CS 188 Spring 2023 Introduction to Artificial Intelligence

Exam Prep 8 Solutions

Q1. Policy Evaluation

In this question, you will be working in an MDP with states S, actions A, discount factor γ, transition function T, and reward function R.

We have some fixed policy $\pi : S \to A$, which returns an action $a = \pi(s)$ for each state $s \in S$. We want to learn the Q function $Q^\pi(s, a)$ for this policy: the expected discounted reward from taking action a in state s and then continuing to act according to π: $Q^\pi(s, a) = \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma Q^\pi(s', \pi(s'))]$. The policy π will not change while running any of the algorithms below.

(a) Can we guarantee anything about how the values Q^π compare to the values Q^* for an optimal policy π^*?

- (●) $Q^\pi(s, a) \leq Q^*(s, a)$ for all s, a
- ( ) $Q^\pi(s, a) = Q^*(s, a)$ for all s, a
- ( ) $Q^\pi(s, a) \geq Q^*(s, a)$ for all s, a
- ( ) None of the above are guaranteed

(b) Suppose T and R are *unknown*. You will develop sample-based methods to estimate Q^π. You obtain a series of *samples* $(s_1, a_1, r_1), (s_2, a_2, r_2), \ldots (s_T, a_T, r_T)$ from acting according to this policy (where $a_t = \pi(s_t)$, for all t).

(i) Recall the update equation for the Temporal Difference algorithm, performed on each sample in sequence:

$$V(s_t) \leftarrow (1 - \alpha)V(s_t) + \alpha(r_t + \gamma V(s_{t+1}))$$

which approximates the expected discounted reward $V^\pi(s)$ for following policy π from each state s, for a learning rate α.

Fill in the blank below to create a similar update equation which will approximate Q^π using the samples.

You can use any of the terms $Q, s_t, s_{t+1}, a_t, a_{t+1}, r_t, r_{t+1}, \gamma, \alpha, \pi$ in your equation, as well as $\sum$ and max with any index variables (i.e. you could write $\max_a$, or $\sum_a$ and then use a somewhere else), but no other terms.

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha \left[r_t + \gamma Q(s_{t+1}, a_{t+1})\right]$$

(ii) Now, we will approximate Q^π using a linear function: $Q(s, a) = \mathbf{w}^\top \mathbf{f}(s, a)$ for a weight vector $\mathbf{w}$ and feature function $\mathbf{f}(s, a)$.

To decouple this part from the previous part, use Q_{samp} for the value in the blank in part (i) (i.e. $Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha Q_{samp}$).

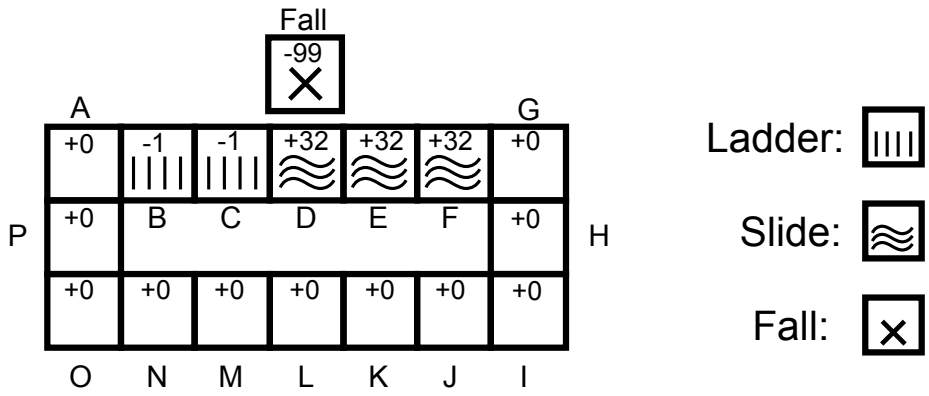
Which of the following is the correct sample-based update for $\mathbf{w}$?

- ( ) $\mathbf{w} \leftarrow \mathbf{w} + \alpha[Q(s_t, a_t) - Q_{samp}]$
- ( ) $\mathbf{w} \leftarrow \mathbf{w} - \alpha[Q(s_t, a_t) - Q_{samp}]$
- ( ) $\mathbf{w} \leftarrow \mathbf{w} + \alpha[Q(s_t, a_t) - Q_{samp}]\mathbf{f}(s_t, a_t)$
- (●) $\mathbf{w} \leftarrow \mathbf{w} - \alpha[Q(s_t, a_t) - Q_{samp}]\mathbf{f}(s_t, a_t)$
- ( ) $\mathbf{w} \leftarrow \mathbf{w} + \alpha[Q(s_t, a_t) - Q_{samp}]\mathbf{w}$
- ( ) $\mathbf{w} \leftarrow \mathbf{w} - \alpha[Q(s_t, a_t) - Q_{samp}]\mathbf{w}$

(iii) The algorithms in the previous parts (part i and ii) are:

- [ ] model-based
- [■] model-free

Q2. RL: Dangerous Water Slide



Suppose now that several years have passed and the water park has not received adequate maintenance. It has become a dangerous water park! Now, each time you choose to move to (or remain at) one of the ladder or slide states (states B-F) there is a chance that, instead of ending up where you intended, you fall off the slide and hurt yourself. The cost of falling is -99 and results in you getting removed from the water park via ambulance. The new MDP is depicted above.

Unfortunately, you don't know how likely you are to fall if you choose to use the slide, and therefore you're not sure whether the fun of the ride outweighs the potential harm. You use reinforcement learning to figure it out!

For the rest of this problem assume $\gamma = 1.0$ (i.e. no future reward discounting). You will use the following two trajectories through the state space to perform your updates. Each trajectory is a sequence of samples, each with the following form: (s, a, s', r).

Trajectory 1: (A, East, B, -1), (B, East, C, -1), (C, East, D, +32)
Trajectory 2: (A, East, B, -1), (B, East, Fall, -99)

(a) What are the values of states A, B, and C after performing temporal difference learning with a learning rate of $\alpha = 0.5$ using only Trajectory 1?

$V(A) =$ -0.5 $V(B) =$ -0.5 $V(C) =$ 16

(b) What are the values of states A, B, and C after performing temporal difference learning with a learning rate of $\alpha = 0.5$ using both Trajectory 1 and Trajectory 2?

$V(A) =$ -1.0 $V(B) =$ -49.75 $V(C) =$ 16

(c) What are the values of states/action pairs (A, South), (A, East), and (B, East) after performing Q-learning with a learning rate of $\alpha = 0.5$ using both Trajectory 1 and Trajectory 2?

$Q(A, \text{South}) =$ 0.0 $Q(A, \text{East}) =$ -0.75 $Q(B, \text{East}) =$ -49.75