Assignment 4: Reinforcement Learning
Machine Learning

Deadline: Sunday, 9th of December 2018, Time 21:00

Introduction

In this assignment, you will further deepen your understanding of Reinforcement Learning (RL). Please provide a latex based report in the PDF format. We provide you a sample latex project you might use for writing and generating your report. If latex is new to you, we recommend using overleaf.

Your report, code and all generated files must be archived in a file named firstname.lastname and uploaded to the iCorsi website before the deadline expires. Late submissions will result in 0 points.

Where to get help

We encourage you to use the tutorials to ask questions or to discuss exercises with other students. However, do not look at any report written by others or share your report with others. Violation of that rule will result in 0 points for all students involved. For further questions you can send email to *xingdong@idsia.ch*.

Tasks

Basic probability

1. Suppose that a migrating lizard that rests at Ticino can be in four different states: Eating (E), Sleeping (S), Fighting (F) and Mating (M), for example protecting its territory against other lizards. Each lizard spends 30% of time on sleeping, 40% of time on eating, 20% of time on fighting and rest of time on mating. A biologist collects a population of lizards and puts them in a cage to study their behaviors. Suppose the probabilities of being caught for the lizards that are eating are 0.1, for the lizards that are sleeping are 0.4, for the lizards that are fighting are 0.8 and for the lizards that are mating are 0.2, respectively.

(a) What is the relative frequency of the lizards for being caught in the cage ? (7p)

(b) What is the proportion of lizards that are fighting of those that were caught in the cage ? (3p)

Markov Decision Process (MDP)

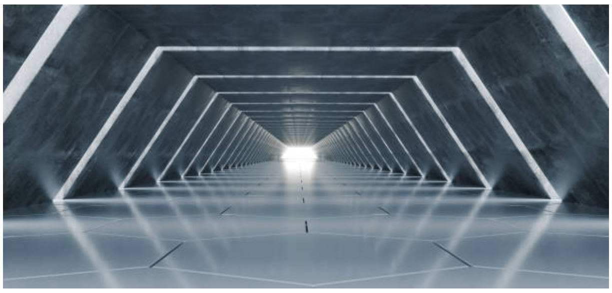

Figure 1: A maze with long corridor.

1. Suppose a robot is put in a maze with long corridor, depicted in Figure 1. The corridor is 1 kilometers long and 5 meters wide. The available actions to the robot are moving forward for 1 meter, moving backward for 1 meter, turning left for 90 degrees and turning right for 90 degrees. If the robot moves and hit the wall, then it will stay in its position and orientation. The goal of the robot is to escape from this maze by reaching the end of the long corridor.

Note: the answers in the following questions should not exceed 5 sentences.

(a) Assume the robot receives +1 for reward signal for each time step taken in the maze and +1000 for reaching the final goal (the end of the long corridor). Then you train the robot for a while, but it seems it still does not perform well at all for navigating to the end of the corridor in the maze. What is happening? Is there something wrong with the reward function? (5p)

(b) If there is something wrong with the reward function, how could you fix it? If not, how to resolve the failure for the training? (5p)

4. The discounted return for the continuing task is defined as:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \tag{1}$$

where $\gamma \in [0, 1]$ is the discounted factor.

(a) Rewrite Equation 1 such that G_t is on the left hand side and G_{t+1} is on the right hand side. (5p)

(b) What is the sufficient condition for this infinite series to be a convergent series ? (5p)

(c) Suppose this infinite series is a convergent series, and each reward in the series is a constant of +1. We know the series is bounded, what is the simple formula for this bound ? Write it down without using summation. (Hint: Geometric series) (5p)

(d) Let the task be an episodic setting and the robot is running for $T = 5$ time steps. Suppose $\gamma = 0.3$, and the robot receives rewards along the way $R_1 = -1, R_2 = -0.5, R_3 = 2, R_4 = 1, R_5 = 6$. What are the values for $G_0, G_1, G_2, G_3, G_4, G_5$? (5p)

(e) Suppose each reward in the series is increased by a constant c, i.e. $R_t \leftarrow R_t + c$. Then how does it change G_t ? (3p)

(f) Now consider episodic tasks, and similar to (e), we add a constant c to each reward, how does it change G_t ? (3p)

Dynamic Programming

1. Write down Bellman optimality equation for state value function without using expectation notation, but using probability distributions instead. (2p)

2. Write down Bellman optimality equation for state-action value function without using expectation notation, but using probability distributions instead. (2p)

3. Consider a 4×4 gridworld depicted in the following table

0	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

The the non-terminal states are $S = \{1, 2, \ldots, 14\}$ and the terminal states are $0, 15$. There are four available actions for each state, that is $A = \{\text{up}, \text{down}, \text{left}, \text{right}\}$. Assume the state transitions are deterministic and all transitions result in a negative reward -1. If the agent hits the boundary, then its state will remain unchanged, e.g. $p(8, -1|s = 8, a = \text{left}) = 1$. **Note: In this exercise, we assume the policy is a deterministic function and it initially maps each state to an arbitrary action.**

(a). Sketch policy iteration algorithm for 3 iterations. Assume the initial state value for all 16 cells are 0.0, in each iteration, write down the equations and detailed numerical computations for the updated values of each cell. Also, write down the greedy action of the improved policy for each state. (25p)

(b). Sketch value iteration algorithm for 3 iterations. Assume the initial state value for all 16 cells are 0.0, in each iteration, write down the equations and detailed numerical computations for the updated values of each cell. (25p)

Good luck !