MSAI-349, Fall 2024
Homework #4: Reinforcement Learning and Ethics
Due Date: Monday, December 9th @ 11:59PM
Total Points: 10.0

In this assignment, you will work with your group to implement Q-learning for the multi-arm bandit and frozen lake environments. You may discuss the homework with other groups, but do not take any written record from these discussions. Also, do not copy any source code from the Web. Your submission must be your own.

1. **Coding (2.0 points):** Your task is to implement two reinforcement learning algorithms:
   - Multi-armed Bandits (in `src/multi_armed_bandits.py`)
   - Q-Learning (in `src/q_learning.py`)

   Note that while these reinforcement learning methods inherently depend on randomization, we provide a `src/random.py` package that will randomize things in the same way for all students. Please use `src.random` anywhere that you might have otherwise used `np.random`. Your goal is to pass the test suite (contained in tests/). You can test your code by running `pytest tests/test_bandit.py` and `pytest tests/test_q_learning.py`. Once the tests are passed, you will use your code to answer FRQ 2 and 3.

2. **Bandits vs. Q-Learning (2.0 points):**
   a. Run `python -m q2.py`; it will create three plots: `2a_SlotMachines_Comparison.png`, `2a_FrozenLake_Comparison.png`, and `2a_SlipperyFrozenLake_Comparison.png`. Please read about the FrozenLake environment (https://gymnasium.farama.org/environments/toy_text/frozen_lake/) . Each plot will show a comparison of your MultiArmedBandit and QLearning models on the named environment (e.g., SlotMachines). Include those plots here. For each plot, provide a one-sentence description of the most notable trend. Pay attention to the scale on the y-axis.
   b. In which of the above plots does Q-Learning appear to receive higher rewards on average than MultiArmedBandit? Provide an explanation for why that happens, based on your understanding of Q-Learning.
   c. Following b.: in the environment(s) where MultiArmedBandit was the worse model, is there any way you could change your choice of hyperparameters so that MultiArmedBandit would perform as well as Q-Learning? Why or why not?
   d. In which of the above plots does MultiArmedBandit appear to receive higher rewards on average than Q-Learning? Provide an explanation for why that happens, based on your understanding of MultiArmedBandit.
   e. Following d.: in the environment(s) where Q-Learning was the worse model, is there any way you could change your choice of hyperparameters so that QLearning would perform as well as MultiArmedBandit? Why or why not?

3. **Exploration vs. Exploitation (2.0 points):**
    a. Look at the code in `q3.py` and run `python -m q3.py` and include the plot it creates (`free_response/3a_g0.9_a0.2.png`) as your answer to this part. In your own words, describe what this code is doing.
    b. Using the above plot, describe what you notice. What seems to be the ``best'' value of epsilon? What explains this result?
    c. The above plot trains agents for 50,000 timesteps each. Suppose we instead trained them for 500,000 or 5,000,000 timesteps. How would you expect the trends to change or remain the same for each of the three values of epsilon? Give a one-sentence explanation for each value.
    d. When people use reinforcement learning in practice, it can be difficult to choose epsilon and other hyperparameters. Instead of trying three options like we did above, suppose we tried 30 or 300 different choices. What might be the danger of choosing epsilon this way if we wanted to use our agent in a new domain?

4. **Tic-Tac-Toe (2.0 points):** Suppose we want to train a Reinforcement Learning agent to play the game of Tic-Tac-Toe (see https://en.wikipedia.org/wiki/Tic-tac-toe), and need to construct an environment with states and actions. Assume our agent will simply choose actions based on the current state of the game, rather than trying to guess what the opponent will do next.
    a. What should be the states and actions within the Tic-Tac-Toe Reinforcement Learning environment? Don't try to list them all, just describe how the rules of the game define what states and actions are possible. How does the current state of the game affect the actions you can take?
    b. Design a reward function for teaching a Reinforcement Learning agent to play optimally in the Tic-Tac-Toe environment. Your reward function should specify a reward value for each of the 3 possible ways that a game can end (win, loss, or draw) as well as a single reward value for actions that do not result in the end of the game (e.g., your starting move). Explain your choices.
    c. Suppose you were playing a more complicated game with a larger board, and you want the agent to learn to win as fast as possible. How might you change your reward function to encourage speed?

5. **Fair ML in the Real World (2.0 points):** Read Joy Buolamwini and Timnit Gebru, 2018. "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on fairness, accountability and transparency, then use it to help answer the following questions (see https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf).
    a. Buolamwini and Gebru use PPV and FPR as metrics to measure fairness. Find the definition of these in the paper, then look up the corresponding definition for NPV and FNR (these appear in the slides). Assuming you were applying for a loan and you know a ML classifier is deciding whether to grant it to you: would you rather have that decision made by a system with a high FPR or a high FNR? Why? Provide a detailed justification.
    b. Assuming you were applying for a loan and you know a ML classifier is deciding whether to grant it to you: would you rather have that decision made by a system with a high PPV or a high NPV? Why? Provide a detailed justification.
    c. What recommendations do Buolamwini and Gebru make regarding accountability and transparency of ML systems? How does this relate to specific metrics such as PPV or FPR?
    d. What is intersectional about the analysis conducted by the authors? What does that analysis show?
    e. In Section 4.7, the authors say that their "findings ... do not appear to be confounded by the quality of sensor readings." What do they mean by "confounded" in this context? Why is it important to the thesis of this paper to check whether their findings are confounded?

Submission Instructions

Turn in your homework as a single zip file, in Canvas. Specifically:

1. Create a single pdf file hw4.pdf with the answers to the questions above and summaries of your results.
2. Create a single ZIP file containing:
    - `hw4.pdf`
    - All of your `.py` code files
3. Turn the zip file in under Homework #4 in Canvas.

Good luck, and have fun!