Final Project: Reinforcement Learning

CS 301 (004), Spring 2021, Introduction to Data Science

Due Date: April 25, 11:59 PM (EST)

WARNING: this project might be hard for some of you: please start as soon as possible!

Remarks. You are expected to write a short essay, which covers in detail your approaches and answers to the below questions. It is highly recommended that you first state your approaches and ideas at a high level and then show how your ideas apply to the two concrete examples as shown here. Your score of this project will be evaluated against both your answers to specific questions and the overall writing skills.

Consider such an interesting game as follows. There is a special die with N sides, where the ith side has the number i for each $1 \leq i \leq N$. Let $[N] \doteq \{1, 2, 3, \ldots, N\}$, the set of integers ranging from 1 to N. Let $\mathbf{p} \in [0, 1]^N$ be a vector of length N such that the ith entry of $\mathbf{p}$, denoted by p_i, represents the probability that we will end with the ith side (thus, we will see the number i) if rolling the die once. For example, $N = 4$ and $\mathbf{p} = (0, 1/2, 1/4, 1/4)$, which means that if we roll the die once, we will see the number 1, 2, 3, and 4, with probability 0, 1/2, 1/4 and 1/4, respectively. There is another binary vector $\mathbf{q} \in \{0, 1\}^N$, where the ith entry of $\mathbf{q}$, denoted by q_i, indicates if the ith side is BAD ($q_i = 1$) or not ($q_i = 0$).

Game Rules. At the beginning, you have \$0 at hand. Suppose at some time, you have $x < K$ dollars at hand, where K is a parameter known in advance. You have two choices to make, either "accept" the challenge or "quit". **(Case 1)** If your choice is "quit", then game is over and you walk away with x dollars. **(Case 2)** If your choice is "accept", then you will roll the die once and see a random number $X \in [N]$ with a probability specified by $\mathbf{p}$. Here are two subcases. (1) If $q_X = 1$, *i.e.,* the Xth side is BAD, then you lose all current money at hand; (2) If $q_X = 0$, *i.e.,* the Xth side is not BAD, then you will get a reward of $f(X)$ where f is a function of X. In this case, you will have $x + f(X)$ dollars. Here is a tricky part: if $x + f(X) \geq K$ (bear in mind that K is a parameter known in advance), then game is over, and you take $x + f(X)$ dollars and go away; otherwise, you will continue the game with $x + f(X)$ dollars at hand. Attention: If you accept the challenge, roll the die, and get X such that $q_X = 1$, you lose all the money at hand but Game is NOT over: you can still continue to play the game with \$0 at hand. Game is over only when you choose to quit or you have at least K dollars at hand. Note that the following key components uniquely define the game: (N, $\mathbf{p}$, $\mathbf{q}$, f, K).

(**Question 1**) Consider a simple case where $N = 6$, $\mathbf{p} = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$. In other words, we have a "normal" die with six sides, and each side will appear with the same chance if we roll once. Let $\mathbf{q} = (1, 1, 0, 1, 0, 0)$, $f(X) = X$, and $K = 100$. You are asked to do the following.

(a) Formulate the above game as a reinforcement learning system. Please specify the key components in the game $(\mathcal{S}, \mathcal{A}, \mathbf{P}, R)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathbf{P}$ is the transition probability matrix, R is the reward function. For simplicity, you can assume the discounted factor $\gamma = 1$. Please specify clearly the terminal state space ($\mathcal{S}_T$) and the non-terminal state space ($\mathcal{S}_N$).

(b) Compute the optimal value function V^* and the optimal policy π^*. You can try either the value iteration method or the dynamic programming method. Please make sure to state explicitly the values of $V^*(s)$ and $\pi^*(s)$ for all $s \in \mathcal{S}_N$, where $\mathcal{S}_N$ refers to the non-terminal state space. Based on your results, state explicitly the maximum expected total rewards you will get in this game when starting with \$0. (If you use the value iteration method, please try different tolerance parameters ϵ to make sure your algorithm converges properly.)

(c) Please try the approach of linear programming (LP) to compute the optimal value function V^* and the optimal policy π^*. You should explicitly specify the following elements in the LP: variables, objective function, and constraints. Again, please state explicitly the values of $V^*(s)$ and $\pi^*(s)$ for all $s \in \mathcal{S}_N$. Based on your results, state explicitly the maximum expected total rewards you will get in this game when starting with \$0.

(**Question 2**) Consider a special case where $N = 5$, $\mathbf{p} = (1/2, 1/4, 1/8, 1/16, 1/16)$, $\mathbf{q} = (1, 0, 0, 0, 1)$, $f(X) = X^2$, and $K = 100$. Answer the same questions (a), (b), and (c), as shown in **Question 1**.