

سپهر مقیسه

۹۸۳۱۱۰۳

تکلیف سوم

۱-الف) مقدار مورد نظر شروع در گره S و به صورت بهینه حرکت کردن

ب) در گره S بهینه ترین حرکت را انتخاب کنیم

ج) بله میتوان با فرمول زیر:

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

میتوان به POLICY هر S با کمک مقدار ها و پاداش های STATE های قبلی رسید.

۲-الف) درست- چرا که نیاز به یک policy داریم که در هر لحظه تصمیم درست را لحاظ کنی

ب) درست- چرا که حضور مقدار v^* افزایش پیدا میکند

ج) نادرست- به هزینه transition نیز بستگی دارد

د) بله چرا که $\text{policy} = \max \arg q$ است و پالیسی در هر state بهینه ترین حرکت را میدهد.

-۳

Subject:
Date:

$$h=1: \begin{cases} V(s_1) = \max [1, 1/2 + 1/1] = 3/2 \text{ (الف)} \\ V(s_2) = \max [1, 1] = 1 \\ V(s_3) = \max [1, 0] = 1 \end{cases}$$

$$h=2: \begin{cases} V(s_1) = \max \left[(1 + 0.5 \times 1) \text{ و } (0.5 \times (3/2 + 1)) + (0.5 \times (1 + 0.5 \times 1)) \right] \\ \quad = \max [1.5 \text{ و } 1.25] = 1.5 \\ V(s_2) = \max [1 \text{ و } (0.5 \times 1 + 0.5 \times 1)] = 1 \\ V(s_3) = \max [1 \text{ و } (0.5 \times 1 + 0.5 \times 0)] = 1 \end{cases}$$

$$h=3: \begin{cases} V(s_1) = \max \left[(1 + 0.5 \times 1.5) \text{ و } (0.5 \times (1.5 + 0.5 \times 1.5)) + (0.5 \times (1 + 0.5 \times 1.5)) \right] \\ \quad = \max [1.75 \text{ و } 1.625] = 1.75 \\ V(s_2) = \max [1 \text{ و } (0.5 \times 1 + 0.5 \times 1)] = 1 \\ V(s_3) = \max [1 \text{ و } (0.5 \times 1 + 0.5 \times 0)] = 1 \end{cases}$$

(ب) باید می شود چرا که از یک جا به بعد دیگر تغییر نمی کند

$$\begin{matrix} s_1 & \xrightarrow{0.5} & s_2 & \xrightarrow{0.5} & s_3 \\ V(s_1) = 1.75 & & V(s_2) = 1 & & V(s_3) = 1 \end{matrix} \quad \text{(ج)}$$

۴- الف) ۱- تربیت گربه در خانه

محیط: خانه

عامل: گربه

۲- app recommendation : محیط : برنامه

عامل: استفاده کننده برنامه

۳- بازی شطرنج : محیط : خانه های شطرنج

عامل : بازیکن ها

ب) ۱- model free - چرا که هر گربه ممکن است شخصیت متفاوتی داشته باشد با توجه به تجربه های مربی های دیگر گربه ها

۲-model based - چرا که سلايق افرادی که ویدیو های مشترک زیادی میبینند تقریباً یکسان است و میتوان بر اساس تجربیات پیشنهاد شده به یک فرد همان هارا به دیگری پیشنهاد داد

۳-model free : چرا که شیوه بازی هر فرد متفاوت است و نمیتوان بر اساس شناخت قبلی عمل کرد

ج) ۱- استفاده از جایزه جدید برای تربیت گربه و یا استفاده از همان جایزه قبلی

۲- پیشنهاد کانال های جدید تر به فرد یا پیشنهاد همان کانال های قبلی

۳- بازی با استفاده از شیوه قبلی و سعی در کیش کردن و یا امتحان شیوه جدید مانند خوردن بعضی مهره های طرف در اول

-۵

$$\begin{aligned}
 V(A_1) &= (-0.1 \times 1) + (0.2 \times -0.1) + (1 \times -0.1) + (10 \times -0.1) \\
 &\quad (9 \times -0.1) + (10 \times -0.1) \\
 &= -0.51 \\
 V(A_2) &= (1 \times -0.1) + (9 \times -0.1) + (9 \times -0.1) + (10 \times -0.1) + (9 \times -0.1) \\
 &\quad 1 \quad 9 \\
 &= -0.25 \\
 V(A_3) &= (2 \times -0.1) + (2 \times -0.1) + (7 \times -0.1) + (-0.1) + (10 \times -0.1) \\
 &\quad 1 \quad 2 \quad 7 \quad 1 \quad 10
 \end{aligned}$$



Subject: _____
Date: _____

$$V(A_1) = \frac{-0/1 - 0/1}{2} = -0/1$$

$$V(B_1) = \frac{-0/1}{1} = -0/1$$

$$V(B_2) = 0$$

$$V(B_3) = \frac{-0/1 + (6 \times -0/1) + (1 \times 0/1)}{8} = -0/15$$

$$V(C_1) = \frac{(6 \times -0/1)}{6} = -0/1$$

$$V(C_2) = \frac{-0/1}{2} = -0/2 \quad V(C_3) = \frac{-0/1 + -0/1}{2} = -0/1$$

۶- در off policy - policy learner بدون توجه به action ای که عامل انجام میدهد سیاست بهینه را یاد میگیرد- SARSA (state-action-reward-state-action) - در این الگوریتم سیاستی که برای اپدیت ارزش هر حالت به کار میرود مانند سیاستی هست که برای عمل حرکت انجام میشود و این با q learning که این دو مقدار یکسان نیست تفاوت دارد.

۷- الف) فاصله خودرو تا چراغ قرمز - زیرا محل عبور عابرین پیاده در زیر چراغ قرمز واقع شده است.

فاصله آن محل از محل های پیاده رو - چرا که هرچه فاصله با محل های پیاده رو کمتر باشد احتمال وجود عابرین پیاده بیشتر است .

فاصله تا پل هوایی - هرچه فاصله بیشتر باشد احتمال رد شدن محل پیاده از وسط خیابان بیشتر است هر کدام از این ویژگی ها وزن خاص خود را دارند

ب) محل پشت چراغ قرمز و محل های دورتر از چراغ