2. Home World Game

Extension Note: Project 5 due date has been extended by 1 **more** day to **September 6 23:59UTC** .

In this project, we will consider a text-based game represented by the tuple $< H, C, P, R, \gamma, \Psi >$. Here H is the set of all possible game states. The actions taken by the player are multi-word natural language **commands** such as **eat apple** or **go east** . In this project we limit ourselves to consider commands consisting of one action (e.g., **eat**) and one argument object (e.g. **apple**).

$C = \{(a,b)\}$ is the set of all commands (action-object pairs).

$P : H \times C \times H \to [0,1]$ is the transition matrix: $P(h'|h,a,b)$ is the probability of reaching state h' if command $c = (a,b)$ is taken in state h.

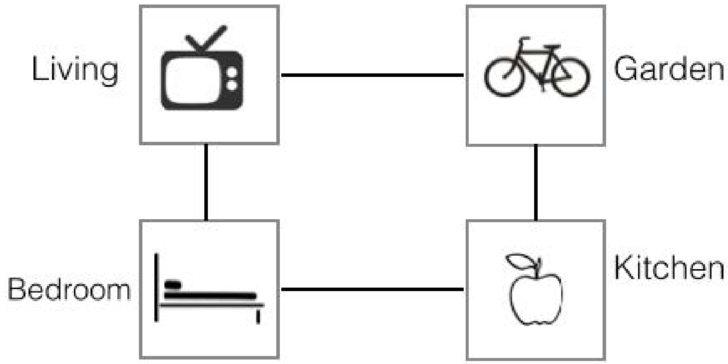
$R : H \times C \to \mathbb{R}$ is the deterministic reward function: $R(h,a,b)$ is the immediate reward the player obtains when taking command (a,b) in state h. We consider discounted accumulated rewards where γ is the discount factor. In particular, the

Generating Speech Output

game state h is **hidden** from the player, who only receives a varying textual description. Let S denote the space of all possible text descriptions. The text descriptions s observed by the player are produced by a stochastic function $\Psi : H \to S$. Assume that each observable state $s \in S$ is associated a **unique** hidden state, denoted by $h(s) \in H$.

You will conduct experiments on a small Home World, which mimic the environment of a typical house. The world consists of four rooms- a living room, a bed room, a kitchen and a garden with connecting pathways (illustrated in figure below). Transitions between the rooms are **deterministic**. Each room contains a representative object that the player can interact with. For instance, the living room has a **TV** that the player can **watch** , and the kitchen has an **apple** that the player can **eat.** Each room has several descriptions, invoked randomly on each visit by the player.

Rooms and objects in the Home world with connecting pathways



Reward Structure

Positive	Negative
Quest goal: $+1$	Negative per step: -0.01
	Invalid command: -0.1

Generating Speech Output

At the beginning of each episode, the player is placed at a random room and provided with a randomly selected quest. An example of a quest given to the player in text is *You are hungry now*. To complete this quest, the player has to navigate through the house to reach the kitchen and eat the apple (i.e., type in command *eat apple*). In this game, the room is *hidden* from the player, who only receives a description of the underlying room. The underlying game state is given by $h = (r, q)$, where r is the index of room and q is the index of quest. At each step, the text description s provided to the player contains two part $s = (s_r, s_q)$, where s_r is the room description (which are varied and randomly provided) and s_q is the quest description. The player receives a positive reward on completing a quest, and negative rewards for invalid command (e.g., *eat TV*). Each non-terminating step incurs a small deterministic negative rewards, which incentives the player to learn policies that solve quests in fewer steps. (see the **Table 1**) An episode ends when the player finishes the quest or has taken more steps than a fixed maximum number of steps.

Each episode produces a full record of interaction $(h_0, s_0, a_0, b_0, r_0, \ldots, h_t, s_t, a_t, b_t, r_t, h_{t+1} \ldots)$ where $h_0 = (h_{r,0}, h_{q,0}) \sim \Gamma_0$ (Γ_0 denotes an initial state distribution), $h_t \sim P(\cdot|h_{t-1}, a_{t-1}, b_{t-1})$, $s_t \sim \Psi(h_t)$, $r_t = R(h_t, a_t, b_t)$ and all commands (a_t, b_t) are chosen by the player. The record of interaction observed by the player is $(s_0, a_0, b_0, r_0, \ldots, s_t, a_t, b_t, r_t, \ldots)$. Within each episode, the quest remains unchanged, i.e., $h_{q,t} = h_{q,0}$ (so as the quest description $s_{q,t} = s_{q,0}$). When the player finishes the quest at time K, all rewards after time K are assumed to be zero, i.e., $r_t = 0$ for $t > K$. Over the course of the episode, the total discounted reward obtained by the player is

$$\sum_{t=0}^{\infty} \gamma^t r_t.$$

We emphasize that the hidden state $h_0, \ldots, h_T$ are unobservable to the player.

Generating Speech Output

The learning goal of the player is to find a policy that $\pi : S \to C$ that maximizes the expected cumulative discounted reward $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(h_t, a_t, b_t) \mid (a_t, b_t) \sim \pi\right]$, where the expectation accounts for all randomness in the model and the player. Let π^* denote the optimal policy. For each observable state $s \in S$, let $h(s)$ be the associated hidden state. The optimal expected reward achievable is defined as

$$V^* = \mathbb{E}_{h \sim \Gamma_0, s \sim \Psi(h)}\left[V^*(s)\right]$$

where

$$V^*(s) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(h_t, a_t, b_t) \,|\, h_0 = h(s), s_0 = s, (a_t, b_t) \sim \pi\right].$$

We can define the optimal Q-function as

$$Q^*(s, a, b) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(h_t, a_t, b_t) \,|\, h_0 = h(s), s_0 = s, a_0 = a, b_0 = b, (a_t, b_t) \sim \pi \text{ for } t \geq 1\right].$$

Note that given $Q^*(s, a, b)$, we can obtain an optimal policy:

$$\pi^*(s) = \arg\max_{(a,b) \in C} Q^*(s, a, b).$$

Generating Speech Output

The commands set C contain all $(action, object)$ pairs. Note that some commands are invalid. For instance, **(eat,TV)** is invalid for any state, and **(eat, apple)** is valid only when the player is in the kitchen (i.e., h_r corresponds to the index of kitchen). When an invalid command is taken, the system state remains unchanged and a negative reward is incurred. Recall that there are **four** rooms in this game. Assume that there are **four** quests in this game, each of which would be finished only if the player takes a particular **command** in a particular room. For example, the quest “You are sleepy" requires the player navigates through rooms to bedroom (with commands such as **go east/west/south/north**) and then take a nap on the bed there. For each room, there is a corresponding quest that can be finished there.

Note that in this game, the transition between states is deterministic. Since the player is placed at a random room and provided a randomly selected quest at the beginning of each episode, the distribution Γ_0 of the initial state h_0 is uniform over the hidden state space H.

Episodic reward

1.0/1 point (graded)

For an episode with $T+1$ steps (starting from $t=0$), where the agent obtains a reward R_t at time step t. What is the total discounted reward for this episode with a discounted factor $\gamma \in (0,1)$?

Important: If needed, please enter $\sum_{t=0}^{T}(\ldots)$ as a function `sum_t(...)`, including the parentheses.

STANDARD NOTATION

sum_t(gamma^t*R_t)

✔ **Answer:** sum_t(gamma^t*R_t)

Generating Speech Output ave used 1 of 6 attempts

Answers are displayed within the problem

Relation between value function and Q-function

1/1 point (graded)
Which of the following equation gives the correct relation between Q^* and V^*?

- $Q^*(s,a,b) = \gamma \mathbb{E}\left[V^*(s_0)\,|h_0 = h(s), s_0 = s, a_0 = a, b_0 = b\right]$
- $Q^*(s,a,b) = \gamma \mathbb{E}\left[V^*(s_1)\,|h_0 = h(s), s_0 = s, a_0 = a, b_0 = b\right]$
- $Q^*(s,a,b) = R(s,a,b) + \mathbb{E}\left[V^*(s_0)\,|h_0 = h(s), s_0 = s, a_0 = a, b_0 = b\right]$
- $Q^*(s,a,b) = R(s,a,b) + \mathbb{E}\left[V^*(s_1)\,|h_0 = h(s), s_0 = s, a_0 = a, b_0 = b\right]$
- $Q^*(s,a,b) = R(s,a,b) + \gamma \mathbb{E}\left[V^*(s_0)\,|h_0 = h(s), s_0 = s, a_0 = a, b_0 = b\right]$
- $Q^*(s,a,b) = R(s,a,b) + \gamma \mathbb{E}\left[V^*(s_1)\,|h_0 = h(s), s_0 = s, a_0 = a, b_0 = b\right]$ ✔

Submit You have used 1 of 4 attempts

Generating Speech Output

Answers are displayed within the problem

Optimal episodic reward

1/1 point (graded)

Assume that the reward function $R(s,a,b)$ is given in Table 1. At the beginning of each game episode, the player is placed in a random room and provided with a randomly selected quest. Let $V^*(h_0)$ be the optimal value function for an initial state h_0, i.e.,

$$V^*(h_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(h_t, a_t, b_t) | \pi^*\right]$$

Please compute the expected optimal reward for each episode $\mathbb{E}[V^*(h_0)]$. Note that the initial state h_0 is uniformly distributed in the state space $H = (r,q) : 0 \leq r \leq 3, 0 \leq q \leq 3$. In other words, there are four quests each mapping to a unique room.Assume that the discounted factor is $\gamma = 0.5$

0.55375

Answer: 0.55375

Solution:

We can categorize the states $S = \{(s_r, s_q)\}$ into three types:

1. The quest s_q requests a command in the initial room with description s_r. An example of such initial states is **(This room has a fridge, oven, and a sink; you are hungry)** . The optimal policy for such a state is to take the

Generating Speech Output ng command to finish the quest and get a reward 1.

2. The quest s_q requests a command in a room next to the initial room with description s_r. An example is **(This area has a bed, desk and a dresser; you are hungry)** . The optimal policy for such a state is first take one step towards the goal room (e.g., **go west,** and get a penalty reward -0.01), and then take the corresponding command to finish the quest (e.g., **eat apple,** and get a positive reward 1). The total discounted reward is: $-0.01 + \gamma \times 1 = 0.49$.

3. The quest s_q requests a command in a room that is not next to the initial room with description s_r, for instance, **(You have arrived at the garden. You can exercise here; you are hungry)** . It is easy to see that the optimal policy would be taking the first steps to arrive at the quested room and then finishing the quest. The total discounted reward would be: $-0.01 + \gamma \times (-0.01) + \gamma^2 \times 1 = 0.235$.

Since the room and the quest are selected randomly for the initial state, the probabilities for the above three types of states are $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ respectively. Therefore,

$$\mathbb{E}\left[V^*\left(h_0\right)\right] = \frac{1}{4} \times 1 + \frac{1}{2} \times 0.49 + \frac{1}{4} \times 0.235 = 0.55375.$$

Submit You have used 1 of 6 attempts

Answers are displayed within the problem

Discussion

Hide Discussion

Topic: Unit 5 Reinforcement Learning (2 weeks) :Project 5: Text-Based Game / 2. Home World Game

Generating Speech Output

Add a Post

Show all posts ▼ by recent activity ▼

? [STAFF]Answer for Optimal episodic reward is "invalid math syntax" 1
I submitted right answer but the system return "invalid math syntax". I checked my answer by "Show Answer" and my answer was right. Did this ...

? [STAFF] there is a mistake equation of optimal policy 4
hi the equation for optimal policy says π∗(s) = max(a,b)∈C Q∗(s,a,b) . I feel that that it should be arg max since the output of a policy is an action, p...

? [Staff] - Please add extra attempt for Episodic reward question 2
I entered the answers meant for optimistic episodic reward in episodic reward so please add extra attempt.

Optimal episodic reward (thought process) 43 new 48

? Optimal episodic reward: latest staff edits broke formatting 3

Optimal episodic reward question. 5

? ??? Relation between value function and Q-function 3
Three of the choices appear to be ill defined in that they rely on an unbound symbol, yet the grader marks as Incorrect each of the other three c...

Increase of max attempt numbers 6
Would it be possible to increase the numbers of max attempt of this tab? Thanks in advance.

? [STAFF] Extension Possibility? 9
I know it is a little early to be asking for this but can we get an extension on this assignment until a few days after the Homework becomes due s...

Notation Overload 7
After this page half of my brain is dedicated to store memory on notations for q-learning.

[staff] Table 1? 3
Generating Speech Output
Hi, probably Table 1. is the only picture/table in the tab, but there is nowhere name of it written (or maybe there is, but I don't see it :))

? Episodic reward
1. Why is "Invalid Input: gamma^{t} not permitted in answer as a variable"? Doesn't discount depend on "t" and when is it 0 and when 1? 2. What ...
3

? [STAFF] "Optimal episodic reward" add more attempts?
I have used up my attempts and think I realize the error in my thinking for the problem "Optimal episodic reward". I would appreciate getting on...
5

Optimal Episodic Reward: steps?
If I start in the kitchen and give a command "eat apple" as soon as the game begins and finish the quest right away... 1) Did it take 0 step or 1 ste
9

Generating Speech Output