محمدرضا احمدی تشنیزی
۹۸۱۷۰۴۴۶

سؤال ۱:

الف) نادرست ؛ زیرا ممکن است داده‌ها از قبل به دست آمده باشند و روی این داده‌ها یادگیری تقویتی را انجام دهیم.

ب) نادرست ؛ اگر در حالتی بیش از یک حالت باعث به دست آوردن بیشینه پاداش شود در این حالت سیاست بهینه یکتا نیست.

ج) نادرست ؛ این الگوریتم هم می‌تواند model based و هم model free باشد.

د) درست ؛

$$V^*(s) = \max_a \sum_{s'} T(s,a,s')\{R(s,a,s') + \gamma V^*(s')\}$$

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

چون در حالت اول بیشینه پاداش را انتخاب می‌کنیم $\Rightarrow V^*(s) \geq V^\pi(s)$

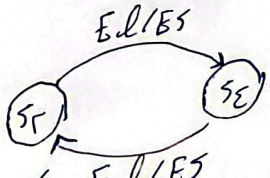
سؤال ۲:

الف)

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

$$\Rightarrow \pi(s_1) = E\ell,\ \pi(s_2) = E\ell,\ \pi(s_3) = E5$$

ب) و ج)



$$\pi(s_2) = E5,\ \pi(s_4) = E5$$

د) در حالت دوم که مسئله را دیده شده است و اطلاعاتی از بین رفته است، با توجه به اینکه مارکوف باید بتوانیم با توجه به اطلاعات state فعلی به طور مستقل از state های قبلی برای state بعدی تصمیم بگیریم ولی در اینجا اطلاع نداریم که چگونه به s_4 رسیده‌ایم پس سیاست بهینه نیز تغییر خواهد کرد.

۹۸۱۷۰۴۲۶

محمدرضا احمدی تشنیزی

سوال ۳:

$$E_{s\sim p(s),\, a\sim \pi_0(s,a)} \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} R(s,a) = ?$$

$$= \sum \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} R(s,a)\, p(s)\, p(a|s) =$$

$$= \sum \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} R(s,a)\, p(s)\, \pi_0(a|s) = \sum \pi_1(s,a) R(s,a) p(s)$$

$$= E_{s\sim p(s),\, a\sim \pi_1(s,a)}$$

ب)

$$E_{s\sim p(s),\, a\sim \pi_0(s,a)} \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} =$$

$$= \sum \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} p(s)\, \pi_0(a|s) = \sum p(s)\, \pi_1(s,a) = 1$$

$$\Rightarrow \frac{E_{s\sim p(s),\, a\sim \pi_0(s,a)} \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} R(s,a)}{\frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)}} =$$

$$= E_{s\sim p(s),\, a\sim \pi_1(s,a)}$$

ج)

$\pi_0 \rightarrow a=0$ $\quad R(s,a)=a$

$\pi_1 \rightarrow a=1$

$\Rightarrow E_{\pi_1} = 1$ but مقدار خروجی تخمین گر = تعریف نشده

سؤال ٤:

الف)

$V^{\pi}(M) = 0.5(3 + 0.7\,V^{\pi}(R)) + 0.5(-1 + 0.7\,V^{\pi}(D))$

$V^{\pi}(R) = (2 + 0.7\,V^{\pi}(R))$

$V^{\pi}(D) = (-1 + 0.7\,V^{\pi}(D))$

$\Rightarrow V^{\pi}(M) = \frac{1}{0.3(1-0.7)}$, $V^{\pi}(R) = \frac{2}{1-0.7}$, $V^{\pi}(D) = \frac{-1}{1-0.7}$

ب)

$\gamma = 0.9 \Rightarrow V^{\pi}(M) = 5$, $V^{\pi}(R) = 20$, $V^{\pi}(D) = -10$

$\Rightarrow Q^{\pi}(M, P) = \underline{5}$, $Q^{\pi}(R, P) = \underline{20}$, $Q^{\pi}(D, P) = \underline{-10}$

$Q^{\pi}(M, W) = 0.1(1 + 0.9\,V^{\pi}(M)) + 0.3(-1 + 0.9\,V^{\pi}(D)) + 0.6(2 + 0.9\,V^{\pi}(R))$

$= \underline{12.55}$

طرف: $Q(R, W) = \underline{18.4}$, $Q(D, W) = \underline{5.5}$

$\Rightarrow V^{\pi_1}(M) = war$, $V^{\pi_1}(R) = peace$, $V^{\pi_1}(D) = war$

ج)

$1 \to 0 + 0.5(-2 + 1(0-0)) = -1$

$2 \to -1 + 0.5(3 + 1(0)) = 1$

$3 \to 0 + 0.5(1 + 1(0-0)) = 0.5$

$4 \to 0 + 0.5(1 + 1(0.5-0)) = 0.75$

$\Rightarrow$

M-P	R-P	D-W
0	0	0
0	0	-1
0	0	1
0	0.5	1
0.75	0.5	1