hw3 4710

jh4ctf

October 2021

1 Problem 1

1.1 a

States:healthy,sick
actions:candy,vegetables
T(healthy,vegetables,healthy)=1
T(healthy,candy,healthy)=1/4
T(healthy,candy,sick)=3/4
T(sick,vegetables,healthy)=1/4
T(sick,vegetables,sick)=3/4
T(sick,candy,sick)=7/8
T(sick,candy,toothless)=1/8
R(vegetables)=4 R(candy)=10

1.2 b

π_1:
V(healthy)=max*T(healthy,candy,healthy)*[10+γ*V(healthy)]
V(healthy)=max*T(healthy,candy,sick)*[10+γ*V(sick)]
V(sick)=max*T(sick,candy,sick)*[10+γ*V(sick)]
V(sick)=max*T(sick,candy,toothless)*[10+γ*V(toothless)]

π_2:
V(healthy)=max*T(healthy,vegetables,healthy)*[4+γ*V(healthy)]
V(sick)=max*T(sick,vegetables,healthy)*[4+γ*V(healthy)]
V(sick)=max*T(sick,vegetables,sick)*[4+γ*V(sick)]

1.3 c

$V_0(healthy)$ =0
$V_0(sick)$ =0

$V_1(healthy)$ =T(healthy,candy,healthy)*[10+γ*V_0(healthy)]

$V_1(sick)$ = T(healthy,candy,sick)*[10+γ*V_0(sick)]

$V_1(healthy)$ = 3/4*(10+0.9*0)=2.5
$V_1(sick)$ =3/4*(10+0.9*0) = 7.5

$V_2(healthy)$ = T(healthy,candy,healthy)*[10+γ*V_1(healthy)]
$V_2(sick)$ = T(sick,candy,sick)*[10+γ*V_1(sick)]+T(healthy,candy,sick)*[10+γ*V_1(healthy)]

$V_2(healthy)$ = 1/4*(10+0.9*2.5)=2.78
$V_2(sick)$ =7/8*(10+0.9*7.5)+3/4*(10+0.9*2.5=14.6+9.18=23.78

1.4 d

i. True because the policy will always converge before value iteration does. When value converges, the policy step before provides policy that help it converge.
v. There can be no more than one optimal value function but there can be more than one optimal policy to lead to that one optimal value function's result.

2 Problem 2

2.1 a

$V^{\pi}(s)$=(1-α)[(1-α)$V^{\pi}(s)$+αR(s,$\pi(s), s') + \gamma V^{\pi}(s')$]+$\alpha$R(s',$\pi(s'), s'') + \gamma V^{\pi}$(s")

2.2 b

i.

Since Q(S3,A2) = 9.075, we take Q(s3,A2) to get to S5. Then we take A1 to S6 by Q(S5,A1) = 5.76. Then we take A3 from S6 to S5 because Q(S6,A3)=3.075. Then we repeat (S5,A1) as it is the only possible way to traverse. It will turn into a cycle as it cycles back forth between S5 and S6.

The pairs are following:

(S3,A2),(S5,A1),(S6,A3),(S5,A1)

ii.

Because the best Q value only emphasizes the exploration of a problem. Hence we can improve it by emphasizing more of exploitation by implementing the ϵ-greedy function to better sort the best case for the problem.

2.3 c

i.

Because:
$Q^*(S_2, A_1) = 6$
$Q^*(S_2, A_2) = 8$
$Q^*(S_3, A_1) = 16$
In the formula:$Q^*(S_1, A_1) = 0.5 * maxQ(S_2, a) + 0.5 * maxQ(S_3, a)$
We can plug in the value to get:
Thus, $Q^*(S_1, A_1) = 0.5 * 8 + 0.5 * 16 = 12$

ii.

$\pi(S_1) = A_2$, $\pi(S_2) = A_2$, $\pi(S_3) = A_1$

iii.

$[(S1, A1, S2) - (S2, A2, S4)]$: $Q = 6$
$[(S1, A2, S3) - (S3, A1, S5)]$: $Q = 10$
$[(S1, A2, S2) - (S2, A1, S4)]$: $Q = 8.5$

3 Problem 3

For ants to do local beam search, they will start from their home to to go out to search for good and they go find k(number of locations that ants go out to find randomly) number of locations that contains food. When they reach to each location, they go around each location to see if they have found exactly what they want. If not, then they go through all k locations' surroundings and decide between all of the surroundings around all k locations to be their wanted food place.