CSE 150. Assignment 6

Out: *Thu Mar 05*
Due: *Thu Mar 12* **(No extensions allowed!)**
Reading: *Sutton & Barto*, Chapters 1-4.

6.1 CAPE Survey

You should have received an email from CAPE asking you to evaluate this course. **Please complete the online survey if you have not already done so.** Your answers not only affect future offerings of this course, but also other related courses in artificial intelligence.

6.2 Policy improvement

Consider the Markov decision process (MDP) with two states $s \in \{0, 1\}$, two actions $a \in \{0, 1\}$, discount factor $\gamma = \frac{3}{4}$, and rewards and transition matrices as shown below:

s	$R(s)$
0	-6
1	6

s	s'	$P(s'\|s, a=0)$
0	0	$\frac{1}{2}$
0	1	$\frac{1}{2}$
1	0	$\frac{1}{2}$
1	1	$\frac{1}{2}$

s	s'	$P(s'\|s, a=1)$
0	0	$\frac{2}{3}$
0	1	$\frac{1}{3}$
1	0	$\frac{1}{3}$
1	1	$\frac{2}{3}$

(a) Consider the policy π that chooses the action $a = 0$ in each state. For this policy, solve the linear system of Bellman equations (by hand) to compute the state-value function $V^\pi(s)$ for $s \in \{0, 1\}$. Your answers should complete the following table. **Show your work for full credit.**

s	$\pi(s)$	$V^\pi(s)$
0	0	
1	0	

(b) Compute the greedy policy $\pi'(s)$ with respect to the state-value function $V^\pi(s)$ from part (a). Your answers should complete the following table. **Show your work for full credit.**

s	$\pi(s)$	$\pi'(s)$
0	0	
1	0	

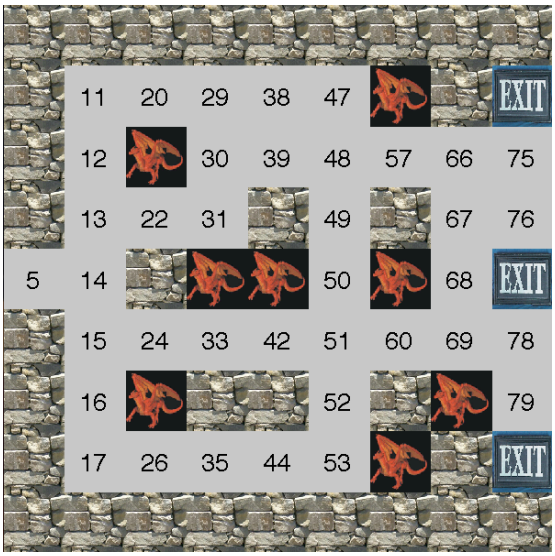
6.3 Value iteration

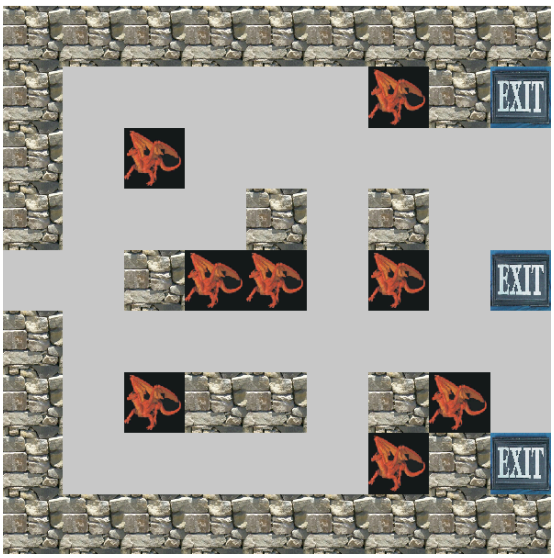
In this problem, you will use value iteration to find the optimal policy of the MDP demonstrated in class. This MDP has $|\mathcal{S}| = 81$ states and $|\mathcal{A}| = 4$ actions, and discount factor $\gamma = 0.99$. Download the ASCII files on the course web site that store the transition matrices and reward function for this MDP. The transition matrices are stored in a sparse format, listing only the row and column indices with non-zero values; if loaded correctly, the rows of these matrices should sum to one.

(a) Compute the optimal state value function $V^*(s)$ using the method of value iteration. Print out a list of the non-zero values of $V^*(s)$. Compare your answer to the numbered maze shown below. The correct value function will have positive values at all the numbered squares and negative values at the all squares with dragons.

(b) Compute the optimal policy $\pi^*(s)$ from your answer in part (a). Interpret the four actions in this MDP as moves to the WEST, NORTH, EAST, and SOUTH. Fill in the correspondingly numbered squares of the maze with arrows that point in the directions prescribed by the optimal policy. Turn in a copy of your solution for the optimal policy, as visualized in this way.

(c) **Turn in your source code along with your answers to the above questions.**





6.4 The unreasonable effectiveness of data

This talk by Peter Norvig, the Director of Research at Google, showcases many real-world applications of the models that we covered in CSE 150. It is optional, but well worth an hour of your time:

`http://www.youtube.com/watch?v=yvDCzhbjYWs`