CS444 HW #9

Solutions to this assignment should be submitted through Blackboard as .pdf files.

1. Consider the MDP represented by the following simple maze:

A .75			S						B 1.0

The robot in this maze starts at location S. At each time step he has the option of taking one of two actions LEFT, or RIGHT. These actions deterministically move the robot one space to the left or right. When the robot is in state A, he receives a reward of 0.75. When he is in state B he receives a reward of 1.0. Any action taken in states A or B end the trial: the robot exits the maze and no more steps are taken. Assume a discount factor of $\gamma = .9$. As a reminder, here is the equation for the utility of a state under the optimal policy in an MDP:

$$U(s) = R(s) + \gamma \max_a \sum_{s'} P(s'|s,a)U(s')$$

- In each square of the maze, write the utility (expected discounted reward under the optimal policy) of being in that state. States A and B are already completed. (1pt)
- In each square of the maze (other than A and B) indicate the optimal action for the robot to take.

2. Fill out the following tables with the appropriate Q-values for each action in each state. Assume the same reward function and transition model as the previous question. As a reminder, the constraint equation for Q-values is:

$$Q(s,a) = R(s) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q(s',a')$$

A .75|.75	|	|	S |	|	|	|	|	|	B 1.0|1.0

3. For this question, assume that we are using Q-learning to find a policy for the environment described in Question 1. Assume a learning rate of $\alpha = .1$. All Q-values are initialized to 0:

A			S						B
0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0

As a reminder, the update equation for Q-learning is:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

- In the first trial, the agent randomly follows the following action sequence (starting from S):
  ⟨LEFT, RIGHT, LEFT, LEFT, LEFT, LEFT⟩
  What are the estimated Q-values after this trial?

A			S						B
|	|	|	|	|	|	|	|	|	|

- In the second trial, the agent randomly follows the following action sequence (starting from S):
  ⟨LEFT, LEFT, LEFT, LEFT⟩
  What are the estimated Q-values after this trial?

A			S						B
|	|	|	|	|	|	|	|	|	|

- In the third trial, the agent randomly follows the following action sequence (starting from S):
  ⟨LEFT, LEFT, RIGHT, RIGHT, RIGHT, RIGHT, RIGHT, RIGHT, RIGHT, RIGHT, RIGHT⟩
  What are the estimated Q-values after this trial?

A			S						B
|	|	|	|	|	|	|	|	|	|

4. Consider the problem of using a genetic algorithm to search for an optimal policy for the MDP described above.

   - How could you encode policies for this problem as bit-strings? Be specific. How many bits does your encoding take?
   - Describe an appropriate fitness function for this problem.