CS 445/545

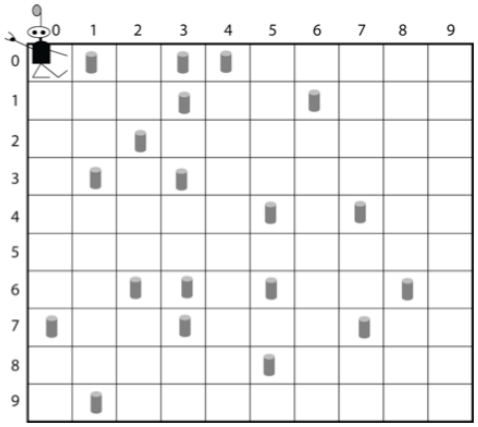
Machine Learning, Winter 2017

Homework 6: Reinforcement Learning

Due Thursday, March 16, 2017, 2pm

In this homework you will write code to have Robby the Robot use Q-learning to learn to correctly pick up cans and avoid walls in his grid world.

Robby the Robot:
As was described in class, Robby the Robot lives in a 10 x 10 grid, surrounded by a wall. Some of the grid squares contain soda cans.



Robby has five "sensors": Here, North, South, East, and West. At any time step, these each return the "value" of the respective location, where the possible values are Empty, Can, and Wall.

Robby has five possible actions: Move-North, Move-South, Move-East, Move-West, and Pick-Up-Can. Note: if Robby picks up a can, the can is then gone from the grid.

Robby receives a reward of 10 for each can he picks up; a "reward" of −5 if he crashes into a wall (after which he immediately bounces back to the square he was in); and a reward of −1 if he tries to pick up a can in an empty square.

Your Assignment:

Part 1: Write a (simple!) simulator for Robby in which he receives sensor input, can perform actions, and receives rewards. Write a Q-learning method for Robby, using a Q-matrix, in which the rows correspond to states and the columns correspond to actions. (We will discuss in class how to easily map sensor input to state-index in the Q-matrix.) The Q-matrix is initialized to all zeros at the beginning of a run.

During a run, Robby will learn over a series of *N* episodes, during each of which he will perform *M* actions. The initial state of the grid in each episode is a random placement of cans, where each grid square has a probability of 0.5 to contain a can (and 0.5 not to contain a can). Robby is initially placed in a random grid square.

At each time step t during an episode, your code should do the following:

- Observe Robby's current state s_t
- Choose an action a_t, using ε-greedy action selection
- Perform the action
- Receive reward r_t (which is zero except in the cases specified above)
- Observe Robby's new state s_{t+1}
- Update $Q(s_t, a_t) = Q(s_t, a_t) + \eta(r_t + \gamma max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t))$

At the end of each episode, generate a new distribution of cans and place Robby in a random grid square to start the next episode. (Don't reset the Q-matrix — you will keep updating this matrix over the N episodes. Keep track of the total reward gained per episode.

To do a run consisting of N episodes of M steps each, use the following parameter values:

$$N = 5{,}000 \; ; \; M = 200 \; ; \; \eta = \; 0.2; \; \gamma = \; 0.9$$

For choosing actions with ε-greedy action selection, set $\varepsilon = 1$ initially, and reduce it by 0.01 every 50 epochs until it reaches 0.1. After that, it stays at 0.1.

Run the N episodes of learning, and plot the total sum of rewards per episode (plotting a point every 100 episodes). This plot—let's call it the *Training Reward* plot—indicates the extent to which Robby is learning to improve his cumulative reward.

After training is completed, run N *test* episodes using your trained Q-matrix, but with $\varepsilon = 0.1$ for all N episodes. Again regenerate a grid of randomly placed cans at the beginning of each episode and also place Robby in a random grid location. Calculate the average over sum-of-rewards-per-episode, and the standard deviation. For simplicity in this writeup, let's call these values *Test-Average* and *Test-Standard-Deviation*. These values indicate how a *trained* agent performs this task in new environments.

In your report, describe the experiment, give the *Training Reward* plot described above (plotted every 100 episodes), and *Test-Average* and *Test-Standard-Deviation*.

Part 2: Experiment with Learning Rate. Choose 4 different values for the learning rate, η, approximately evenly spaced in the range [0,1], keeping the other parameters set as in Part 1. For each value, give the *Training Reward* plot (plotted every 100 episodes), and the *Test-Average* and *Test-Standard-Deviation*. Discuss how changing the learning rate changes these results.

Part 3: Experiment with Epsilon. Try learning with a constant epsilon (choose a value ε in [0,1]). Give the *Training Reward* plot and *Test-Average* and *Test-Standard-Deviation*. How do your results change when using a constant value of epsilon rather than a decreasing value? Speculate on why you get these results.

Part 4: Experiment with negative reward for each action. Modify your code so that a negative reward (an "action tax") of -0.5 is given in addition to the original rewards for each action. Run learning and testing with the parameter values of Part 1, and give the *Training Reward* plot and the *Test-Average* and *Test-Standard-Deviation*. What differences do you see from the results in Part 1?

Part 5: Devise your own experiment, different from those of Parts 1−4 above. This can involve a change to a parameter value, a change in the rewards, a modification of the actions or sensors, etc. Describe your experiment and give plots or values that show the results. Are the results what you expected? Why or why not?

Here is what you need to turn in:

Your spell-checked, double-spaced report with the information requested above. Also, your commented code with instructions how to run it.

How to turn it in (read carefully!):

- Send these items in electronic format to mm@pdx.edu by 2pm on the due date. No hard copy please!
- The report should be in pdf format and the code should be in its original format (e.g., .py, .m, etc.)
- Put "MACHINE LEARNING HW 6" in the subject line.

If there are any questions, don't hesitate to ask me or e-mail the class mailing list.

Policy on late homework: If you are having trouble completing the assignment on time for any reason, please see me before the due date to find out if you can get an extension. Any homework turned in late without an extension from me will have 5% of the grade subtracted for each day the assignment is late.