4. Tabular Q-learning for Home World game

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

In this section you will evaluate the tabular Q-learning algorithms for the *Home world* game. Recall that the state observable to the player is described in text. Therefore we have to choose a mechanism that maps text descriptions into vector representations.

In this section you will consider a simple approach that assigns a unique index for each text description. In particular, we will build two dictionaries:

- `dict_room_desc` that takes the room description text as the key and returns a unique scalar index
- `dict_quest_desc` that takes the quest description text as the key and returns a unique scalar index.

For instance, consider an observable state $s = (s_r, s_q)$, where s_r and s_q are the text descriptions for the current room and the current request, respectively. Then i_r =`dict_room_desc[`s_r`]` gives the scalar index for s_r and i_q =`dict_quest_desc[`s_q`]` gives the scalar index for s_q. That is, the textual state $s = (s_r, s_q)$ is mapped to a tuple $I = (i_r, i_q)$.

Normally, we would build these dictionaries as we train our agent, collecting descriptions and adding them to the list of known descriptions. For the purpose of this project, these dictionaries will be provided to you.

Evaluating Tabular Q-learning on Home World

1.0/1 point (graded)
The following python files are provided:

- `framework.py` contains various functions for the text-based game environment that the staff has implemented for you. Some functions that you can call to train and testing your reinforcement learning algorithms:
  - `newGame()`
    - Args: None
    - Return: A tuple where the first element is a description of the initial room, the second element is a description of the quest for this new game episode, and the last element is a Boolean variable with value *False* implying that the game is not over.
  - `step_game()`
    - Args:
      - `current_room_desc` : An description of the current room

- `current_quest_desc` : A description of the current quest state
- `action_index` : An integer used to represent the index of the selected action
- `object_index` : An integer used to indicate the index of the selected object

- Return: the system next state when the selected command is applied at the current state.
  - `next_room_desc` : The description of the room of the next state
  - `next_quest_desc` : The description of the next quest
  - `reward` : A real valued number representing the **one-step** reward obtained at this step
  - `terminal` : A boolean valued number indicating whether this episode is over (either quest is finished, or the number of steps reaches the maximum number of steps for each episode).

- `agent_tabular_QL.py` contains various function templates that you will use to implement your learning algorithm.

In this section, you will evaluate your learning algorithm for the Home World game. The metric we use to measure an agent's performance is the cumulative discounted reward obtained per episode averaged over the episodes.

The evaluation procedure is as follows. Each experiment (or run) consists of multiple epochs (the number of epochs is `NUM_EPOCHS`). In each epoch:

1. You first train the agent on `NUM_EPIS_TRAIN` episodes, following an ε-greedy policy with $\varepsilon =$ `TRAINING_EP` and updating the Q values.

2. Then, you have a testing phase of running `NUM_EPIS_TEST` episodes of the game, following an ε-greedy policy with $\varepsilon =$ `TESTING_EP`, which makes the agent choose the best action according to its current Q-values 95% of the time. At the testing phase of each epoch, you will compute the cumulative discounted reward for each episode and then obtain

the average reward over the `NUM_EPIS_TEST` episodes.

Finally, at the end of the experiment, you will get a sequence of data (of size `NUM_EPOCHS`) that represents the testing performance at each epoch.

Note that there is randomness in both the training and testing phase. You will run the experiment `NUM_RUNS` times and then compute the averaged reward performance over `NUM_RUNS` experiments.

Most of these operations are handled by the boilerplate code provided in the `agent_tabular_QL.py` file by functions `run`, `run_epoch` and `main`, but you will need to complete the `run_episode` function.

Write a `run_episode` function that takes a boolean argument (whether the epsiode is a training episode or not) and runs one episode.

Reminder: You should implement this function locally first. Make sure you can achieve reasonable performance on the Home World game before submitting your code

Available Functions: You have access to the NumPy python library as `np`, framework methods `framework.newGame()` and `framework.step_game()`, constants `TRAINING_EP` and `TESTING_EP`, `GAMMA`, dictionaries `dict_room_desc` and `dict_quest_desc` and previously implemented functions `epsilon_greedy` and `tabular_QLearning`

```
def run_episode(for_training):
    """ Runs one episode
    If for training, update Q function
    If for testing, computes and return cumulative discounted reward

    Args:
        for_training (bool): True if for training

    Returns:
        None
```

```
    """
    epsilon = TRAINING_EP if for_training else TESTING_EP

    epi_reward = 0
```

Press ESC then TAB or click outside of the code editor to exit

Correct

```
def run_episode(for_training):
    """ Runs one episode
    If for training, update Q function
    If for testing, computes and return cumulative discounted reward

    Args:
        for_training (bool): True if for training

    Returns:
        None
    """
    epsilon = TRAINING_EP if for_training else TESTING_EP
    gamma_step = 1
    epi_reward = 0

    (current_room_desc, current_quest_desc, terminal) = framework.newGame()
    while not terminal:
        # Choose next action and execute
        cur_room_desc_id = dict_room_desc[current_room_desc]
        cur_quest_desc_id = dict_quest_desc[current_quest_desc]
        (action_index, object_index) = epsilon_greedy(cur_room_desc_id,
                                                      cur_quest_desc_id,
                                                      q_func, epsilon)
        (next_room_desc, next_quest_desc, reward,
         terminal) = framework.step_game(current_room_desc, current_quest_desc,
                                         action_index, object_index)

        if for_training:
            # update Q-function.
            next_room_desc_id = dict_room_desc[next_room_desc]
            next_quest_desc_id = dict_quest_desc[next_quest_desc]
            tabular_q_learning(q_func, cur_room_desc_id, cur_quest_desc_id,
                               action_index, object_index, reward,
```

```
                                next_room_desc_id, next_quest_desc_id, terminal)

        if not for_training:
            # update reward
            epi_reward = epi_reward + gamma_step * reward
            gamma_step = gamma_step * GAMMA

        # prepare next step
        current_room_desc = next_room_desc
        current_quest_desc = next_quest_desc

    if not for_training:
        return epi_reward
```

Test results

Hide output

CORRECT

Test: correctly computes reward

Testing computes correct reward for testing episode

Output:

```
eps greedy  / eps greedy  / eps greedy  / eps greedy  / eps greedy  / eps greedy  /
-0.162500
```

Test: updates qvalue

Testing computes correct reward for testing episode

Output:

```
eps greedy  / ql / eps greedy  / ql / eps greedy  / ql / eps greedy  / ql / eps greedy  / ql /
eps greedy  / ql /
```

Hide output

Submit You have used 0 of 25 attempts

Answers are displayed within the problem

Report performance

2/2 points (graded)

In your Q-learning algorithm, initialize Q at zero. Set `NUM_RUNS` $= 10$, `NUM_EPIS_TRAIN` $= 25$, `NUM_EPIS_TEST` $= 50$, $\gamma = 0.5$, `TRAINING_EP` $= 0.5$, `TESTING_EP` $= 0.05$ and the learning rate $\alpha = 0.1$.

Please enter the number of epochs when the learning algorithm converges. That is, the testing performance become stable.

15 ✔ **Answer:** 15

Please enter the *average episodic rewards* of your Q-learning algorithm when it converges.

0.5110307482421875 ✔ **Answer:** 0.52

Submit You have used 3 of 6 attempts

Answers are displayed within the problem

Discussion

Hide Discussion

Topic: Unit 5 Reinforcement Learning (2 weeks) :Project 5: Text-Based Game / 4. Tabular Q-learning for Home World game

Add a Post

Show all posts ▼ by recent activity ▼

? [Staff] Could you please kindly check my code? 3
Dear Staff, Could you please kindly check my code? Before asking for your help you I did review the discussion forum posts. The issue is in the re...

☑ Need help!!! 5

? Episodic Reward Too Low 3
I have passed the grader on my code and the number of epochs at convergence was accepted (although it seemed too fast to me). However the ...

Read No of epochs out of graph - hint 4
Don't apply every parameter written in the task blindly in your code...wouldn't it be better to have a more exact graph in the relevant area and ru...

?	[STAFF] The last question is super frustrating. Please add more attempts. My algorithm is converging very fast, the first value got accepted. So I calculated the mean value of all the rewards from all episodes for this valu...	5
?	[STAFF] Convergence velocity (can someone look my code?) I read all the discussion about the convergence, I was getting a low convergence number and my code passed the grader even so. I found an err...	4
?	run_episode(for_training)	4
?	Grader accepted but locally code has TypeError While trying to debug my code I decided to run it through the grader as I sometimes learn from the expected outputs it is giving and to my surpri...	5
?	[staff] Tabular Q-learning_Please check my code I am almost there, geting the update qvalue correct but still struggling with the reward. I thought I have upadte the reward correctly, but it seems...	5
💬	[staff] Report Performace - Please, have a look at my code.	5
?	[Staff] Check my run_episode code, pls I have passed grader but I am curious why I am getting very low cumulative reward compared to expected reward? thanks.	5
?	[staff] why do I get staff debug error? Hi, I receive the following message: There was a problem running your solution (Staff debug: L379). We couldn't run your solution (Staff debug: L...	8
💬	run_episode( ) half correct I am stuck on deciding the initial value of epi_reward in the run_episode( ) function. Any hint, please?	12
?	Trying to understand the comment "initialize for each episode" in the beginning part of run_episode() In the skeleton code provided for run_episode the first part is as shown below. What does the last comment "initialize for each episode" refer to?	4