MSc AI

Reinforcement Learning

Module Code: **B9AI105_2021_TMD3**

Module Description: **Reinforcement Learning**

Examiner: **Amit Sharma**

Internal Moderator: **Terri Hoare**

External Examiner: **Dr Svetlana Hensman**

Date: 25 August 2021
Time: 14:00 – 16:00

INSTRUCTIONS TO CANDIDATES

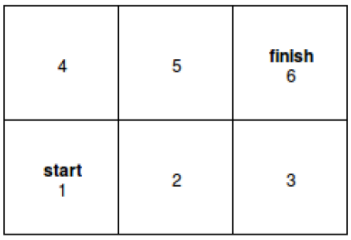
Time allowed is 2 hours
Answer ALL Questions

All answers should reference literature and case studies as appropriate. Use of scientific calculators is permitted.

Question 1 **(35 Marks)**

a)

Assume we are an agent in a 3x2 gridworld, as shown in the below figure. We start at the bottom left node (1) and finish in the top right node (6). When node 6 is reached, we receive a reward of +10 and return to the start for a new episode. On all other actions that not lead to state 6, the reward is -1.



In each state we have four possible actions: up, down, left and right. For each action we move deterministically in the specific direction on the grid. Assume that we cannot take actions that bring us outside the grid.

The current estimates of $Q(s,a)$ are given in the below table:

Q(1,up)=4			Q(1,right)=3
Q(2,up)=6		Q(2,left)=3	Q(2,right)=8
Q(3,up)=9		Q(3,left)=7	
	Q(4,down)=2		Q(4,right)=5
	Q(5,down)=6	Q(5,left)=5	Q(5,right)=8

Since we have full environmental knowledge, we can apply Bellman's equation to further update the Q estimates (i.e., dynamic programming). We take greedy policy. Discount factor = 0.9

$$Q(s,a) = \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma \sum_{a'} \pi(s',a')Q(s',a')]$$

Perform a single update of Q(3,left).

[7]

b) Why were we using Q-values? What is the advantage of learning state-action values (Q) compared to state values (V)? [5]

c) We now no longer assume a model of the environment. The above table was rather created through temporal difference learning, where we sample through the state-space. Why is it not smart to take the greedy policy now (from the start)?

[5]

d) We decide to switch to SoftMax exploration:

$$\pi(s,a) = \frac{e^{Q(s,a)}}{\sum_b e^{Q(s,b)}}$$

We are currently in node 2. Give the probability that we will move right on the next step.

[8]

e) We will continue updating the Q-table with a SARSA (state-action-reward-state-action) algorithm. Starting from node 2, we have sampled the following trajectory: 2 – up - 5 - right - 6, after which the trial ended. Update two Q (s, a) entries for (2, up) and (5, right).

$(\text{take } \alpha = 0.2, \gamma = 0.8)$. The SARSA equation is provided:

$$Q(s,a) = Q(s,a) + \alpha[R^a_{ss'} + \gamma Q(s',a') - Q(s,a)]$$

[10]

Question 2 **(15 Marks)**

Consider following Reinforcement Learning (RL) examples. What are observations, actions and feedback in these examples. How would you setup a reward system for each example.

- Teaching a robot to win a race without going off the tracks [5]
- Teaching a web navigation system to identify human vs bot user through a series of CAPTCHAs. [5]
- Neural Network architecture search [5]

Question 3 **(25 Marks)**

a) Consider a system with two states and two actions. You perform actions and observe the rewards and transitions listed below. Each step lists the current state, reward, action and resulting transition as S_i; R = r; $a_k : S_i \rightarrow S_j$. Perform Q-learning using a learning rate of α = 0.5 and a discount factor of γ = 0.5 for each step. The Q-table entries are initialized to zero.

$$S_1 \quad R = -10 \quad a_1 : S_1 \rightarrow S_1$$

$$S_1 \quad R = -10 \quad a_2 : S_1 \rightarrow S_2$$

$$S_2 \quad R = +20 \quad a_1 : S_2 \rightarrow S_1$$

$$S_1 \quad R = -10 \quad a_2 : S_1 \rightarrow S_2$$

[20]

b) What is the optimal policy at this point?

[5]

Question 4 **(25 Marks)**

You are designing a recycling robot whose job is to collect empty soda cans around the building. The robot has a sensor to detect when a can is in front of it, and a gripper to pick up the can. It also senses the level of its battery. The robot can navigate, as well as pick up a can and throw a can it is holding in the trash. There is a battery charger in the building, and the robot should not run out of battery.

(a) Describe this problem as an MDP. What are the states and actions?

[8]

(b) Suppose that you want the robot to collect as many cans as possible, while not running out of battery. Describe what rewards would enable it to achieve this task

[7]

(c) Instead of thinking about the actions described above, one could describe the task of the robot as choosing between larger activities: walk randomly to find cans, wait for someone to drop a can, or go dock with battery charger. Describe the advantages and disadvantages of this problem formulation compared to the one you gave before.

[10]

END OF EXAMINATION