a-1}(1-\theta)^{b-1}\right) =$$

$$\alpha(a,b)\left[(a-1)\theta^{a-2}(1-\theta)^{b-1} - (b-1)\theta^{a-1}(1-\theta)^{b-2}\right] = 0 \Rightarrow$$

$$(a-1)(1-\theta) = (b-1)\theta \Rightarrow \theta = \frac{a-1}{a+b-2}$$

۴)

مزیت یادگیری تفاضل زمانی مقادیر Q (Q-learning) به
یادگیری تفاضل زمانی ارزش حالات V چیست؟

اگر از یادگیری تفاضل زمانی بر روی ارزش های حالات استفاده شود،
استخراج policy از روی ارزش های یاد گرفته شده، دشوار خواهد بود.
(نیاز به مدل گذر T داریم تا بتوان سیاست را از روی ارزش ها محاسبه کنیم.)
برای یادگیری تفاضل زمانی مقادیر Q، یک سیاست می تواند مستقیماً با معادلهٔ زیر استخراج شود:

$$\pi^*(s) = \arg\max_{a} Q^*(s,a)$$

۵)

در برخی مسایل RL ، پاداش‌ها (rewards) برای اهداف ، مثبت و در دیگر موارد صفر یا منفی هستند.

آیا علامت پاداش‌ها مهم است یا تنها فاصله‌ی میان آنها اهمیت دارد؟

با استفاده از تعریف دریافتی تخفیف یافته R_t (discounted reward) ، ثابت کنید که افزودن یک مقدار ثابت C به همه‌ی پاداش‌های ابتدایی ، ثابت K را به ارزش همه‌ی حالات می‌افزاید و بنابراین بر ارزش‌های نسبی همه‌ی حالات تحت همه‌ی سیاست‌ها تأثیری نمی‌گذارد.

* مقدار K را برحسب C و γ محاسبه نمایید.

راه حل :

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots$$
$$= \sum_{i=0}^{\infty} \gamma^i r_{t+i+1}$$

با اضافه کردن C به همه‌ی پاداش‌های ابتدایی ، داریم :

$$r'_t = r_t + C$$

و از آنجا

$$R'_t = \sum_{i=0}^{\infty} \gamma^i r'_{t+i+1} = \sum_{i=0}^{\infty} \gamma^i (r_{t+i+1} + C) = \underbrace{\sum_{i=0}^{\infty} \gamma^i r_{t+i+1}}_{R_t} + \underbrace{\sum_{i=0}^{\infty} \gamma^i C}_{K}$$

در نتیجه :

$$R'_t = R_t + K$$

که در آن

$$K = C \sum_{i=0}^{\infty} \gamma^i = C \frac{1}{1-\gamma}$$

لذا تنها فاصله‌ی میان پاداش‌ها اهمیت دارند ، نه مقادیر مطلق آنها.

۶)

اثبات همگرایی Q-learning

راه حل: طرح کلی اثبات:

- حالت دنیا را قطعی deterministic در نظر بگیرید که در آن هر (s,a) به تعداد دفعات نامتناهی مشاهده می‌شود.
- یک بازهٔ کامل را به عنوان بازه‌ای که در خلال آن هر (s,a) مشاهده می‌شود در نظر بگیرید. interval
- نشان دهید که در خلال هر چنین بازه‌ای، مقدار قدر مطلق خطا (حداکثر خطا) در جدول Q با ضریب γ کاهش می‌یابد.
- در نتیجه، برای $\gamma < 1$، پس از بی‌نهایت هنگام‌سازی، حداکثر خطا به سمت صفر میل می‌کند.

راه حل ←

فرض کنید $\hat{Q}_n$ جدولی باشد که پس از n هنگام‌سازی به دست آمده است و

e_n حداکثر خطا در این جدول باشد:

$$e_n = \max_s \max_a |\hat{Q}_n(s,a) - Q(s,a)|$$

حداکثر خطا را پس از $(n+1)$ امین هنگام‌سازی می‌یابیم:

$$\begin{aligned}
e_{n+1} &= |\hat{Q}_{n+1}(s,a) - Q(s,a)| \\
&= |(\cancel{r} + \gamma \max_{a'} \hat{Q}_n(s',a')) - (\cancel{r} + \gamma \max_{a'} Q_n(s',a'))| \\
&= \gamma |\max_{a'} \hat{Q}_n(s',a') - \max_{a'} Q_n(s',a')| \\
&\le \gamma \max_{a'} |\hat{Q}_n(s',a') - Q_n(s',a')| \\
&\le \gamma \max_{s''} \max_{a'} |\hat{Q}_n(s'',a') - Q_n(s'',a')| \\
&= \gamma e_n
\end{aligned}$$

بنابراین:

$$e_{n+1} \le \gamma e_n \quad , \quad e_{n+1} \ge 0 \quad \Rightarrow$$

$$\frac{e_{n+1}}{e_n} \le \gamma \Rightarrow \prod_{i=0}^{n} \frac{e_{i+1}}{e_i} \le \prod_{i=0}^{n} \gamma \Rightarrow \frac{e_{n+1}}{e_0} \le \gamma^{n+1}$$

$$\Rightarrow e_{n+1} \le e_0 \gamma^{n+1}$$

$$\Rightarrow e_m \le e_0 \gamma^m \quad , \quad e_m \ge 0$$

$$0 \le \lim_{m\to\infty} e_m \le \lim_{m\to\infty} e_0 \gamma^m = 0 \Rightarrow \qquad 0 \le \gamma < 1$$

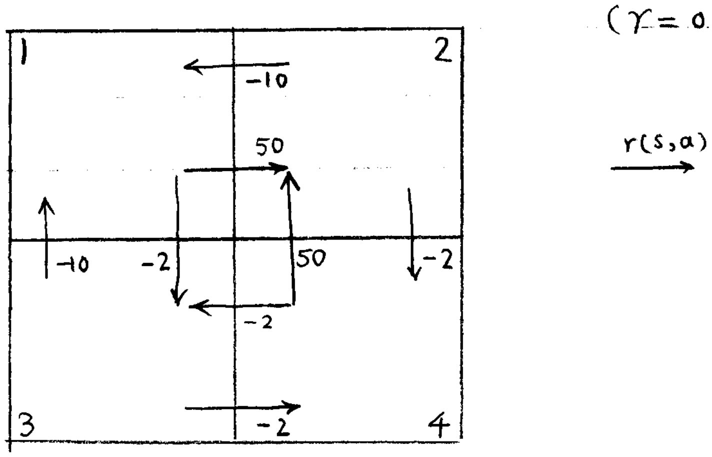
$$\lim_{m\to\infty} e_m = 0$$

توجه می‌نمایید که هیچ فرضی بر روی دنباله‌ی actionها وجود ندارد، بنابراین Q_learning می‌تواند تابع Q را یاد بگیرد (و بنابراین سیاست بهینه را optimal policy)، درحالی که از actionهای انتخاب شده به طور تصادفی آموزش می‌بیند.

مادامی که دنباله‌ی یادگیری حاصل، هر (s,a) را بی‌نهایت مرتبه مشاهده می‌کند.

۷)

شکل زیر یک دنیای شبکه‌ای چهار حالتی را به تصویر کشیده است، که در آن حالت 2، "طلا" را نشان می‌دهد. با استفاده از مقادیر پاداش بلافصل (immediate rewards) نشان داده شده بر روی شکل و بکارگیری الگوریتم Q-learning، بر روی حالات به صورت پادساعتگرد حرکت کنید و جدول state_action را بهنگام نمایید. $(\gamma = 0.9)$



راه حل:

ابتدا همهٔ درایه‌های جدول مقادیر Q را به صفر مقداردهی می‌کنیم:

$$\forall s \forall a \left(s \in \{1, 2, 3, 4\} \wedge a \in \{\rightarrow, \leftarrow, \uparrow, \downarrow\} \Rightarrow Q(s,a) = 0 \right)$$

حال فرمول زیر را تکراری کنیم:

$$Q(s,a) \leftarrow r(s,a) + \gamma \max_{a'} \{ Q(s', a') \}$$

$$(s) \xrightarrow{a} (s') \xrightarrow{a'} *$$

دور اول:

$$Q(3, \rightarrow) = -2 + 0.9 \max\{Q(4, \leftarrow), Q(4, \uparrow)\} = -2$$

$$Q(4, \uparrow) = 50 + 0.9 \max\{Q(2, \downarrow), Q(2, \leftarrow)\} = 50$$

$$Q(2, \leftarrow) = -10 + 0.9 \max\{Q(1, \rightarrow), Q(1, \downarrow)\} = -10$$

$$Q(1, \downarrow) = -2 + 0.9 \max\{Q(3, \uparrow), Q(3, \rightarrow)\} = -2$$

$$Q(3, \rightarrow) = -2 + 0.9 \max\{Q(4, \leftarrow), \underbrace{Q(4, \uparrow)}_{50}\} = 43$$

a

Q	↑	↓	→	←
1	–	-2	0	–
2	–	0	–	-10
3	0	–	43	–
4	50	–	–	0

S

دور دوم :

$Q(4,\uparrow) = 50 + 0.9\max\{Q(2,\downarrow), Q(2,\leftarrow)\} = 50 + 0.9\max\{0,-10\} = 50$

$Q(2,\leftarrow) = -10 + 0.9\max\{Q(1,\rightarrow), Q(1,\downarrow)\} = -10 + 0.9\max\{0,-2\} = -10$

$Q(1,\downarrow) = -2 + 0.9\max\{Q(3,\uparrow), Q(3,\rightarrow)\} = -2 + 0.9\max\{0,43\} = 36.7$

$Q(3,\rightarrow) = -2 + 0.9\max\{Q(4,\leftarrow), Q(4,\uparrow)\} = -2 + 0.9\max\{0,50\} = 43$

a

Q	↑	↓	→	←
1	–	36.7	0	–
2	–	0	–	-10
3	0	–	43	–
4	50	–	–	0

S

دور سوم : تنها $Q(2,\leftarrow)$ تغییر می‌کند :

$Q(2,\leftarrow) = -10 + 0.9\max\{Q(1,\rightarrow), Q(1,\downarrow)\} = -10 + 0.9\max\{0,36.7\} = 23.03$

دور چهارم : تنها $Q(4,\uparrow)$ و $Q(3,\rightarrow)$ تغییر می‌کند :

$Q(4,\uparrow) = 50 + 0.9\max\{Q(2,\downarrow), Q(2,\leftarrow)\} = 50 + 0.9\max\{0,23.03\} = 70.73$

$Q(3,\rightarrow) = -2 + 0.9\max\{Q(4,\leftarrow), Q(4,\uparrow)\} = -2 + 0.9\max\{0,70.73\} = 61.66$

a

Q	↑	↓	→	←
1	–	36.7	0	–
2	–	0	–	23.03
3	0	–	61.66	–
4	70.73	–	–	0

S

۸) MDPs و RL

یک روبات متحرک خودکار را درنظر بگیرید که می‌تواند SLOW یا FAST در هر گام زمانی حرکت کند.

حرکت FAST به طور کلی پاداش +2 را می‌دهد، درحالی که

حرکت SLOW تنها پاداش +1 را می‌دهد.

به هرحال، روبات باید دمای داخلی خود را به حساب آورد، که می‌تواند HOT یا OK باشد.

حرکت کردن در حالت SLOW منجر به دمای کمتری می‌شود، در حالی که

حرکت کردن در حالت FAST منجر به افزایش دما می‌شود.

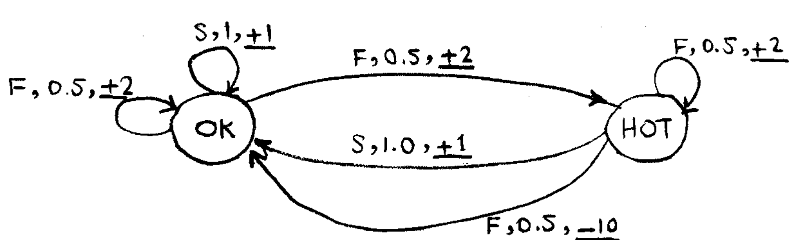
اگر روبات HOT باشد، این خطر وجود دارد که overheat شود که در این نقطه باید بایستد، دمایش را کم کند و تعمیر شود. گذرهای MDP و پاداش‌ها به صورت زیر مشخص می‌شود:

s	a	s′	T(s,a,s′)	R(s,a,s′)
OK	SLOW	OK	1.0	+1
OK	FAST	OK	0.5	+2
OK	FAST	HOT	0.5	+2
HOT	SLOW	OK	1.0	+1
HOT	FAST	HOT	0.5	+2
HOT	FAST	OK	0.5	-10

SLOW S

FAST F

توجه کنید که با اینکه تعمیر وقت‌گیر است، روبات پس از آن OK می‌شود (سطر آخر جدول)



آ) دو تکرار از value iteration در جدول زیر را اجرا کنید. (با استفاده از discount $\gamma = 0.8$)

(خانه‌های هاشور خورده را درنظر نگیرید)

S	V_0	V_1	V_2
OK	0	2	3.2
HOT	0	1	//////

$$V_{i+1}(s) = \max_{a} \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_i(s')]$$

$$V_1(OK) = \max\{1(1+\gamma \times 0)\ ,\ 0.5(2+\gamma \times 0) + 0.5(2+\gamma \times 0)\} = \max\{1, 2\} = 2$$

$$V_1(HOT) = \max\{1(1+\gamma \times 0)\ ,\ 0.5(2+\gamma \times 0) + 0.5(-10+\gamma \times 0)\} = \max\{1, -4\} = 1$$

$$V_2(OK) = \max\{1(1+\gamma \times 2)\ ,\ 0.5(2+\gamma \times 2) + 0.5(2+\gamma \times 1)\} = \max\{2.6, 3.2\} = 3.2$$

ب) Q-learning را با تخفیف $\overset{\gamma}{0.8}$ اجرا کنید و از نرخ یادگیری $\overset{\alpha}{0.5}$، با استفاده از نمونه‌های گذر زیر کمک بگیرید. مقادیر Q ای که در یک گام تغییر نمی‌کند را دوباره نویسی نکنید.

نمونه‌های زیر را با تجربه‌های عامل زیر فرض کنید:

(OK, FAST, HOT), reward +2, calculate Q_1

(HOT, FAST, OK), reward −10, calculate Q_2

(OK, SLOW, OK), reward +1, calculate Q_3

$$Q(s,a) \leftarrow Q(s,a) + 0.5\,[R(s,a,s') + 0.8 \max_{a'}(s',a') - Q(s,a)]$$

$$Q_1(OK, FAST) = 0 + 0.5\,[2 + 0.8 \max_{a'} Q(HOT, a') - 0] = 1$$

$$Q_2(HOT, FAST) = 0 + 0.5\,[-10 + 0.8 \max_{a'} Q(OK, a') - 0] = -4.6$$

$$Q_3(OK, SLOW) = 0 + 0.5\,[1 + 0.8 \max_{a'} Q(OK, a') - 0] = 0.9$$

S	a	Q_0	Q_1	Q_2	Q_3
OK	SLOW	0			0.9
OK	FAST	0	1.0		
HOT	SLOW	0			
HOT	FAST	0		−4.6	