AI/HW3/Q1

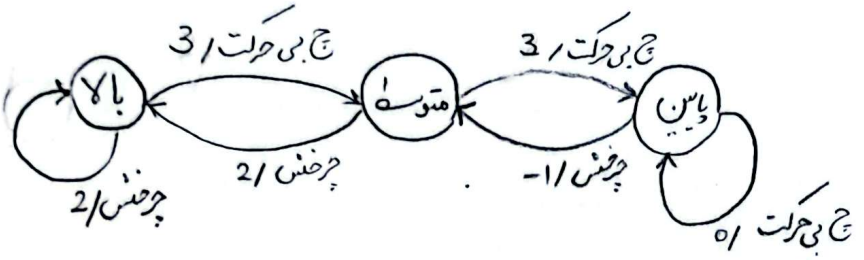
مهدی شاهینی 9923040

تنبید هر حرکت:
- ب → -1
- م → -1
- پ → -1

پاداش خانه‌ها:
- بالا → +3
- متوسط → +3
- پایین → 0

گره‌ها ← محل قرارگیری ربات
یال‌ها ← تغییر محل قرارگیری
نام یال ← نام act/پاداش و تنبیه خانه و act
- چرخش
- بی‌حرکت

گراف ماشین حالت



جدول کلاره‌ها بر حسب شرایط اولیه ∘ حل می‌کنیم.
خانه‌های ستاره‌دار حرکت با max امتیاز همراه با تنبیه بعدی است.

امتیاز اولیه ∘

محل قرارگیری / حرکت بعدی	تکرار ∘ خانه آینده	1	2	3
پایین — چرخش	متوسط	-1 = 0 - 1	LS2 = 1.4 * بالا	LS3 = 2.52 *
پایین — بی‌چرخش	پایین	0 = 0 + 0 *	LD2 = 0.4 پایین	LD3 = 1.12
متوسط — چرخش	بالا	2 = 0 + 2	MS2 = 4.4 * بالا	MS3 = 6.32 *
متوسط — بی‌چرخش	پایین	3 = 0 + 3 *	MD2 = 3.4 پایین	MD3 = 4.12
بالا — چرخش	بالا	2 = 0 + 2	HS2 = 4.4 بالا	HS3 = 6.32
بالا — بی‌چرخش	متوسط	3 = 0 + 3 *	HD2 = 5.4 * متوسط	HD3 = 6.52 *

LS2: پ → م : 0.8 × 3 − 1 = 1.4 LD2: پ → پ 0.8 × 0 + 0 = 0

مثال: 2 ↘ max(M1) ↓ ↘ پ → م

MS2: م → ب : 0.8 × 3 + 2 = 4.4 MD2: م → پ 0.8 × 0 + 3 = 3

HS2: ب → ب : 0.8 × 3 + 2 = 4.4 HD2: 0.8 × 3 + 3 = 5.4

LS3: ب → م : 0.8 × 4.4 − 1 = 2.52 LD3: ب → پ : 0.8 × 1.4 + 0 = 1.12

MS3: م → ب : 0.8 × 5.4 + 2 = 6.32 MD3: م → پ : 0.8 × 1.4 + 3 = 4.12

HS3: ب → ب : 0.8 × 5.4 + 2 = 6.32 HD3: ب → م : 0.8 × 4.4 + 3 = 6.52

سیاست بهینه این است که اگر خانه پایین هستیم ابتدا به خانه متوسط برویم. سپس چرخه زیر را تکرار کنیم

متوسط ⟵ بالا (2 ↺) ⟶ چرخه بهینه

AI/HW3/Q2

مهدی شاهینی 9923040

الف) نادرست. در الگوریتم Q-L. اصلا سیاست برای این است که همواره به exploration و exploitation باشد و با احتمال بزرگتر از صفر برای exploration تا سیاست بهینه یافت شود.

ب) درست. چون مقدار تنبیه برای عامل بسیار زیاد است؛ عامل ناچاراً مجبور است سریعتر مسیر بهینه را بیابد (درصورت وجود جواب).

ج) درست. این ضریب اگر بین ۰ تا ۱ باشد با کاهش آن از exploration می‌کاهیم و به exploitation می‌پردازیم. چون برابر ۱ است عملاً exploitation نخواسته و تماماً exploration است. چون همواره اکشن بهینه انتخاب می‌شود تقریباً به جواب خواهیم رسید.

Q3

الف)

$V^{\pi}(C) = \frac{1}{2}(1+1+1+10) + \frac{1}{2}(4+1+10) = 14$

because $V^{\pi}(s) = \sum_{s'} T(s,\pi(s),s')[R(s,\pi,s') + \gamma V^{\pi}(s')]$

ب) اکشن بهینه * دارد (همان max است)

$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_k(s')]$

تکرار اول V_1: $A \to 1$, $B \to 1$, $C = \frac{1+4}{2} = 2.5$, $D = 1$, $E = 1$, $F = 10$, $G = 0$

تکرار دوم V_2: $1+1 = 2$, $\begin{cases} * \ 1+2.5 = 3.5 \\ 1-2 = -1 \end{cases}$, $\begin{cases} \frac{1}{2}(4+1) + \frac{1}{2}(1+1) = 3.5 \ * \\ 2.5 - 2 = 0.5 \end{cases}$, $\begin{cases} 1+1 = 2 \ * \\ 1-2 = -1 \end{cases}$ $\begin{cases} 1+10 = 11 \ * \\ 1-2 = -1 \end{cases}$ $\begin{cases} 10+0 = 10^{*} \\ 10-2 = 8 \end{cases}$, $G = 0$

$V_2(B) = \max(3.5, -1) = 3.5$ $\quad Q_2(B, right) = 1 + [1 + 2.5] = 3.5$ $\quad Q_2(B, Left) = 1[-2+1] = -1$

$Q_{k+1} = \sum T[R + \gamma \max Q_k]$

نخش ج صفحه بعد

AI/HW3/Q3-C

فهرستی 9923040

—	Q(C,L)	Q(C,J)	Q(E,L)	Q(E,R)	Q(F,L)	Q(F,R)
init	0	0	0	0	0	0
trans.1	0	$\frac{0}{2}+\frac{1}{2}(4+0)=2$	0	0	0	0
" 2	0	0	$0\ \frac{1}{2}(0)+\frac{1}{2}(1)=0.5$	0	0	
" 3	0	0	0	$0\ \frac{0}{2}+\frac{1}{2}(-2+0.5)=-0.75$	0	
" 4	0	0	$0\ \frac{1}{2}(0)+\frac{1}{2}(1+0)=0.5$	0	0	

$$(1-\alpha)\,Q(s,a)+\alpha\left[R+\gamma \max_{a'}(Q',a')\right]=\frac{1}{2}Q_K(s,a)+\frac{1}{2}(R+Q_K(s',a'))=Q_{K+1}(s,a)$$

Q4 $V^{\pi_i}_{K+1}(s)=\sum_{s'} T(s,\pi_i(s),s')\left[R(s,\pi_i,s')+\gamma V^{\pi_i}_K(s')\right]$ همان فرمول Q3 منتی

$U_0=0$ Firststate: "First Rank"

$V_1^{\pi}(FR)=T(FR,R,FR)\left[R(FR,R,FR)+\gamma V(FR)\right]$

$+\ T(FR,R,BR)\left[R(FR,R,BR)+V(BR)\right]+T(FR,W,FR)\left[R(FR,W,BR)+V(FR)\right]$

$+\ \cdots = \underbrace{1(0+0)+0(0+0)}_{=0=Q(FR,R)}+\underbrace{0.9(10+0)+0.1(-10+0)}_{=8=Q(FR,W)}+\underbrace{1(1+0)+0(1+0)}_{=1=Q(FR,Rest)}$

$*V_1^{\pi}(FR)=8\ FR\to W$

$V_1^{\pi}(BR)=\underbrace{.3(0+0)+.7(0+0)}_{=0=Q(BR,Run)}+\underbrace{0.05(10+0)+.95(-10+0)}_{=-9=Q(BR,W)}+\underbrace{0(1+0)+1(1+0)}_{=1=Q(BR,Rest)}$

$*V_1^{\pi}(BR)$

$V_2^{\pi}(FR)=\underbrace{1(0+.5(8))+0(0+.5\times1)}_{4=Q(FR,Run)}+\underbrace{.9(10+.5\times8)+.1(-10+.5\times1)}_{=11.65=Q(FR,W)}+\underbrace{0(1+.5\times8)+1(1+.5\times1)}_{=1.5=Q(FR,Rest)}$

$*V_2^{\pi}(FR)=11.65\ FR\to W$

$V_2^{\pi}(FR)=\underbrace{.3(0+.5\times8)+.7(0+.5\times1)}_{1.55=Q(BR,Run)}+\underbrace{.05(10+.5\times8)+.95(-10+.5\times1)}_{=-8.325=Q(BR,W)}+\underbrace{0(1+.5\times8)+1(1+.5\times1)}_{1.5=Q(BR,Rest)}$

$*V_1^{\pi}(BR)$

ادامه صفحه بعد

AI HW3 IQ4 مهلا

مهدی شاهینی 9923040

در تکرار اول و دوم الویت حرکت در موقعیت First Rank تغییری نمی‌کند و Water می‌ماند

اما در تکرار دوم، اولویت اول و دوم حرکت برای موقعیت Bad Rank تغییر می‌کند اما از تکرار را ادامه دهیم الگوریتم همگرا شده و الویت‌ها همین ترتیب باقی خواهند ماند.

Q5

الف) $Q($ خوشحال , ترن هوایی قدیمی $) = w_0 b + w_1 x_1 + w_2 x_2 + w_3 x_3 = 1 + 2 + (-0.5) = 3.5$

(1) (1) (2) (1) (1) (0) (0.5) (1)

$w^T = [w_0, w_{3\times 1}^T]^T$, $X = [b \; \underset{1\times 3}{x}]$

ب) به صورت قطعی از خوشحال به ناراحت می‌رویم ($T = 1$)، پاداش -10.5 دریافت می‌کنیم؛ داریم: (مطابق روش الگوریتم)

$$-10.5 + \tfrac{1}{2} V^* = -10.5 + \tfrac{1}{2} \max\{3.5, 4, 2.5, 2, 2\} = -8.5 \quad Q_{K+1}^{old}(tran)$$

new w:

$w_0 = 1 + \frac{1}{2}(-12) = -5$ $w_2 = 1 + \frac{1}{2}(-12) = -5$ ← مشابه بروزرسانی V^*

$w_1 = 2 + \frac{1}{2}(-12) = -4$ $w_3 = -0.5 + \frac{1}{2}(-12) = -5.5$

$3.5 = Q_K(old\,train)$

$\Delta Q(old\,train) = -12$

ج) بله تکیلا هشترا در جدول MDP و توصیفات تنبل یا تشویق داشته به فرضی پیدا نشد واتری ندارد.

د)

وزن قدیمی

ترن قدیمی طولانی رفت ← ۳٫۵ ترن جدید کوتاه کند ← $۱ + ۲ + ۱ = ۴$

سقوط کوتاه رفت ← $۱ + ۱ + ۰٫۵ = ۲٫۵$ ← رفتن کوتاه کند ← $۱ + ۱ = ۲$

ترک ← $۱ + ۱ = ۲$

با احتمال $\frac{4\epsilon}{5}$ سایر حرکات و $1 - \epsilon + \frac{\epsilon}{5} = 1 - \frac{4\epsilon}{5}$ حرکت ترن جدید کوتاه کند انتخاب می‌شود.

ه) این تابع را به عنوان exploitation در نظر می‌گیرد و K را تأثیر exploration فرض می‌کند تا $\frac{K}{n}$ بخش exploration و 1 به عنوان پاداش دریافتی از Q (مثلاً ۷) باشد.

AI / HW3 / Q6

مهدی شاهینی 9923040

الف) $Q((UCLA, UPSET), BUS) = w_0 b + w_1 x_1 + w_2 x_2 + w_3 x_3 = 5$

$1 \quad 1 \quad 2.5 \quad 1 \quad .5 \quad 1 \quad 1 \quad 1$

$Q((UCLA, UPSE), Taxi) = 1\times1 + 2.5\times0 + .5\times1 + 1\times0 = 1.5$

nowiti
no stops

$Q((UCLA, UPET), Move) = 1.\times1 + 1W \times zeros(3,1) = 1$

ب) طبیعی به نظر می‌رسد. در پس انتشار خطا برای محاسبه و بروزرسانی وزن‌ها قدیمی، خوشحالی و ناراحتی وزن ندارند پس تأثیری ندارند.

ج) $(1-\epsilon)$ اتوبوس انتخاب می‌شود و با احتمال ϵ رندم چون آن‌ها هم $\frac{\epsilon}{3}$ ممکن است اتوبوس انتخاب شود پس

اتوبوس $\boxed{(1-\frac{2}{3}\epsilon)}$

هرکدام $\frac{\epsilon}{3}$ ← $\boxed{\frac{2}{3}\epsilon \text{ سایرین}}$

د) $w_i = w_i + \alpha \underbrace{[r + \gamma \max_{a'} Q_w(s',a') - Q_w(s,a)]}_{.5(-7 + .5 \max\{5, 1.5, 1\}) = -2.25} \frac{\partial}{\partial w_i} Q_w(s,a)$

2.5

$[w_0 \quad w_1 \quad w_2 \quad w_3] = [1 \quad 2.5 \quad .5 \quad 1] - 2.25 \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}^T = [-2 \quad 2.5 \quad -2.5 \quad 1]$

ه) هنگام تعریف تابع Q و بروزرسانی Q همراه با وزن‌ها پیش روش‌های مختلف وجود دارد همچنین ممکن است علاوه بر تنوع انتخاب، انتخاب الگوریتمی که همگرا شود سخت و نا ممکن باشد. اما مزیت آن این است که در محیط‌هایی که پیچیده‌اند و پارامترهای زیادی هستند، در انتخاب اتیار و کنش مناسب راهگشا باشد.