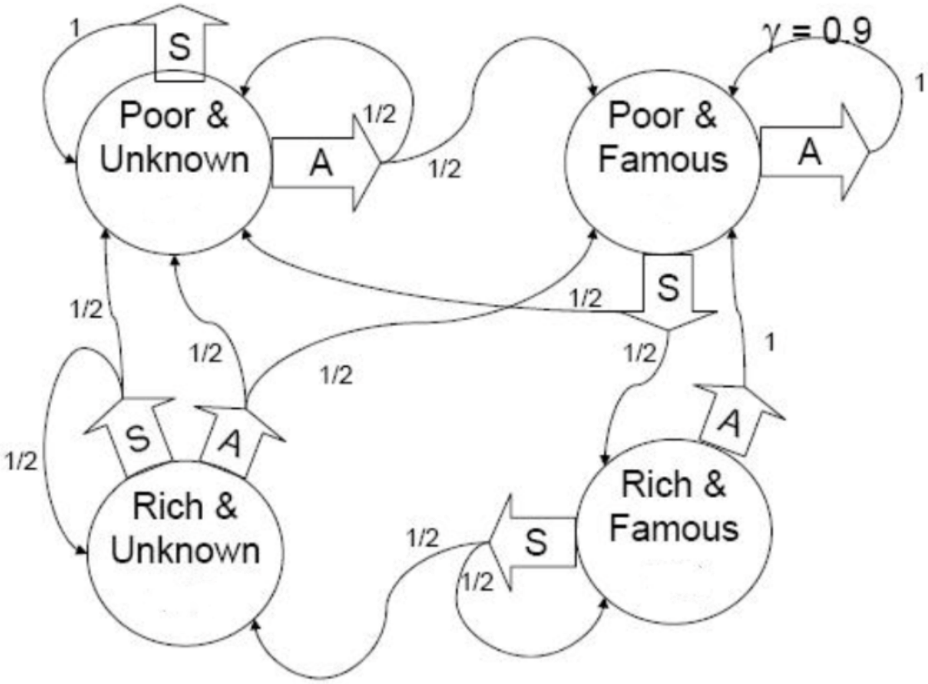
THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY
Department of Computer Science and Engineering
COMP4211: Introduction to Machine Learning
Spring 2022: Assignment 3
Due time and date: 11:59pm, May 6 (Fri), 2022.

IMPORTANT NOTES

- **Your grade will be based on the correctness and clarity.**
- **Late submission: 25 marks will be deducted for every 24 hours after the deadline.**
- **If you have questions, please contact the TA Weiyu Chen at wchenbx@connect.ust.hk.**

The following figure shows a problem from the lecture notes. It has 4 states. At each state, the possible actions are: advertise (A) and save-money (S). When the action takes the agent to the states "Poor & Unknown" or "Poor & Famous", the immediate reward is 0; whereas when the action takes the agent to the states "Rich & Unknown" or "Rich & Famous", the immediate reward is 10. The world is non-deterministic. The numbers shown along the action outcomes are probabilities for the corresponding transitions. The discount factor γ is 0.9.



Q1. Consider the policy π_0: "always advertise" (i.e., always choose the action A). Write down the Bellman equations and obtain V^{π_0} by solving a linear system. For simplicity, in your answer denote the states "Poor & Unknown", "Poor & Famous", "Rich & Unknown", "Rich & Famous" by s_1, s_2, s_3, s_4, respectively.

Q2. Perform one iteration of policy improvement on policy π_0, and obtain the updated π_1.

Q3. Implement policy iteration for the above problem in Python. Partial code for problem definition is provided in assignment3.ipynb. Run the algorithm until the policy is stable. Print the obtained policy in the following format:

```
state: Poor & Unknown, action: ... , value: ...
state: Poor & Famous,  action: ... , value: ...
state: Rich & Unknown, action: ... , value: ...
state: Rich & Famous,  action: ... , value: ...
```

Q4. Implement value iteration for the above problem in Python. Stop the iteration when the maximum value change over all states (i.e., Δ in the lecture notes) is smaller than 10^{-4}. Print the obtained policy in the same format as above.

Q5. In this question, we learn the optimal policy by using Q-learning. As the world is non-deterministic, use a learning rate α of 0.2. The following is a partially learned Q table.

	S	A
Poor & Unknown	0	0
Poor & Famous	2	0
Rich & Unknown	0	0
Rich & Famous	5.88	0.36

Starting from the state "Rich & Unknown", show the Q table after taking each of the following actions. Show your steps clearly, and report the numbers to four decimal places.

1. "Rich & Unknown" $\xrightarrow{A}$ "Poor & Famous"

	S	A
Poor & Unknown	?	?
Poor & Famous	?	?
Rich & Unknown	?	?
Rich & Famous	?	?

2. Then, "Poor & Famous" $\xrightarrow{S}$ "Rich & Famous"

	S	A
Poor & Unknown	?	?
Poor & Famous	?	?
Rich & Unknown	?	?
Rich & Famous	?	?

3. Then, "Rich & Famous" $\xrightarrow{S}$ "Rich & Unknown"

	S	A
Poor & Unknown	?	?
Poor & Famous	?	?
Rich & Unknown	?	?
Rich & Famous	?	?

Submission Guidelines

Please include
(i) a report report.pdf containing your results on Q1, Q2 and Q5.
(ii) an executable Python notebook file (.ipynb file) for Q3 and Q4.

Zip all the files to YourStudentID_assignment3.zip (e.g., 12345678_assignment3.zip). Please submit the assignment by uploading the compressed file to Canvas. Note that the assignment should be clearly legible, otherwise you may lose some points if the assignment is difficult to read. **Plagiarism will lead to zero point on this assignment.**