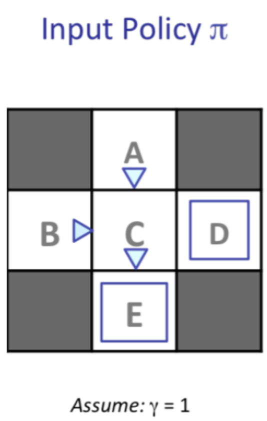
Homework 4 Electronic

June 17th, 2020 at 11:59pm

1 Model-Based RL:Grid



Assuming we have four observed episodes for training,

1. Episode 1:A,south,C,-1;C,south,E,-1;E,exit,x,+10
2. Episode 2:B,east,C,-1;C,south,D,-1;D,exit,x,-10
3. Episode 3:B,east,C,-1;C,south,D,-1;D,exit,x,-10
4. Episode 4:A,south,C,-1;C,south,E,-1;E,exit,x,+10

What model would be learned from the above observed episodes?

1 T(A, south , C) = 1 2/2

2 T(B, east , C) = 1 2/2

3 T(C, south , E) = 0.5 2/4 =0.5

4 T(C, south , D) = 0.5 2/4 = 0.5

(Your answer should be 1,0.5,0.25,0.35 for example)

2 Direct Evaluation

Consider the situations in problem 1, what are the estimates for the following quantities as obtained by direct evaluation:

$\widehat{V}^{\pi}(A) =$ 8
$\widehat{V}^{\pi}(B) =$ -12
$\widehat{V}^{\pi}(C) =$ -1
$\widehat{V}^{\pi}(D) =$ -10
$\widehat{V}^{\pi}(E) =$ 10

(Your answer should be 1,-1,0,0,5 for example)

(i) $\hat{V}^{\pi}(A)$ Ep 1 → $-1-1+10=8$, Ep 4 → $-1-1+10=8$ ⇒ $(8+8)/2=8$

(ii) $\hat{V}^{\pi}(B)$ Ep 2 → $-1-1-10=-12$, Ep 3 → $-1-1-10=-12$ ⇒ $(-12-12)/2=-12$

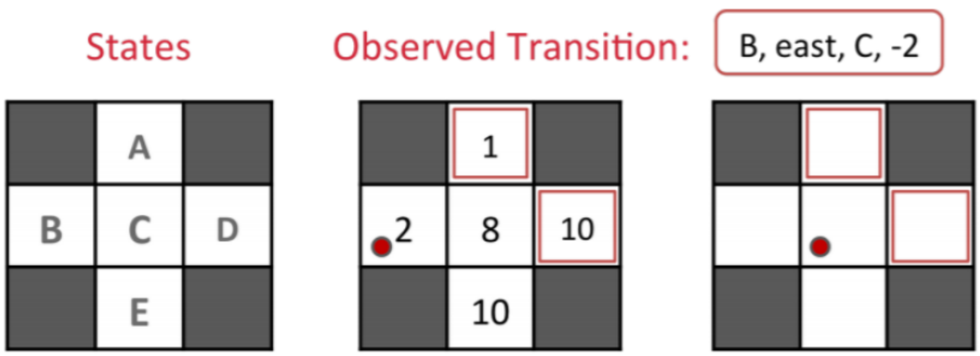
(iii) $\hat{V}^{\pi}(C)$ Ep 1 → $-1+10=9$
Ep 2 → $-1-10=-11$
Ep 3 → $-1+10=9$
Ep 4 → $-1-10=-11$
$\}$ → $(9-11+9-11)/4=-1$

(iv) $\hat{V}^{\pi}(D)$ Ep 2 → -10
Ep 3 → -10
$\}$ → $(-10-10)/2=-10$

(v) $\hat{V}^{\pi}(E)$ Ep 1 → 10
Ep 4 → 10
$\}$ → $(10+10)/2=10$

3 Temporal Difference Learning

Consider the gridworld shown below. The left panel shows the name of each state A through E. The middle panel shows the current estimate of the value function V^π for each state. A transition is observed, that takes the agent from state B through taking action east into state C, and the agent receives a reward of -2. Assuming $\gamma = 1, \alpha = \frac{1}{2}$, what are the value estimates after the TD learning update? (note: the value will change for one of the states only)



Assume: γ = 1, α = 1/2 $\quad V^\pi(s) \leftarrow (1-\alpha)V^\pi(s) + \alpha\left[R(s,\pi(s),s') + \gamma V^\pi(s')\right]$

Only B changes

$\hat{V}^\pi(A) =$ 1
$\hat{V}^\pi(B) =$ 4
$\hat{V}^\pi(C) =$ 8
$\hat{V}^\pi(D) =$ 10
$\hat{V}^\pi(E) =$ 10

$$\hat{V}^\pi(B) = \frac{1}{2}\times 2 + \frac{1}{2}[-2 + 1\cdot 8]$$
$$= 1 + \frac{1}{2}\cdot 6 = 4$$

(Your answers should be 1,-1,0,0,5 for example)

4 Model-Free RL:Cycle

We recommend you work out the solutions to the following questions on a sheet of scratch paper, and then enter your results into the answer boxes.

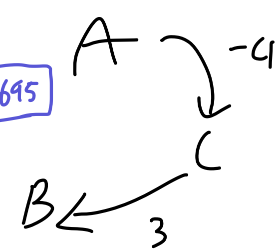
Consider an MDP with 3 states, A, B and C; and 2 actions Clockwise and Counterclockwise. We do not know the transition function or the reward function for the MDP, but instead, we are given with samples of what an agent actually experiences when it interacts with the environment (although, we do know that we do not remain in the same state after taking an action). In this problem, instead of first estimating the transition and reward functions, we will directly estimate the Q function using Q-learning.

Assume, the discount factor, γ is 0.5 and the step size for Q-learning, α is 0.5.

Our current Q function, $Q(s, a)$, is as follows.

$\boxed{-1.68}$ $\boxed{2.9695}$

	A ①	B	C ②
Clockwise	$\boxed{-0.93}$	1.24	$\boxed{0.439}$
Counterclockwise	-5.178	5	3.14



The agent encounters the following samples,

	s	a	s'	r
①	A	clockwise	C	-4
②	C	clockwise	~~D~~ B	3

Process the sample given above. Fill in the Q-values after both samples have been accounted for.

Q(A,clockwise)=____
Q(B,clockwise)=____
Q(C,clockwise)=____
Q(A,counterclockwise)=____
Q(B,counterclockwise)=____
Q(C,counterclockwise)=____

(You answer should be 1,-1,0,0,5,6 for example)

$Q(s,a) \leftarrow (1-\alpha)\,Q(s,a) + \alpha\,[r + \gamma \max_{a'} Q(s',a')]$

① = $Q(A, \text{clockwise}) = \frac{1}{2}(-0.93) + \frac{1}{2}[-4 + \frac{1}{2}\boxed{\max_{a'}(C, a')}]$

↳ (C, clockwise) 0.439 < (C, Counter C...) 3.14 ✓

$= \frac{1}{2}(-0.93) + \frac{1}{2}[-4 + \frac{1}{2}(3.14)] = -1.68$

② $Q(C, \text{clockwise}) = \frac{1}{2}(0.439) + \frac{1}{2}[3 + \frac{1}{2}\boxed{\max_{a'}(B, a')}]$

↳ (B, clockwise) 1.24 < (B, Counter C...) 5

$= \frac{1}{2}(0.439) + \frac{1}{2}[3 + \frac{1}{2}(5)] = 2.9695$

other Q function remains same, because those did not go through Q-learning

5 Q-Learning Properties

In general, for Q-Learning to converge to the optimal Q-values...

A. It is necessary that every state-action pair is visited infinitely often.

B. It is necessary that the learning rate α (weight given to new samples) is decreased to 0 over time.

C. It is necessary that the discount γ is less than 0.5.

D. It is necessary that actions get chosen according to $\arg\max_{\mathrm{a}} Q(s, a)$.

(You answers should be ABCD for example)

6 Exploration and Exploitation

For each of the following action-selection methods, indicate which option describes it best.

A: With probability p, select $\arg\max_{\mathrm{a}} Q(s, a)$. With probability 1-$p$, select a random action. $p = 0.99$.

A. Mostly exploration

B. Mostly exploitation

C. Mix of both

B: Select action a with probability

$$P(a|s) = \frac{e^{\frac{Q(s,a)}{\tau}}}{\sum_{a'} e^{\frac{Q(s,a')}{\tau}}}$$

where τ is a temperature parameter that is decreased over time.

A. Mostly exploration

B. Mostly exploitation

C. Mix of both

C: Always select a random action.

A. Mostly exploration

B. Mostly exploitation

C. Mix of both

D: Keep track of a count, $K_{s,a}$, for each state-action tuple,(s,a), of the number of times that tuple has been seen and select $\arg\max_{\mathrm{a}} [Q(s, a) - K_{s,a}]$.

A. Mostly exploration

B. Mostly exploitation

C. Mix of both

Which method(s) would be advisable to use when doing Q-Learning? BD

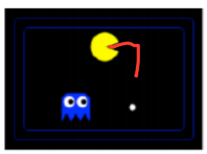
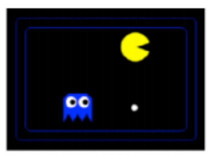
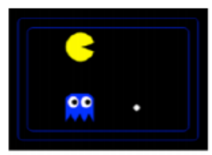
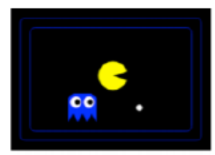
(Your answers should be A,B,C,C,ABCD for example)

7 Feature-Based Representation: Actions

Consider the two Pacman board states presented in two rows below. In each row, the agent considers possible actions to take; these are represented by the images. The agent is using feature-based representation to estimate the $Q(s, a)$ value of taking an action in a state, and the features the agent uses are:

- f_0 = 1/(Manhattan distance to closest food + 1)
- f_1 = 1/(Manhattan distance to closest ghost + 1)

For example, the feature representation $f(s = A, a = \textbf{STOP}) = [1/4, 1/4]$.

State	a=STOP	a=RIGHT	a=LEFT	a=DOWN
A				
$f(s, a)$	[0.25, 0.25]	[1/3, 0.2]	[0.2, 1/3]	[1/3, 1/3]

The agent picks the action according to $\arg\max_a Q(s, a) = w^T f(s, a) = w_0 f_0(s, a) + w_1 f_1(s, a)$, where the features $f_i(s, a)$ are as defined above, and w is a weight vector. Using the weight vector $w = [0.2, 0.5]$, which action, of the ones shown above, would the agent take from state A?

A. STOP $0.2 \cdot 0.25 + 0.5 \cdot 0.25 = 0.175$

B. RIGHT $0.2 \cdot 1/3 + 0.5 \cdot 0.2 = 0.166$

C. LEFT $0.2 \cdot 0.2 + 0.5 \cdot 1/3 = 0.206$

D. DOWN $0.2 \cdot 1/3 + 0.5 \cdot 1/3 = \boxed{0.233}$

max

Using the weight vector w = [0.2,-1], which action, of the ones shown above, would the agent take from state A?

$0.2 \cdot 0.25 + (-1) \cdot 0.25 = -0.2$

A. STOP

$0.2 \cdot 1/3 + (-1) \cdot 0.2 = \boxed{-0.133}$ — max

B. RIGHT

$0.2 \cdot 0.2 + (-1) \cdot 1/3 = -0.293$

$0.2 \cdot 1/3 + (-1) \cdot 1/3 = -0.266$

C. LEFT

D. DOWN

(Your answer should be A,D for example)

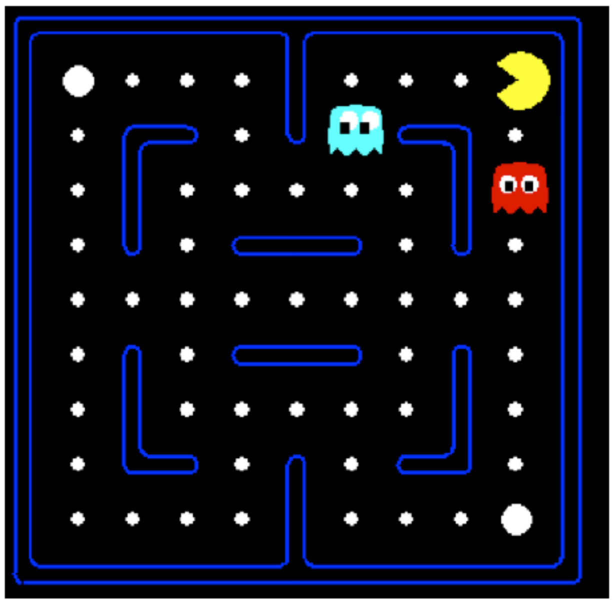
8 Feature-Based Representation: Update

Consider the following feature based representation of the Q-function:

$$Q(s, a) = w_1 f_1(s, a) + w_2 f_2(s, a)$$

with:
$f_1(s, a) = 1/$(Manhattan distance to nearest dot after having executed action a in state s)

$f_2(s, a) =$(Manhattan distance to nearest ghost after having executed action a in state s)
Part 1
Assume $w_1 = 2$, $w_2 = 5$. For the state s shown below, find the following quantities. Assume that the red and blue ghost are both setting on top of a dot.



Q(s,West)=_____ 17 $2 \cdot 1 + 5 \cdot 3$

Q(s,South)=______ 7 $2 \cdot 1 + 5 \cdot 1$

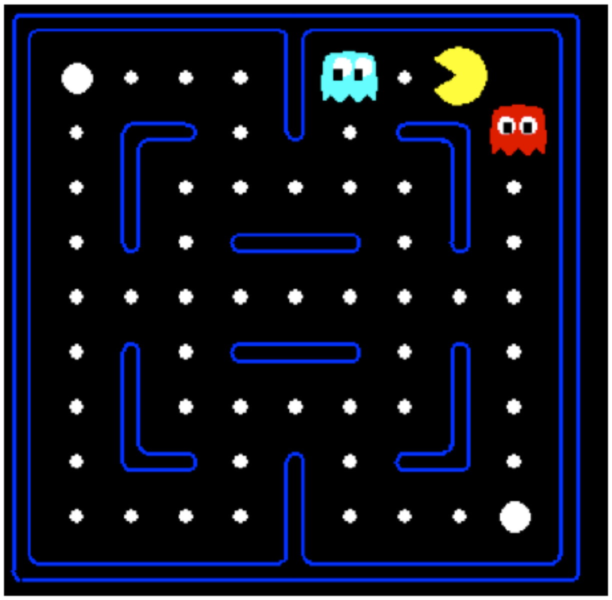
Based on this approximate Q-function, which action would be chosen:

A.West

B.South

Part 2

Assume Pac-Man moves West. This results in the state s' shown below.



The reward for this transition is $r = +10 - 1 = 9$ (+10: for food pellet eating, -1 for time passed). Fill in the following quantities. Assume that the red and blue ghosts are both sitting on top of a dot.

Q(s',West)=7 $2 \cdot 1 + 5 \cdot 1$

Q(s',East)=7 $2 \cdot 1 + 5 \cdot 1$

What is the sample value (assuming γ=1)?

Sample $= [r + \gamma \max_{\mathbf{a}'} Q(s', a')] =$ 16

$9 + 1 \cdot 7$

Part 3

Now let's compute the update to the weights. Let $\alpha = 0.5$

difference $= = [r + \gamma \max_{a'} Q(s', a')] - Q(s, a) =$ -1   16 - 17

$w_1 \leftarrow w_1 + \alpha($ difference $) f_1(s, a) ==$ 1.5 $2 + 0.5(-1) \cdot 1$

$w_2 \leftarrow w_2 + \alpha($ difference $) f_2(s, a) ==$ 3.5 $5 + 0.5(-1) \cdot 3$

(Your answer should be 1,2,A,1,2,3,1,2,3 for example)