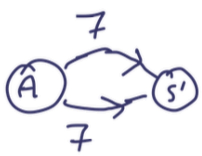
400108547

پرهام رضائی

(1)

الف) غلط - می‌تواند آفلاین هم از ریواردها قبلی درصورتی که محیط تغییر نکند استفاده کند، برای مثال با importance sampling از ریواردهای یک پالیسی دیگری استفاده کند و خود را ترین کند. و جدا از این کلی الگوریتم‌های offline RL هم نیز برقرار اند.



ب) غلط - ریوارد ناشی از دو اکشن مختلف می‌تواند یکسان باشد ← که در این صورت هر دو اکشن پالیسی بهینه در این حالت اند.

ج) غلط - REINFORCE که نمونه کارهای پالیسی گرادیانت هست مستقیماً خود پالیسی را تغییر می‌دهد و استفاده‌ای از مدل مثل ترنزیشن‌ها و ... ندارد، پس model-free است.

د) صحیح - اثبات در note های لکچر کورس CS234 2019 صفحه 15,16 و همچنین در کتاب sutton&barto موجود است. (اثبات از آنجا آورده شود)

[در صورت سوال ذکر شده برای غلط‌ها دلیل بیاوریم ← نیاوردم برای صحیح روش هم فقط گذاردم]

(2)

الف)

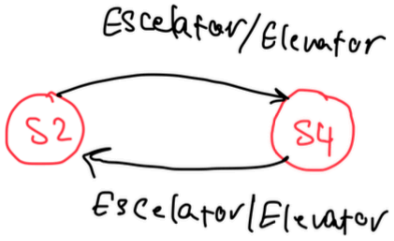
$$\pi(S) = \arg\max_a Q(S,a)$$

$$\pi(S_1) = \arg\max [1/a, -1/a] = \text{Elevator}$$

$$\pi(S_2) = \arg\max [+2/1, -0/3] = \text{Elevator}$$

$$\pi(S_3) = \arg\max [+0/8, +0/9] = \text{Escalator}$$

ب)



ج)

$$\pi(S_2) = \arg\max [+0/08, +2/08] = \text{Escalator}$$

$$\pi(S_4) = \arg\max [+1/18, +1/38] = \text{Escalator}$$

د) ریوارد حرکت S4 به S2 بیشتر از حالتی که بین S1 یا S3 بودیم فرق می‌کرد و به نوعی اکشن قبلی طی S2 به S4 که با elevator یا escelator بوده روی ریوارد حرکت S4 به S2 تأثیر داشته پس MDP فرض مارکوین نیست و در طول این تغییر information loss رخ داده که در نتیجه سیاست بهینه هم لزوماً برابر نمی‌شد. (که نشده است)

3

الف)

$$\mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)} R(s,a) = \sum R(s,a)\,p(s)\,p(a|s) = \sum R(s,a)\,p(s)\,\pi_1(s,a)$$

$$= \sum R(s,a)\,p(s)\,\pi_0(s,a)\,\frac{\pi_1(s,a)}{\pi_0(s,a)} = \sum p(s)\,\pi_0(s,a)\left(R(s,a)\,\frac{\pi_1(s,a)}{\pi_0(s,a)}\right)$$

$$\approx \mathbb{E}_{s\sim p(s),a\sim\pi_0(s,a)} R(s,a)\,\frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)}$$

$\infty \to$ تعداد سمپل ها

نسبت ادرج
طبق تابع امید ریاضی

ب)

با میل سمپل ها به ∞، صورت کسر مثل الف $\mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)} R(s,a)$ میشه

باید ثابت کنیم $\mathbb{E}$ خارج کسر می‌شود

$$\mathbb{E}_{s\sim p(s),a\sim\pi_0(s,a)} \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} \underset{\text{بی‌نهایت سمپل}}{\approx} \sum p(s)\,\pi_0(s,a)\,\frac{\pi_1(s,a)}{\pi_0(s,a)}$$

$$= \sum p(s)\,\pi_1(s,a) = \sum p(s)\,p(a|s) = \mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)} 1 = 1$$

$\mathbb{E}(C)=C$ است.
C ثابت

$\Rightarrow$ Q.E.D

پرهام رضائی
400108547

(۵

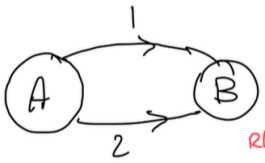
اگر شرط صرفا یک دنیا داشته باشیم . (S_1, a_1)

$$\frac{E_{s\sim p(s|,a\sim\pi_0(s,a)} \frac{\pi_1(s,a)}{\pi_0(s,a)} R(s,a)}{E_{s\sim p(s),a\sim\pi_0(s,a)} \frac{\pi_1(s,a)}{\pi_0(s,a)}} = \frac{\frac{\pi_1(s_1,a_1)}{\pi_0(s_1,a_1)} R(s_1,a_1)}{\frac{\pi_1(s_1,a_1)}{\pi_0(s_1,a_1)}} = R(s_1,a_1)$$

$E(R(s_1,a_1)) = R(s_1,a_1)$

برای اینکه $R(s,a)$ برمبنای π_0 صرفا داریم R برمبنای π_0 آمده کم

می دهد که لزوما برای π_1 نیست. مثلا در این است



$R(B)=0$

stochastic

حتما A باشد

$R(A,1)=4$

$R(A,2)=0$

$\pi_0(1|A)=0.8$

$\pi_0(2|A)=0.2$

$\pi_1(1|A)=0.5$

$\pi_1(2|A)=0.5$

تنها گرفتن از π_0

$(A,1,4)$

آنکه استیتر 4 شده.

$\Rightarrow R(s,a)=0.5\times 0+0.5\times 4=2$

π_1 طبق

$2\neq 4$

(۲ $\pi(s) =$ ملع

$$v = R + \gamma P v \implies v = (I - \gamma P)^{-1} R \quad \star$$

ماتریس Transition احتمالات Peace ها در جدول نوشته شده است.

$$\begin{bmatrix} v(M) \\ v(R) \\ v(D) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v(M) \\ v(R) \\ v(D) \end{bmatrix}$$

$$\star \implies \begin{bmatrix} v(M) \\ v(R) \\ v(D) \end{bmatrix} = \begin{bmatrix} 1 & -0.5\gamma & -0.5\gamma \\ 0 & 1-\gamma & 0 \\ 0 & 0 & 1-\gamma \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}$$

ایجاد با استفاده از فرم کرمر یا صافت ایتریتیو درایه‌ها (که [نشود)

$$\begin{bmatrix} 1 & -0.5\gamma & -0.5\gamma \\ 0 & 1-\gamma & 0 \\ 0 & 0 & 1-\gamma \end{bmatrix} \begin{bmatrix} 1 & \frac{\gamma}{2-2\gamma} & \frac{\gamma}{2-2\gamma} \\ 0 & \frac{1}{1-\gamma} & 0 \\ 0 & 0 & \frac{1}{1-\gamma} \end{bmatrix} = I$$

$$\implies \begin{bmatrix} v(M) \\ v(R) \\ v(D) \end{bmatrix} = \begin{bmatrix} 1 & \frac{\gamma}{2-2\gamma} & \frac{\gamma}{2-2\gamma} \\ 0 & \frac{1}{1-\gamma} & 0 \\ 0 & 0 & \frac{1}{1-\gamma} \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 + \frac{\gamma}{1-\gamma} - \frac{\gamma}{2-2\gamma} \\ \frac{2}{1-\gamma} \\ \frac{1}{\gamma - 1} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{2-\gamma}{2-2\gamma} \\ \frac{2}{1-\gamma} \\ \frac{-1}{1-\gamma} \end{bmatrix}$$

$$\left[\begin{array}{ccc|ccc} 1 & -0.5\gamma & -0.5\gamma & 1 & 0 & 0 \\ 0 & 1-\gamma & 0 & 0 & 1 & 0 \\ 0 & 0 & 1-\gamma & 0 & 0 & 1 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|ccc} 1 & 0 & -0.5\gamma & 1 & \frac{-0.5\gamma}{1-\gamma} & 0 \\ 0 & 1-\gamma & 0 & 0 & 1 & 0 \\ 0 & 0 & 1-\gamma & 0 & 0 & 1 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & \frac{-0.5\gamma}{1-\gamma} & \frac{-0.5\gamma}{1-\gamma} \\ 0 & 1-\gamma & 0 & 0 & 1 & 0 \\ 0 & 0 & 1-\gamma & 0 & 0 & 1 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & \frac{-0.5\gamma}{1-\gamma} & \frac{-0.5\gamma}{1-\gamma} \\ 0 & 1 & 0 & 0 & \frac{1}{1-\gamma} & 0 \\ 0 & 0 & 1-\gamma & 0 & 0 & 1 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & \frac{-0.5\gamma}{1-\gamma} & \frac{-0.5\gamma}{1-\gamma} \\ 0 & 1 & 0 & 0 & \frac{1}{1-\gamma} & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{1-\gamma} \end{array}\right]$$

ب)

$\gamma = 0.9$

V ها را که طبق الف داریم. یک مرحله policy improvement را اجرا می کنیم.

$$\hat{\pi}(s) = \underset{a \in A}{\arg\max} \left[R(s) + \gamma \sum_{s' \in S} P(s' \mid s, a)\, V^{\pi}(s') \right]$$

$$V(M) = \frac{2 - 0.9}{2 - 2 \times 0.9} = \frac{1.1}{0.2} = 5.5$$

$$V(R) = \frac{2}{1 - 0.9} = 20$$

$$V(D) = \frac{-1}{1 - 0.9} = -10$$

$$\hat{\pi}(\text{Mountain}) = \underset{P, W}{\arg\max} \left[1 + \gamma \left[(0.5) V(R) + 0.5\, V(D) \right],\ 1 + \gamma \left[0.1 V(M) + 0.2 V(D) + 0.7 V(R) \right] \right]$$

$$= \underset{P, W}{\arg\max} \left[1 + \frac{0.9}{2}(20 - 10),\ 1 + 0.9(.55 + (-2) + 14) \right]$$

$$= \underset{P, W}{\arg\max} \left[5.5,\ 12.295 \right] = \text{War}$$

$$\hat{\pi}(\text{Riverside}) = \underset{P, W}{\arg\max} \left[2 + \gamma (1 \cdot V(R)),\ 2 + \gamma (0.2 V(R) + 0.8 V(M)) \right]$$

$$= \underset{P, W}{\arg\max} \left[2 + 0.9(20),\ 2 + 0.9(0.2 \times 20 + 0.8 \times 5.5) \right]$$

$$= \underset{P, W}{\arg\max} \left[20,\ 9.56 \right] = \text{Peace}$$

$$\hat{\pi}(\text{Desert}) = \underset{P, W}{\arg\max} \left[-1 + \gamma (1 \cdot V(D)),\ -1 + \gamma (1 \cdot V(M)) \right] =$$

$$\underset{P, W}{\arg\max} \left[-1 + 0.9 \times (-10),\ -1 + 0.9(5.5) \right] = \underset{P, W}{\arg\max} \left[-10,\ 3.95 \right] = \text{War}$$

(c

M-P	M-W	R-P	R-W	D-P	D-W	
0	0	0	0	0	0	0
0	0	0	0	0	-1	1
0	0	0	0	0	1	2
0	0	0.5	0	0	1	3
0.75	0	0.5	0	0	1	4

1: $Q(D,W) = Q(D,W) + \alpha(R + \lambda \max_a Q(D,a) - Q(D,W))$

$Q(D,W) = 0 + 0.5(-2 + 1 \max(0,0)) = 0.5 \times 2 = -1$

2: $Q(D,W) = Q(D,W) + \alpha(R + \lambda \max_a Q(R,a) - Q(D,W))$

$Q(D,W) = -1 + 0.5(3 + 1 \max(0,0) - (-1)) =$

$-1 + 0.5(3+1) = -1 + 2 = 1$

3: $Q(R,P) = Q(R,P) + \alpha(R + \lambda \max_a Q(M,a) - Q(R,P))$

$Q(R,P) = 0 + 0.5(1 + 1 \max(0,0) - 0) = 0.5$

4: $Q(M,P) = Q(M,P) + \alpha(R + \lambda \max_a Q(R,a) - Q(M,P))$

$Q(M,P) = 0 + 0.5(1 + 1 \max(0.5,0) - 0) =$

$0.5(1 + 0.5) = 0.5(1.5) = 0.75$

Mountain-peace	Riverside-peace	Desert-war
.	.	.
0	0	-1
0	0	1
0	0.5	1
0.75	0.5	1