Worksheet Week 22

Problems

Q1. Consider the gridworld in figure 1

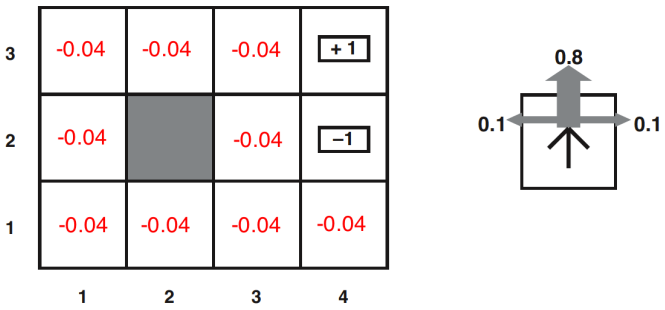

Figure 1: Gridworld

Suppose we wish to use value iteration to compute the utility of each state.

(a) Do we need to wait until the algorithm has converged until we know the utility of each state, or are there some states whose utility we already know?

The equation for value iteration is:

$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|a,s) U_i(s')$$

Since the terminal states have no successors, we already know the utility of the terminal states.

(b) Suppose we initialise the utility of every state to 0, and then perform one iteration of the value iteration algorithm. What is the utility of each state?

Since the utility at step 0 is set to be equal to 0, after one round of value iteration, the utility of each state will be

$$U_1(s) = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|a,s) U_0(s') \quad = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|a,s) * 0 \tag{1}$$

$$= R(s) + 0 = R(s) \tag{2}$$

The utility of each state is equal to the reward at that state

U	U	L	X
U	X	L	X
U	U	L	D

Table 1: Policy after one iteration. The action in (3, 1) could also be R

(c) Suppose we wish to use policy iteration to discover the optimal policy, and suppose our initial policy sets the action in every cell to Up. After one round of policy iteration, what is the resulting policy?

r	-1	10 G
-1	-1	-1
-1	-1	-1

Table 2: 3x3 gridworld

Q2. Consider the 3x3 gridworld shown in table 2. The transition model is as follows: 80% of the time the agent goes in the direction it selects; the rest of the time it moves at right angles to the intended direction, each with a probability of 10%. (i.e., the same as in the previous question).

The r in the top left corner is a reward value. For different values of r, state what policy results. You don't need to use value iteration or policy iteration, you can just work it out from common sense. Use discounted rewards where $\gamma = 0.99$.

(a) $r = -3$

R	R	X
R	R	U
R	R	U

Table 3: Policy with $r = -3$

Here, the agent again tries to reach the goal as fast as possible while attempting to avoid the square (1, 3), but the penalty for square (1, 3) is not so great that the agent will try to actively avoid it at all costs. Thus, the agent will choose to move Right in square (1, 2) in order to try to get closer to the goal even if it occasionally will result in a transition to square (1, 3).

(b) $r = +3$

U	L	X
U	L	D
U	L	L

Table 4: Policy with $r = 3$

Here the agent tries to avoid the goal as best as possible as the reward for staying in square (1,3), with the discount of 0.99 is greater than that of reaching the terminal state. The agent will try to avoid leaving at all costs, hence why it takes the down action in square (3,2) rather than the left action.

Q3. Figure 2 shows a narrow bridge represented as a gridworld environment. A robot starts at the left hand side, in the middle row (marked with a reward of 1). The goal is the middle row on the right hand side, marked with a reward of 10. Squares marked with a reward of -100 are terminal nodes, and represent the robot falling off the bridge. The robot can move one square up, down, left or right. When told to move in a specified direction, it moves in the intended direction with probability 0.8 or at 90 degrees to the intended direction with probability 0.1, or at -90 degrees to the intended direction with probability 0.1.

wall	-100	-100	-100	-100	-100	wall
1	0	0	0	0	0	10
wall	-100	-100	-100	-100	-100	wall

wall	-100	-100	-100	-100	-100	wall
1	-17.28 ←	-30.44 ←	-36.56 →	-25.78 →	-10.8 →	10
wall	-100	-100	-100	-100	-100	wall

Figure 2: (a) rewards for the bridge-crossing problem in gridworld. (b) utilities after 5 iterations, and the corresponding optimal policy

(a) Using a discount value of 0.9, calculate the utility of each non-terminal grid square after one and two moves.

Example calculation for one square:

$$U_1(1) = R(s) + \gamma \max_{a \in A(1)} \sum_{1'} P(1'|a,1)U_0(1') \tag{3}$$

$$= 0 + 0.9 \max_{a \in \{U,D,L,R\}} \sum_{1'} P(1'|a,1)U_0(1') \tag{4}$$

$$\text{Up, Down} : \sum_{1'} P(1'|a,1)U_0(1') = 0.8 * -100 + 0.1 * 1 + 0.1 * 0 = -80 + 0.1 = -79.9 \tag{5}$$

$$\text{Left} : \sum_{1'} P(1'|a,1)U_0(1') = 0.8 * 1 + 0.1 * -100 + 0.1 * -100 = 0.8 - 20 = 19.2 \tag{6}$$

$$\text{Right} : \sum_{1'} P(1'|a,1)U_0(1') = 0.8 * 0 + 0.1 * -100 + 0.1 * -100 = -20 \tag{7}$$

So we see that the action with highest utility is Left (8)

$$U_1(1) = 0 + 0.9 * -19.2 = -17.28 \tag{9}$$

Utilities of the non-terminal squares at each timestep:

0	0	0	0	0
-17.28	-18.	-18.	-18.	-10.8
-17.28	-30.4416	-30.96	-25.776	-10.8

(b) The problem leads to the optimal policy shown in Figure 2b, which fails to cross the bridge. What would be the effect on the policy of decreasing the discount value ?
Decreasing the discount value does not lead to being able to cross the bridge, for example with discount value 0.001 we have

0	0	0	0	0
-0.0192	-0.02	-0.02	-0.02	-0.012
-0.0192	-0.0200154	-0.020016	-0.0200096	-0.012
-0.0192	-0.0200154	-0.020016	-0.0200096	-0.012
-0.0192	-0.0200154	-0.020016	-0.0200096	-0.012

(c) What would be the effect on the policy of increasing the utility of the goal ? Choose a new value for the utility of the goal state so that the optimal policy is to cross the bridge from left to right, and show the utility of each non-terminal grid square after 3 iterations.

If we increase the value of the goal then the agent can succeed in crossing the bridge. For example, setting the goal to 179 enables the agent to cross the bridge, since the agent must move to the first square and thereafter the score should be bigger than 1.

0	0	0	0	0
-17.28	-18.	-18.	-18.	110.88
-17.28	-30.4416	-30.96	61.8336	110.88
-17.28	-30.4416	26.5202	61.8336	110.88
-17.28	1.09454	26.5202	61.8336	110.88
-17.2119	1.09454	26.5202	61.8336	110.88