Collab Worksheet 11

CS440/ECE448, Spring 2021

Week of 4/26 - 4/30, 2021

Question 1

ATARA is an Automatic Telephone-based Airplane Reservation Agent.

In order to make an airplane reservation, ATARA needs to learn the user's starting city, ending city, and date of travel (she always asks in that order). When she starts each dialog, she knows none of these things.

During each of her dialog turns, ATARA has the option of asking for 1 or 2 pieces of information. Unfortunately, her speech recognizer makes mistakes. If she asks for 1 piece of information, she always gets it. If she asks for 2 pieces of information, then she gets both pieces of information with probability $\left(\frac{1}{2}\right)$, but with probability $\left(\frac{1}{2}\right)$, she gets nothing.

ATARA receives a reward of $R(s) = 10$, and ends the dialog, when she has correctly recognized all 3 pieces of information. Otherwise, she gets a reward of $R(s) = -1$ for each dialog turn during which she has not finished the dialog.

(a) What is the set of states for this Markov decision process?

(b) What is the set of actions?

(c) Write the transition probability table $P(s'|s, a)$.

(d) Use value iteration to find $U(s)$, the utility of each state, assuming a discount factor of $\gamma = 1$.

Question 2

A cat lives in a two-room apartment; its current state is given by the room number it currently occupies ($s \in \{1, 2\}$). It has two possible actions: $a \in \{\text{walk}, \text{purr}\}$. The cat attempts to determine the optimum policy using Q-learning. It starts out with an empty Q-table ($Q_0(s, a) = 0 \;\; \forall s, a$). Starting in state $s_1 = 1$, it receives the following rewards, performs the following actions, and observes the following resulting states:

t	s_t	R_t	a_t	s_{t+1}
1	1	2	purr	1
2	1	2	purr	1

The cat performs one iteration of TD-learning with each of these two observations, using a learning rate of $\alpha = 0.1$ and a discount factor of $\gamma = 1$.

(a) After these two iterations of Q-learning, what values in the Q-table have changed?

(b) Define $Q_t(s, a)$ to be the value of the Q-function after the t^{th} step of TD-learning. What is $Q_2(1, \text{purr})$?