ECS 170: Problem Set 6

March 11, 2017

This assignment will not be graded! We're providing it to practice for the final.

1. Consider a cumulative discount reward (the objective function of Q-learning) with a $\gamma = 0$ and $\gamma = 1$. What type of behavior would each of these reward functions encourage?

2. We've told you that γ needs to be in $[0, 1]$. Besides potential floating point overflow issues, name a problem that might occur if we set $\gamma >= 1$.

3. Give one reason for why you would prefer to use Q-learning to implement a black-jack player rather than the mini-max algorithm.

4. Why did we learn Q table entries for each action/state combination instead of learning the V function for each state?

5. One assumption that the Q-learning algorithm makes is that whenever an action is taken, an immediate reward is given. Suppose we have a problem where the only positive reward is for a single sequence of actions. For example, a robot could be given a reward if starting at location X it moves 3 spaces to the left, but receives a reward of 0 if it moves only 1 or 2 spaces to the left and then takes some other action. How could you modify the states and actions to transform this problem for Q-learning?

6. Consider the grid in Figure 1. The arrows indicate all possible actions where labels correspond to the reward of the action (unlabeled actions have 0 reward). Apply the Q learning algorithm to this grid, assuming the table of $\hat{Q}$ is initialized to zero. In the table below are the Q values you will update. Note that the updates start in cell $B1$, go clockwise until reaching $B2$, then repeats starting in $B1$. For this problem only fill out the updates in the table - you do not need to present a table of the final action rewards. For this problem let $\gamma = .9$.

Step	Start State	Subsequent State	Update to Q(s,a)
1	B1	A1	Q(B1,Up)=
2	A1	A2	Q(A1,Right)=
3	A2	A3	Q(A2,Right)=
4	A3	B3	Q(A3,Down)=
5	B3	B2	Q(B3,Left)=
6	B1	A1	Q(B1,Up)=
7	A1	A2	Q(A1,Right)=
8	A2	A3	Q(A2,Right)=
9	A3	B3	Q(A3,Down)=
10	B3	B2	Q(B3,Left)=

7. Suppose we run Q-learning twice on the grid in Figure 1, once with $\gamma = .99$ and once with $\gamma = .01$. How will the policies generated by the rewards differ with respect to how the agent reaches B2? Assume the agent starts at B1. Note that we're not asking you to manually run Q-learning, we just want to know how the two policies will qualitatively differ.

8. Suppose now that the values in Figure 1 are not the state-action rewards, but rather the values of a learnt Q-table (where missing values indicate a reward of 0). Construct a policy using these values, assuming the agent starts in cell A2.

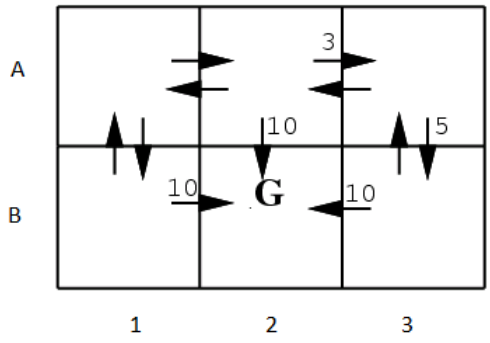

Figure 1: Grid for problems 7-9