باسمه تعالی

تکلیف ۳ درس هوش مصنوعی – دکتر جوانمردی

سید امیرمهدی میرشریفی - ۹۸۳۱۱۰۵

سوال ۱ (۲۰ نمره)

صحیح یا غلط بودن موارد زیر را با ذکر دلیل بیان کنید.

۱- در فرایند تصمیم گیری مارکوف نتیجه هر عمل به حالت کنونی و قبلی وابسته است.

۲- در direct evaluation بعد از انجام هر انتقال، ارزش آن حالت را دوباره محاسبه می‌کنیم.

۳- اگر تنها تفاوت بین دو MDP، مقدار discount factor باشد، این دو قطعا سیاست بهینه (optimal policy) یکسان دارند.

۴- Q-learning یک شیوه یادگیری تقویتی model-free و off-policy است.

پاسخ:

۱- غلط - در تصمیم گیری مارکوف نتیجه هر عمل ما صرفا به حالت کنونی وابسته است.

۲- غلط- اگر مقصود جمله این است که بعد از هر نمونه نتیجه را به روز رسانی میکنیم ، جواب غلط است در این روش ما در تمام بخش های نمونه مان ، دنباله هایی که شروعشان state مدنظرمان هست را به عنوان یک نمونه در نظر میگیریم و ارزش آن را با میانگیری آن ها حساب می کنیم.

۳- غلط – در مسئله خوردن دات ها این قضیه نقض خواهد شد.

۴- درست

سوال ۲) (۱۰نمره)

فرض کنید میخواهیم از روش یادگیری Q تخمینی در یک هلیکوپتر کوچک برای خودداری از برخورد با درختان و ساختمان ها استفاده کنیم. ابتدا مشخص کنید از کدام یک از ویژگی های محیط را برای تابع ارزش خطی استفاده می کنید و سپس دو حالتی که بر اساس ویژگی های گفته شده مشابه هستند اما ارزش بسیار متفاوتی دارند را بیان کنید.

پاسخ:

ویژگی ها همچون ، ارتفاع ، فاصله تا درخت ، سرعت ، جهت ، تراکم درختان اطراف و ....
برای ساخت تابع ارزش خطی میتوان از پارامتر های سرعت(عامل خنثی یا مثبت) ، پراکندگی (عامل منفی) و فاصله تا درخت(عامل منفی) استفاده کرد. اگر هلی کوپتر در حالتی باشد که پراکندگی محیط زیاد است اما فاصله اش تا نزدیک ترین درخت متوسط باشد ارزش آن برابر خواهد شد با حالتی که پراکندگی کم است اما فاصله اش تا نزدیک ترین درخت بسیار بسیار کم باشد. به این صورت شرایط متفاوت اما ارزش برآورد شده شان یکی خواهد بود.

سوال۳) (۱۰ نمره)

الف) اگر در طول اجرای policy iteration، فقط یک iteration از policy evaluation را به جای اجرای آن تا زمان همگرایی انجام دهیم، آیا همچنان به سیاست بهینه می‌رسیم؟ توضیح دهید.

ب) فرض کنید در Q-learning، مقدار $\varepsilon=1$ باشد. در این صورت آیا تضمینی وجود دارد که به سیاست بهینه همگرا شود؟ توضیح دهید.

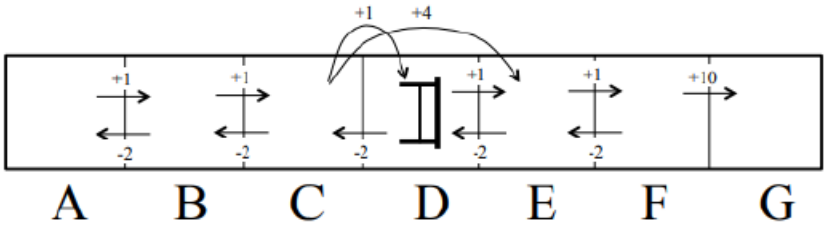
پاسخ:

الف) وقتی یک دور را بخواهیم مدام اجرا کنیم یعنی با یک سیاست یکسان هر دفعه مقادیر حالات را حساب کنیم ، بنابراین بعد از دور اول که مقادیر با سیاست ها همگرا شد سراغ دور دوم میرویم با همان سیاست ها که طبق همگرایی مرحله آخر مجددا همگرا است بنابراین لزوما بهینه نیست.

ب) وقتی اپسیلون برابر یک باشد یعنی همیشه رندوم عمل کند و در نهایت که همگرا میشود ، بهینه هم هست

سوال۴) (۲۰ نمره)

مدل MDP شکل زیر را در نظر بگیرید. در خانه D یک مانع وجود دارد و عامل ما می‌تواند به صورت deterministic به چپ (left) یا راست (right) حرکت کند. خانه G نیز حالت ترمینال است. اگر عامل در خانه C باشد، نمی‌تواند به راست حرکت کند و باید بپرد (jump). پرش ممکن است موفقیت‌آمیز باشد و به خانه E برسد و $T(C, jump, E) = 0.5$. همچنین ممکن است شکست خورده و به مانع خانه D برخورد کند و $T(C, jump, D) = 0.5$. به سوالات زیر با فرض $\gamma = 1$ پاسخ دهید.



الف) دو مرحله value iteration را انجام دهید و مقادیر زیر را محاسبه کرده و جدول را کامل کنید. Iteration صفر برای تخصیص مقدار اولیه صفر برای همه مقادیر است.

1) $V_2(B)$

2) $Q_2(B, right)$

3) $Q_2(B, left)$

پاسخ:

In v1:

Va= vb +1= 0+1= 1

Vb= max ((vc +1) , va -2) = max(1,-1)= 1

Vc= max ((vd +1)/2+(ve +4)/2, -2+1) = max(5/2,2,-1)=5/ 2

Vd= max(vc -2 , ve +1)= max(0/5, 1)=1

Ve= max ((vf +1) , vd -2) = max(1,-2)= 1

Vf= max ((vg +10) , ve -2) = max(10,-1)= 10

In v2:

Va= vb +1= 1+1= 2

Vb= max ((vc +1) , va -2) = max(3.5,0)= 3.5

Vc= max ((vd +1)/2+(ve +4)/2, -2+3.5) = max(7/2,2,1.5)=7/ 2

Vd= max(vc -2 , ve +1)= max(1.5, 2)=2

Ve= max ((vf +1) , vd -2) = max(11,0)= 11

Vf= max ((vg +10) , ve -2) = max(10,-9)= 10 => V(B)=3.5 / Q(B,left)=0 / Q(B,right) =3.5

	A	B	C	D	E	F	G
V_1	1	1	5/2	1	1	10	0
V_2	2	3.5	7/2	2	11	10	0

ب) با توجه به اپیزود زیر و چهار انتقال انجام شده، به روز رسانی Q-learning را اعمال کنید و Q-valueهای به دست آمده را در خانه‌های جدول پر کنید. خانه‌هایی که تحت تاثیر قرار نمی‌گیرند را خالی بگذارید.

Episode

s	a	r	s	a	r	s	a	r	s	a	r	s
C	*jump*	+4	E	*right*	+1	F	*left*	-2	E	*right*	+1	F

بر فرض آن که آلفا برابر نیم باشد:

	Q(C, left)	Q(C, jump)	Q(E, left)	Q(E, right)	Q(F, left)	Q(F, right)
Initial	0	0	0	0	0	0
Transition 1	۰	۲	۰	۰	۰	۰
Transition 2	۰	۲	۰	/۵	۰	۰
Transition 3	۰	۲	۰	/۵	-۰,۷۵	۰
Transition 4	۰	۲	۰	/۷۵	-۱	۰

سوال۵) (۲۰ نمره)

فرض کنید یک عامل هوشمند داریم که میتواند درس بخواند! به جهت سادگی، این عامل فقط دارای حالت وضعیت درسی {High, Low} میباشد. در هر کدام از این حالت ها، عامل میتواند یکی از عمل های زیر را انجام دهد:

- درس بخواند
- امتحان بدهد
- به تماشای Netflix بنشیند.

این عامل را می توانیم با MDP زیر مدل کنیم:

S	A	S'	T(S, A, S')	R (S, A, S')
High	study	High	1.0	0
High	study	Low	0.0	0
High	exam	High	0.9	10
High	exam	Low	0.1	-10
High	Netflix	High	1.0	1
High	Netflix	Low	0.0	1
Low	study	High	0.3	0
Low	study	Low	0.7	0
Low	exam	High	0.05	10
Low	exam	Low	0.95	-10
Low	Netflix	High	0.0	1
Low	Netflix	Low	1.0	1

Policy iteration را برای این MDP تا دو iteration اعمال کنید. سیاست اولیه را عمل study برای هر حالت در نظر بگیرید. Utility اولیه هر دو حالت را برابر با صفر در نظر بگیرید. مقدار $\gamma = 0.5$ است. راه حل خود را شرح دهید. آیا در انتها سیاست ها همگرا می‌شوند؟ توضیح دهید.

پاسخ:

Π_0=study and v1 =0 and v2 =0

V1 = 1*(0+0) + 0*(0 + 0)=0

V2 = .3*(0+0) + 0*(0+0) =0

⇨ Now check Q for Each V :

Q(V1,study) = 0

Q(V1,exam)=.9*(10+0) + .1*(-10+0)=8 => best action = exam

Q(V1,Netflix)=1+0=1

Q(V2,study) = 0

Q(V2,exam)=.05*(10+0) + .95*(-10+0)=-9 => best action = netflix

Q(V1,Netflix)=1+0=1

⇨ Π_1 = exam, netflix

V1 = 8

V2 = 1+0=1

⇨ Now check Q for Each V :

Q(V1,study) = 1*(0 +8/2)=4

Q(V1,exam)=.9*(10+8/2) + .1*(-10+1/2)= 12.6 -.95 = 11.65 => best action=exam

Q(V1,Netflix)=1+8/2=5

Q(V2,study) = .3*(8/2) + .7(1/2)=1.2 + .35 = -1.55

Q(V2,exam)=.05*(10+8/2) + .95*(-10+1/2)=.7 – 9.025=-8.325 => best action= study

Q(V1,Netflix)=1-1/2=-.5

⇨ Π2= for v1 = exam & for v2 = study

برای حل این مسئله همانطور که در بالا نوشته شده است اول با سیاست این که هر دو حالت درس جلو میرویم. ارزش حالت ها را به دست می آوریم. سپس بهترین اکشن را به عنوان سیاست بعدی پیدا میکنیم و همین مراحل را دو بار انجام میدهیم تا در نهایت برای حالت یک سیاست آزمون و حالت دو سیاست درس انتخاب میشود. برای این که بفهمیم همگرا شده است یا نه نیاز به یک مرحله دیگر دارد زیرا تا آخرین مرحله همگرا نبود.

سوال ۶) (۲۰ نمره)

یک مسئله MDP را در نظر بگیرید که در آن سه حالت [A,B,C] و دو اکشن حرکت (Go) و توقف (Stop) وجود دارد. نمونه‌های جدول زیر از انجام اکشن‌های مختلف در این MDP تولید شده‌اند. همچنین فرض کنید $\gamma = 1$ و $\alpha = 0.5$. به سوالات زیر پاسخ دهید.

الف) Q-learning را روی نمونه‌های زیر اجرا کنید. با فرض مقدار اولیه Q-value صفر، مقادیر Q-value زیر که از Q-learning به دست آمده‌اند را محاسبه کنید.

s	a	s'	r
A	Go	B	2
C	Stop	A	0
B	Stop	A	-2
B	Go	C	-6
C	Go	A	2
A	Go	A	-2

1) Q (C, Stop)

2) Q (C, Go)

پاسخ:

First (A,Go,B,2) => Q(A,Go) += .5 * (2 + max Q(B) – Q(A,Go)) => Q(A,GO)= 1

Then (C,stop,A,0) => Q(C,stop) += .5*(0 + maxQ(A) – Q(C,stop)) => Q(C,stop)=0.5

Then (B,stop,A,-2) => Q(B,stop) += .5* (-2 + maxQ(A) - Q(B,stop))=> Q(B,stop) =-0.5

Then (B,Go,C,-6) => Q(B,Go)+=0.5*(-6 + maxQ(c)- Q(B,Go)) => Q(B,Go)=-0.275

Then (C ,GO ,A ,2)=> Q(C,Go)+=0.5*(2+maxQ(A)- Q(C,Go))=> Q(C,Go)=1.5

Finally (A,GO,A,-2)=> Q(A,Go) += 0.5*(-2 + maxQ(A)- Q(A,Go)) => Q(A,Go)=0

ب) در این بخش از دو ویژگی داده شده زیر استفاده کنید.

- $f_1(s,a) = 1$
- $f_2(s,a) = \begin{cases} 1 & a = \text{Go} \\ -1 & a = \text{Stop} \end{cases}$

وزن‌های w_1 وw_2 را با توجه به نمونه‌های مشاهده شده پس از اولین به روزرسانی (با استفاده از نمونه اول) و دومین بروزرسانی (با استفاده از نمونه دوم) بنویسید. وزن‌های اولیه صفر هستند.

s	a	s'	r
A	Go	B	4
B	Stop	A	0

پاسخ:

با توجه به آن که آلفا بیان نشده است مقدار آن را نیم در نظر میگیریم:

Q(S,A) = w1*f1(s,a) + w2*f2(s,a)

W1=w2=0

First: Q(A,Go) = 0 + 0 = 0

R + maxQ(B) − Q(A,Go) = 4 + 0 − 0=0 -> w1 = w1 + alpha*difference*fi = 0 + .5* 4 *1 =2

W2 = w2 + alpha*difference*f2 = 0 + .5*1 = 2

W1 = W2 = 2

Then:

Q(B,Stop) = 2* 1 + 2*-1= 0

Difference = 0 + maxQ(A) − Q(B,stop) = 0 + 3 − 0 = 3 => w1 = 2 + 0.5*3*1 = 3.5

W2= 2 + .5*3 * (-1) = 0.5

W1 = 3.5 W2 = 0.5