کلاغی عبدی 09036162048

Subject : Year : Month : Date :

① الف) حالت: موقعیت گوی $(X-Y)$ در صفحه

اکشن: چه حرکتی شود که باعث تغییر حالت صفحه نشود. (زاویه رول و پیچ صفحه)

پاداش: قدر دادن یک پاداش مثبت به ضربه به خط برد

فاصله گوی از مرکز $(-\sqrt{x^2+y^2})$

و همچنین یک پاداش زیاد به قدر که منفی گوی در مرکز (حالت تعادل) و یک پاداش منفی زیاد

برد افتادن گوی از صفحه

ب) حالت: مقدار سرمایه (موجودی)

اکشن: خرید یا فروش کالا و مقدار کالا

پاداش: پاداش مثبت برد اندازه‌ای که سرمایه و پاداش منفی برد کاهش آن (Δx)

② ب)

Right -1 Y Right +4

X, Right, +1, X, ... , terminal

X, Left, 0, X, Left, 0, ...

الف)

$G_{02} R_1 + \gamma R_2 + \gamma^2 R_3 = 1 + \frac{1}{2}(1) + \frac{1}{2}^2(4)$

$= 1.5$

SABA

Subject : Year : Month : Date :

$$V_\pi(s) = \sum_a P(a|s) \sum_{s',r} P(s',r|s,a)\left[r + \gamma V_\pi(s')\right]$$

$$V_{\pi_1}(Y) = 4 + 0.5 \times V_{\pi_1}(\text{terminal}) = 4$$

$$V_{\pi_2}(X) = \frac{3}{4}\left(1 + 0.5 \times V_{\pi_2}(X)\right) + \frac{1}{4}\left(-1 + 0.5 \times V_{\pi_2}(Y)\right)$$

$$V_{\pi_2}(Y) = 4 + 0.5 \times V_{\pi_2}(\text{terminal}) = 4$$

$$\Rightarrow V_{\pi_2}(X) = \frac{3}{4} + \frac{3}{8} V_{\pi_2}(X) + \frac{1}{4}(1)$$

$$\Rightarrow \frac{5}{8} V_{\pi_2}(X) = 1 \Rightarrow V_{\pi_2}(X) = \frac{8}{5} = 1.6$$

(۳)

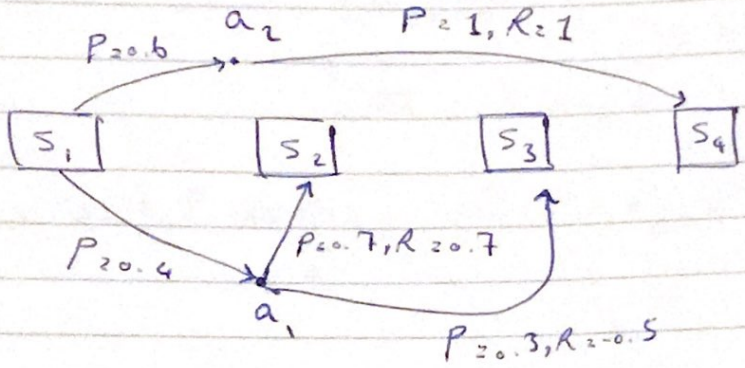
$$G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \cdots$$

$$= 5 + 10(\gamma + \gamma^2 + \gamma^3 + \cdots) = 5 + 10\,\frac{\gamma}{1-\gamma}$$

$$= 5 + 10\,\frac{0.9}{0.1} = 95$$

$$G_0 = \underbrace{R_1}_{5} + \overset{0.9}{\gamma} G_1 = 95 \Rightarrow 90 = 0.9\, G_1 \Rightarrow G_1 = 100$$

(۴)



$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} P(s',r|s,a)\left[r + \gamma V_\pi(s')\right]$$

$$V_\pi(s_1) = 0.4 \times \left[0.7\left(0.5 + \gamma V_\pi(s_2)\right) + 0.3\left(-0.5 + \gamma V_\pi(s_3)\right)\right] + 0.6\left(1 + \gamma V_\pi(s_4)\right)$$

$$q_\pi(s_1, a_1) = 0.7\left(0.5 + \gamma V_\pi(s_2)\right) + 0.3\left(-0.5 + \gamma V_\pi(s_3)\right)$$

$$q_\pi(s_1, a_2) = 1\left(1 + \gamma V_\pi(s_4)\right)$$

$$\Rightarrow V_\pi(s_1) = 0.6\, q_\pi(s_1, a_2) + 0.4\, q_\pi(s_1, a_1)$$

(۹)

$q_*(s,a) = q_\pi(s,a) + f(s)$ که $f(s) = f(s): S \to \mathbb{R}$

تابع $f(s)$، تابعی از حالت است و تغییری بر روی سیاست حاصل از آن نخواهد داشت.

حال با استفاده از رابطه زیر می‌توان به سیاست بهینه دست پیدا کرد.

Subject : Year : Month : Date :

$\pi(s) = \arg\max_a q_\pi(s,a)$

$\arg\max_a q'_*(s,a) = \pi'_*$

$\arg\max_a q_*(s,a) + f(s) = \arg\max_a q_*(s,a) = \pi_*$

$\Rightarrow \pi'_* = \pi_*$

(4)

$$V_\pi(x) = \frac{1}{3}\Big[\underbrace{0.25(-1.6 + 0.9\times 8) + 0.75(0.2 + 0.9\times 2)}_{a} + \underbrace{(-2.7 + 0.9\times 3) + 0.2(-5 + 0.9\times 0)}_{b} + \underbrace{0.8(1.6 + 0.81\times V_\pi(x1))}_{c}\Big]$$

$\Rightarrow V_\pi(x) = \dots$

$V_\pi^*(x) = \max\left[\frac{1}{3}a, \frac{1}{3}b, \frac{1}{3}c\right] = \max[0.96, 0, 0.85]$

$= 0.96 \Rightarrow$ L

(5)

$$V_\pi(s_{13}) = \frac{1}{4}\Big[0 + \underbrace{0.92\times 2.3}_{\text{up}} + \underbrace{0.92\times 0.4}_{\text{right}} + \underbrace{0.92\times 0.7}_{\text{down}} + \underbrace{0.92\times(-0.41)}_{\text{left}}\Big]$$

$\simeq 0.7$