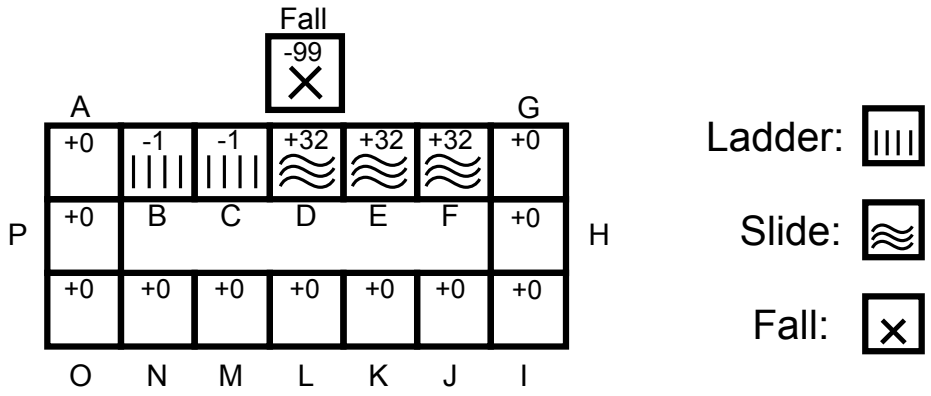
CS 188
Fall 2023

Introduction to
Artificial Intelligence

Exam Prep 6 Solutions

Q1. 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 $\quad V(B) =$ -0.5 $\quad V(C) =$ 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 $\quad V(B) =$ -49.75 $\quad V(C) =$ 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 $\quad Q(A, \text{East}) =$ -0.75 $\quad Q(B, \text{East}) =$ -49.75

Q2. RL: Amusement Park

After the disastrous waterslide experience you decide to go to an amusement park instead. In the previous questions the MDP was based on a single ride (a water slide). Here our MDP is about choosing a ride from a set of many rides.

You start off feeling well, getting positive rewards from rides, some larger than others. However, there is some chance of each ride making you sick. If you continue going on rides while sick there is some chance of becoming well again, but you don't enjoy the rides as much, receiving lower rewards (possibly negative).

You have never been to an amusement park before, so you don't know how much reward you will get from each ride (while well or sick). You also don't know how likely you are to get sick on each ride, or how likely you are to become well again. What you do know about the rides is:

Actions / Rides	Type	Wait	Speed
Big Dipper	Rollercoaster	Long	Fast
Wild Mouse	Rollercoaster	Short	Slow
Hair Raiser	Drop tower	Short	Fast
Moon Ranger	Pendulum	Short	Slow
Leave the Park	Leave	Short	Slow

We will formulate this as an MDP with two states, well and sick. Each ride corresponds to an action. The 'Leave the Park' action ends the current run through the MDP. Taking a ride will lead back to the same state with some probability or take you to the other state. We will use a feature based approximation to the Q-values, defined by the following four features and associated weights:

Features	Initial Weights
$f_0(state, action) = 1$ (this is a bias feature that is always 1)	$w_0 = 1$
$f_1(state, action) = \begin{cases} 1 & \text{if } action \text{ type is Rollercoaster} \\ 0 & \text{otherwise} \end{cases}$	$w_1 = 2$
$f_2(state, action) = \begin{cases} 1 & \text{if } action \text{ wait is Short} \\ 0 & \text{otherwise} \end{cases}$	$w_2 = 1$
$f_3(state, action) = \begin{cases} 1 & \text{if } action \text{ speed is Fast} \\ 0 & \text{otherwise} \end{cases}$	$w_3 = 0.5$

(a) Calculate Q(*'Well'*, *'Big Dipper'*):

$1 + 2 + 0 + 0.5 = 3.5$

(b) Apply a Q-learning update based on the sample (*'Well'*, *'Big Dipper'*, *'Sick'*, -10.5), using a learning rate of $\alpha = 0.5$ and discount of $\gamma = 0.5$. What are the new weights?

Difference $= -10.5 + 0.5 * max(4, 3.5, 2.5, 2.0, 2.0) - 3.5 = -12$
$w_0 = 1 - 6 * 1 = -5$
$w_1 = 2 - 6 * 1 = -4$
$w_2 = 1 - 6 * 0 = 1$
$w_3 = 0.5 - 6 * 1 = -5.5$

(c) Using our approximation, are the Q-values that involve the sick state the same or different from the corresponding Q-values that involve the well state? In other words, is Q(*'Well'*, action) = Q(*'Sick'*, action) for each possible action? Why / Why not? (in just one sentence)

Same
They are the same because we have no features that distinguish between the two states.

Now we will consider the exploration / exploitation tradeoff in this amusement park.

(d) Assume we have the original weights from the table on the previous page. What action will an ϵ-greedy approach choose from the well state? If multiple actions could be chosen, give each action and its probability.

With probability $(1 - \epsilon\frac{4}{5})$ we will choose the Wild Mouse. Each other action will be chosen with probability $\frac{\epsilon}{5}$

(e) When running Q-learning another approach to dealing with this tradeoff is using an exploration function:

$$f(u, n) = u + \frac{k}{n}$$

(i) How is this function used in the Q-learning equations? (a single sentence)
The update replaces the max over Q values with a max over this function (with Q and N as arguments)

What are each of the following? (a single sentence each)

(ii) u :
The utility, given by Q

(iii) n :
The number of times this state-action pair has been visited

(iv) k :
A constant, by adjusting it we can change how optimistic we are about states we haven't visited much.