CS486: Artificial Intelligence

Homework 6 (15 pts)
Reinforcement Learning
Due 23 October @ 1630

Instructions

This is an individual assignment; however, *you may receive assistance and/or collaborate without penalty, so long as you properly document such assistance and/or collaboration in accordance with DAW.*

Answer the questions below and submit a hardcopy with DAW coversheet and acknowledgment statement to your instructor by the due date.

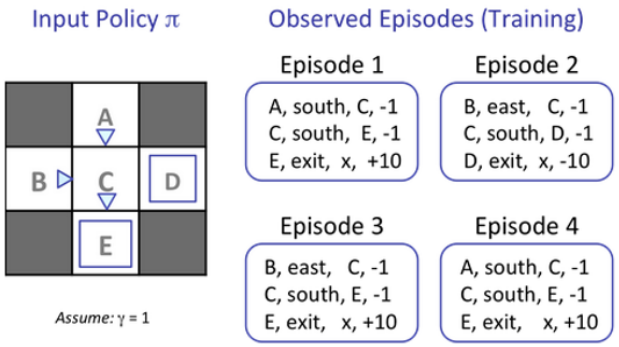


Figure 1: Observed episodes for problems 1 and 2.

Problem 1: Model-Based Learning

What model would be learned from the episodes observed in figure 1?

- $\hat{T}$(A, south, C) = 1
  $T(A, \text{south})$ results in C 2 out of 2 times.
- $\hat{T}$(B, east, C) = 1
  $T(B, \text{east})$ results in C 2 out of 2 times.

- $\hat{T}$(C, south, E) = 0.75
  $T(C, \text{south})$ results in E 3 out of 4 times.
- $\hat{T}$(C, south, D) = 0.25
  $T(C, \text{south})$ results in D 1 out of 4 times.

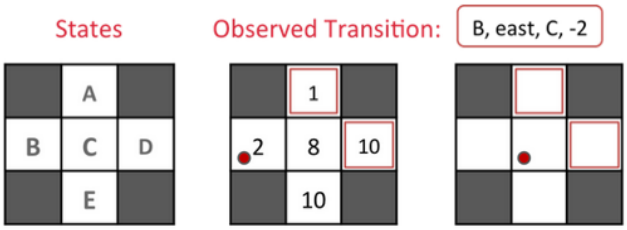
Problem 2: Direct Evaluation

Using figure 1 again, what are the estimates for the following quantities as obtained by direct evaluation?

- $\hat{V}^{\pi}$(A) = 8
  Episode 1 (A to exit): $-1 + -1 + 10 = 8$
  Episode 4 (A to exit): $-1 + -1 + 10 = 8$
  Expected value: $(8 + 8)/2 = 8$
- $\hat{V}^{\pi}$(B) = -2
  Episode 2 (B to exit): $-1 + -1 + -10 = -12$
  Episode 3 (B to exit): $-1 + -1 + 10 = 8$
  Expected value: $(-12 + 8)/2 = -2$
- $\hat{V}^{\pi}$(C) = 4
  Episode 1 (C to exit): $-1 + 10 = 9$
  Episode 2 (C to exit): $-1 + -10 = -11$
  Episode 3 (C to exit): $-1 + 10 = 9$
  Episode 4 (C to exit): $-1 + 10 = 9$
  Expected value: $(9 + -11 + 9 + 9)/4 = 4$
- $\hat{V}^{\pi}$(D) =-10
  Episode 2 (D to exit): -10
  Expected value: $-10/1 = 8$
- $\hat{V}^{\pi}$(E) =10
  Episode 1 (E to exit): 10
  Episode 3 (E to exit): 10
  Episode 4 (E to exit): 10
  Expected value: $(10 + 10 + 10)/3 = 10$

Problem 3: TD Learning

Consider the GridWorld above. The left panel shows the name of each state A through E. The middle panel shows the current estimate of the value function $\hat{V}^{\pi}$ 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 $\gamma = 1, \alpha = \frac{1}{2}$, which state will get an updated value, and what is the new value estimate after the TD learning update?



$$\hat{V}^{\pi}(B) = 4$$

TD learning updates the originating state (B).

$$\begin{aligned}\hat{V}^{\pi}(B) &= (1-\alpha)\hat{V}^{\pi}(B) + \alpha(r + \gamma\hat{V}^{\pi}(C)) \\ &= \frac{1}{2}\cdot 2 + \frac{1}{2}(-2+8) \\ &= 4\end{aligned}$$

Problem 4: Approximate Q-Learning

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

$f_1(s,a) = 1/(\text{Manhattan distance to nearest dot after having executed action } a \text{ in state } s)$
$f_2(s,a) = \text{Manhattan distance to nearest ghost after having executed action } a \text{ in state } s$

a. Assume $w_1 = 1, w_2 = 10$ and that the red and blue ghosts are both sitting on top of a dot. For the state s shown below:

find the following quantities:

- $Q(s, \text{west}) = 31$
  $f_1(s, \text{west}) = 1/1 = 1$
  $f_2(s, \text{west}) = 3$
  $Q(s, \text{west}) = 1 \cdot 1 + 10 \cdot 3 = 31$
- $Q(s, \text{south}) = 11$
  $f_1(s, \text{south}) = 1/1 = 1$
  $f_2(s, \text{south}) = 1$
  $Q(s, \text{south}) = 1 \cdot 1 + 10 \cdot 1 = 11$
- Based on this approximate q-function, which action would be chosen?
  West $(31 > 11)$

b. Assume Pac-man moves west and that the red and blue ghosts are still both sitting on top of a dot. 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). Find the following quantities:

- $Q(s', \text{west}) = 11$
  $f_1(s, \text{west}) = 1/1 = 1$
  $f_2(s, \text{west}) = 1$
  $Q(s, \text{west}) = 1 \cdot 1 + 10 \cdot 1 = 11$
- $Q(s', \text{east}) = 11$
  $f_1(s, \text{east}) = 1/1 = 1$
  $f_2(s, \text{east}) = 1$
  $Q(s, \text{east}) = 1 \cdot 1 + 10 \cdot 1 = 11$
- What is the sample value (assuming $\gamma = 1$)? 20

$$\begin{aligned} \text{sample} &= \left[r + \gamma \max_{a'} Q(s', a') \right] \\ &= 9 + 11 \\ &= 20 \end{aligned}$$

c. Compute the update to the weights. Let $\alpha = 0.5$.

- What is the difference between the received q-value and the expected q-value? -11

$$\begin{aligned}\text{difference} &= \left[r + \gamma \max_{a'} Q\left(s', a'\right)\right] - Q(s, a) \\ &= [9 + 11] - 31 \\ &= -11\end{aligned}$$

- What is the new value for w_1? -4.5

$$\begin{aligned}w_1 &\leftarrow w_1 + \alpha(\text{difference}) f_1(s, a) \\ &= 1 + 0.5(-11) \cdot 1 \\ &= -4.5\end{aligned}$$

- What is the new value for w_2? = -6.5

$$\begin{aligned}w_2 &\leftarrow w_2 + \alpha(\text{difference}) f_2(s, a) \\ &= 10 + 0.5(-11) \cdot 3 \\ &= -6.5\end{aligned}$$