Predictive Analytics World / Deep Learning World
Exercises - Q-Learning

1. Create a graphical 7x7 maze RL environment with horizontal and vertical walls. Define the observation as the unique integer ID of the current cell. Define actions that try to move in each of the 4 directions. Let the user enter each action manually. Define one goal state. Moving away from the goal state should give a reward of 1.0, and should place the agent at a random location. Output each observation and reward.

2. Connect a random agent to the environment. Output the mean reward received over 10,000 steps. Disable the graphical display for fast evaluation.

3. Modify the agent to use Q-learning with epsilon-greedy exploration, according to the equation below. Set epsilon=0.1, learning rate alpha = 0.1, and discount factor gamma = 0.9. Reevaluate.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \Big]$$

4. Tune these three hyper-parameters to maximize total reward over 10,000 steps.

5. Evaluate your best set of hyper-parameters on the random seed provided by the instructor.

6. Which gives the best definition of cumulative discounted reward?
a. The sum of rewards which have not yet been received.
b. The difference between an actual reward and a potential reward.
c. The sum of rewards multiplied by increasing powers of the discount factor.

7. What is the difference between a value function and a Q function?
a. Both refer to values that depend on states, but Q values also depend on actions.
b. They are the same thing.
c. Q functions can determine the best actions to take, but value functions cannot.

8. What is an example of a violation of the Markov property for states?
a. When it is impossible to leave a particular state.
b. When the reward received upon leaving a state depends on the previous state.
c. When a set of states and their actions all yield zero reward.