MDP HW_ML

Thursday, September 30, 2021 4:44 AM

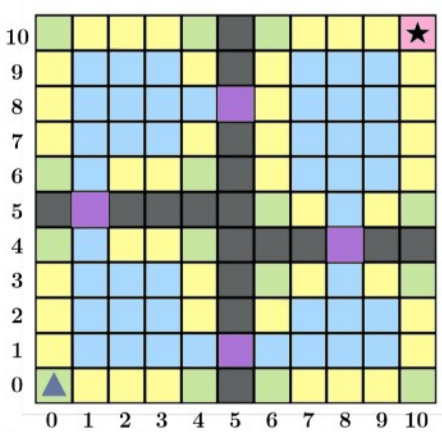
Q1 : (a) Action = { left, right, up, down }.

State = (x, y) ; $x, y \in [0, 10]$.

$(x, y) \neq (5, i), (j, 5), (k, 4)$

$$\begin{cases} i = 0, 2, 3, 4, 5, 6, 7, 9, 10 \\ j = 0, 2, 3, 4. \\ k = 6, 7, 9, 10 \quad \text{(except "Walls")} \end{cases}$$

(b)



As show in figure :

it have 44 states that can move in 4 direction (blue),

40 states that can move in 3 direction (yellow)

(15+4) states that can move in 2 direction (green & purple)

and 1 goal state that return to the start state.

Therefore,

$$44 \times 4 \times 3 + 40 \times 4 \times 3 + 15 \times 3 \times 2 + 15 \times 2 \times 2 + 4 \times 2 \times 2 + 4 \times 2 \times 3 + 1 = 1199.$$

I think there are 1199 non-zero rows in this conditional probability table.

Q2 : (a) episodic function : $G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$.

with discounting : $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$.

$$= \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1} = -\gamma^{T-t-1}$$

continuing function : $G_t' = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = \gamma^k$

Because "episodic" has only 1 failure, it returns the last G_t,

while "continuing" will have many failures and the return value

will always be updated.

(b) $R_1 = R_2 = \cdots = R_{t+k} = 0$; $R_{t+k+1} = 1$.

Then the G_t will always be 1, so it will never have improvement.

Q3: (a) $G_5 = R_6 + \gamma G_6 = 0$

$G_4 = R_5 + \gamma G_5 = 2$

$G_3 = R_4 + \gamma G_4 = 3 + 0.5 \times 2 = 4$

$G_2 = R_3 + \gamma G_3 = 6 + 0.5 \times 4 = 8$

$G_1 = R_2 + \gamma G_2 = 2 + 0.5 \times 8 = 6$

$G_0 = R_1 + \gamma G_1 = -1 + 0.5 \times 6 = 2$

(b) $G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \cdots + \gamma^n R_{n+2}$

$= \sum_{i=0}^{\infty} \gamma^i R_{n+2}$. $(R_1 = 2,\ R_2 = R_3 = \cdots = R_n = 7)$

$= \frac{1}{1-\gamma} \times 7 = \frac{1}{1-0.9} \times 7 = 70$

$G_0 = R_1 + 0.9 \times 70 = 2 + 63 = 65.$

Q4: $\sum_{i=1}^{100} \gamma^i = \gamma^1 + \gamma^2 + \cdots + \gamma^{100} = \frac{(\gamma^{100}-1)\gamma}{\gamma - 1}$

$$\begin{cases} G_0 = R_1 + \sum_{i=1}^{100} \gamma^i R_{101} = 50 + \sum_{i=1}^{100} \gamma^i \times (-1) = 50 - \frac{(\gamma^{100}-1)\gamma}{\gamma-1} \\ G_0' = R_1' + \sum_{i=1}^{100} \gamma^i R_{101} = (-50) + \sum_{i=1}^{100} \gamma^i \times 1 = (-50) + \frac{(\gamma^{100}-1)\gamma}{\gamma-1} \end{cases}$$

When $G_0 > G_0'$: $\quad 50 - \frac{(\gamma^{100}-1)\gamma}{\gamma-1} > -50 + \frac{(\gamma^{100}-1)\gamma}{\gamma-1}$

$$50 > \frac{(\gamma^{100}-1)\gamma}{\gamma-1}$$

$$-1.047 < \gamma < 0.9844$$

Thus, while $0 < \gamma < 0.9844$, choose [UP]; else, choose [DOWN]

Q5: (a) $G_t = \sum_{i=0}^{\infty} \gamma^i R_{t+k+1}$

$V_{\pi}(s) \doteq E_\pi[G_t \mid S_t = s] = E_\pi\left[\sum_{i=0}^{\infty} \gamma^i R_{t+k+1} \mid S_t = s\right]$

$V_{\pi(s)}(c) \doteq E_\pi\left[\sum_{i=0}^{\infty} \gamma^i (R_{t+k+1} + c) \mid S_t = s\right]$

$= E_\pi\left[\sum_{i=0}^{\infty} \gamma^i R_{t+k+1} + \sum_{i=0}^{\infty} \gamma^i c \mid S_t = s\right]$

$= E_\pi\left[\sum_{i=0}^{\infty} \gamma^i R_{t+k+1} \mid S_t = s\right] + E_\pi\left[\sum_{i=0}^{\infty} \gamma^i c \mid S_t = s\right]$

$$= E_\pi[\sum_{i=0} \gamma^i R_{t+k+1} + \sum_{i=0} \gamma^i c \mid S_t = s]$$
$$= E_\pi[\sum_{i=0}^{\infty} \gamma^i R_{t+k+1} \mid S_t = s] + E_\pi[\sum_{i=0}^{\infty} \gamma^i c \mid S_t = s]$$
$$= V_\pi(s) + \sum_{i=0}^{\infty} \gamma^i c$$
$$= V_\pi(s) + \frac{c}{1-\gamma}$$

(b) If we make the reward after each action a positive number, then the result will remain the same, it will not move. Since the first time it already got a positive result. The requirement to add a constant without bringing a effect is that the reward remains negative after adding the constant, and if it becomes positive after adding the constant, it will have the effect I mentioned before.

Q6: (a) $V_\pi(s) = \frac{1}{4} \times 1 \times (0 + 0.9 \times 2.3) + \frac{1}{4} \times 1 \times (0 + 0.9 \times 0.4)$

$+ \frac{1}{4} \times 1 \times (0 + 0.9 \times (-0.4)) + \frac{1}{4} \times 1 \times (0 + 0.9 \times 0.7)$

$= \frac{1}{4} \times 0.9 \times 2.3 + \frac{1}{4} \times 0.9 \times 0.7$

$= 0.5175 + 0.1575 = 0.675$

(b) $V_\pi(s) = \frac{1}{2} \times 1 \times (0 + 0.9 \times 19.8) + \frac{1}{2} \times 1 \times (0 + 0.9 \times 19.8) = 17.82$.

Q7: (a) I guess the value function $= \frac{1}{2} \times 0 + \frac{1}{2} \times 1 = \frac{1}{2}$.

Verify: $V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma V_\pi(s')]$

$= \frac{1}{2} \times 1 \times [1 + 1 \times 0] + \frac{1}{2} \times 1 \times [0 + 1 \times 0]$

$= \frac{1}{2} \times 1 + \frac{1}{2} \times 0 = \frac{1}{2}$.

(b) $V_{\pi(s)}(A) = \frac{1}{2} \times 1 \times [0 + 1 \times 0] + \frac{1}{2} \times 1 \times [0 + 1 \times V_{\pi(s)}(B)] = \frac{1}{2} V_{\pi(s)}(B)$

$V_{\pi(s)}(B) = \frac{1}{2} \times 1 \times [0 + 1 \times V_{\pi(s)}(A)] + \frac{1}{2} \times 1 \times [0 + 1 \times V_{\pi(s)}(C)] = \frac{1}{2} V_{\pi(s)}(A) + \frac{1}{2} V_{\pi(s)}(C)$

$V_{\pi(s)}(C) = \frac{1}{2} \times 1 \times [0 + 1 \times V_{\pi(s)}(B)] + \frac{1}{2} \times 1 \times [0 + 1 \times V_{\pi(s)}(D)] = \frac{1}{2} V_{\pi(s)}(B) + \frac{1}{2} V_{\pi(s)}(D)$

$V_{\pi(s)}(D) = \frac{1}{2} \times 1 \times [0 + 1 \times V_{\pi(s)}(C)] + \frac{1}{2} \times 1 \times [0 + 1 \times V_{\pi(s)}(E)] = \frac{1}{2} V_{\pi(s)}(C) + \frac{1}{2} V_{\pi(s)}(E)$

$$V_{\pi(s)}(D) = \frac{1}{2}\times 1\times[0+1\times V_{\pi(s)}(C)] + \frac{1}{2}\times 1\times[0+1\times V_{\pi(s)}(E)] = \frac{1}{2}V_{\pi(s)}(C) + \frac{1}{2}V_{\pi(s)}(E)$$

$$V_{\pi(s)}(E) = \frac{1}{2}\times 1\times[0+1\times V_{\pi(s)}(D)] + \frac{1}{2}\times 1\times[1+1\times 0] = \frac{1}{2}V_{\pi(s)}(D) + \frac{1}{2}$$

$$\begin{cases} V_A = \frac{1}{6}; & V_D = 4V_A = \frac{2}{3}; \\ V_B = 2V_A = \frac{1}{3}; & V_E = 5V_A = \frac{5}{6}. \\ V_C = 3V_A = \frac{1}{2}; & \end{cases}$$

(c) $V_{\pi(s)}(i) = \frac{i}{n-1}$ $[i = 1, 2, 3 \cdots (n-2)]$

Q8: (a) $S(high) = \{search, wait\}$.

$S(low) = \{search, wait, recharge\}$.

$v(s) = E[R_{t+1} + \gamma(S_{t+1}) | S_t = s]$.

$$v(high) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma V_\pi(s')]$$

$$= \pi(search|high)[r_{search} + \alpha\times\gamma\times v(high) + (1-\alpha)\times\gamma\times v(low)]$$

$$+ \pi(wait|high)[1\times(r_{wait} + \gamma\cdot v(high))]$$

$$v(low) = \pi(search|low)[r_{search} + \beta\times\gamma\times v(low) + (1-\beta)\times\gamma\times v(high)]$$

$$+ \pi(wait|low)[1\times(r_{wait} + \gamma v(low))]$$

$$+ \pi(recharge|low)[0 + 1\times\gamma v(high)]$$

(b) $v(high) = 1\times[10 + 0.8\times 0.9\times v(high) + 0.2\times 0.9 v(low)]$

$$= 10 + 0.72 v(high) + 0.18 v(low)$$

$$v(low) = 0.5\times[3 + 0.9 v(low)] + 0.5\times 0.9 v(high)$$

$$= 1.5 + 0.45 v(low) + 0.45 v(high).$$

$$\begin{cases} 0.28 v(high) = 10 + 0.18 v(low) \\ 0.55 v(low) = 1.5 + 0.45 v(high) \end{cases}$$

Thus $\{ v(high) = 79.04$

Thus $\begin{cases} v(\text{high}) = 79.04 \\ v(\text{low}) = 67.40. \end{cases}$

Q9: (a) $V_\pi(s) \doteq E_\pi[G_t | S_t = s]$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right]$$

$$q_\pi(s,a) \doteq E_\pi[G_t | S_t = s, A_t = a]$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s, A_t = a\right]$$

$$V_\pi(s) = \sum_a \pi(a|s)\, q(s,a)$$

(b) $q_\pi(s,a) \doteq E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s, A_t = a\right]$

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

(c) $V_\pi(s') = \sum_{a'} \pi(a'|s')\, q(s',a')$

$$q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma\left(\sum_{a'} \pi(a'|s')\, q(s',a')\right)\right]$$