CIS 391/521: HW 8 - Reinforcement Learning

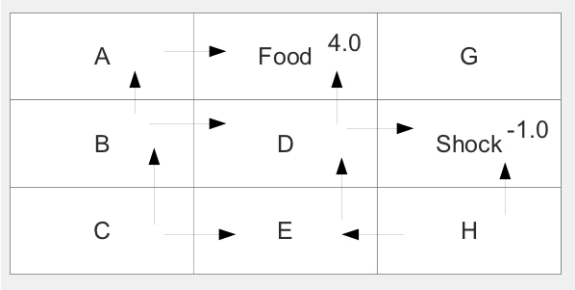
This homework consists of a written portion and a programming portion. Please submit your written responses and final version of your code on **Blackboard** at the **beginning** of class on **Tuesday, April 17**.

Let us know if you have any questions, check the discussion board, and remember that you can always come to Office Hours, even if only to keep the TAs company.

1 Written Portion (20 points)

This part must be done individually.

1. (20 points) Consider the following maze, a 3 x 3 set of rooms, with connections between adjacent rooms



Imagine you put a mouse in such a maze for the very first time, so that the values associated with each square are all 0. When the mouse enters the food or shock square, the mouse is removed to its cage and receives either food or a shock depending on which of the two squares was entered. Assign a value $V(exit) = 0$ to the state where the mouse is in its cage; this value never changes. Assume a learning rate α of 0.35. Allowed moves are shown by arrows, and the numbers indicate the reward for performing each action. If there is no number, the reward is zero.

   (a) Suppose the mouse starts on square C and takes the path $C \rightarrow E \rightarrow D \rightarrow Food$. Using the reinforcement learning rule discussed in class and in the handout, provide the new values of squares C, E, D, and $Food$.

   (b) On a second trial, the mouse starts on square H and takes the path $H \rightarrow E \rightarrow D \rightarrow Food$. Provide the new values for squares H, E, D, and $Food$.

   (c) On a third trial, the mouse starts on square B and takes the path $B \rightarrow D \rightarrow Shock$. Provide the new values for squares B, D and $Shock$.

   (d) After a very large number of trials, assuming that there is a certain amount of randomness in the path taken, would you expect square A or square H to have a higher value? Why?

(e) Explain the role of the learning rate in reinforcement learning. When there are probabilistic transitions, will you need a more complicated formulation of the learning rate or not? Explain why.

(f) Instead of using temporal difference learning, let's use Q-learning. We use the same learning rate $\alpha = 0.35$ and discount factor $\gamma = 0.5$. When choosing actions during learning, the mouse picks the action with the highest Q-value with probability 0.9 and picks randomly with probability 0.1. If all moves have the same Q-value, the mouse picks the action randomly. Initially the Q-values are all zeros. At the beginning of each episode, mouse starts at C. What is the expected value of all the Q values after one, two, and three episodes? (An "episode" is a mouse being put in and moving through the maze until it is taken out.)

(g) If you performed this Q-learning for a large number of episodes, what policy would Q learning produce?

2 Programming Portion (35 points)

In this section, you will learn to program a basic ant in Python and implement reinforcement learning on locating food. This part should be done in pair or alone.

1. Getting Started with Ants - How to Play Ants on Competition Server

   (a) Go to the competition website and register an account under your SEAS username (If you are a pair, use as a username your two SEAS usernames joined with '_'). Please only register a single account for now.

   (b) Extract the attached zip file on Blackboard.

   (c) Edit MyBot.py to import the bot you'd like to run, and to create an instance of the bot you'd like to run.

   (d) Once you are done, zip the contents of the base directory into a submit.zip file inside the directory. (You must not create a zip with a "base" directory inside.)

   (e) Upload your submit.zip file to the competition using the competition server website.