Homework 4

1. Consider the following grid environment in which (1, 1) is the start state, (3, 4) and (2, 4) are the terminal states. Given the reward value for every non-terminal state R(s) = -.03, the reward values for the terminal states +1 and -1 respectively and the transition model as illustrated below, calculate the utility values of A, B, and C states up to the second iteration. Assume that the discount factor γ = 1 and that the initial utility value of each non-terminal state is zero. Show your calculation to find the updated utility values.

|  |  |  |  |
| --- | --- | --- | --- |
| Episode 1 for A |  | Episode 2 for A |  |
| 1, 3 | -0.03 | 1, 3 | -0.03 |
| 2, 3 | -0.03 | 2, 2 | -0.03 |
| 3, 3 | -0.03 | 3, 3 | -0.03 |
| 2, 3 | -0.03 | 4,3 | 1 |
| 3, 3 | -0.03 |  |  |
| 4, 3 | +1 |  |  |
| Total | 0.85 |  | 0.91 |

|  |  |  |  |
| --- | --- | --- | --- |
| Episode 1 for B |  | Episode 2 for B |  |
| 2, 3 | -0.03 | 2, 3 | -0.03 |
| 3, 3 | -0.03 | 3, 3 | -0.03 |
| 2, 3 | -0.03 | 4,3 | 1 |
| 3, 3 | -0.03 |  |  |
| 4, 3 | +1 |  |  |
| Total | 0.88 |  | 0.94 |

|  |  |  |  |
| --- | --- | --- | --- |
| Episode 1 for C |  | Episode 2 for C |  |
| 3, 3 | -0.03 | 3, 3 | -0.03 |
| 3, 3 | -0.03 | 3, 4 | 1 |
| 3, 4 | +1 |  |  |
| Total | 0.94 |  | 0.97 |

1. Consider the following grid world in which we would like to use TD learning and Q-learning to find the values of these states.

Suppose that we have the following observed transitions: (B, East, C, 2), (C, South, E, 6), (C, East, D, 5), (C, North, A, 4) The initial value of each state is 0. Assume that γ = 1 and α = 0.5.

* What are the learned values from TD learning after all four observations?

B = (1 – 0.5) 0 + 0.5 (2 + (1 \* 0)) = 1

C1 = (1 – 0.5) 0 + 0.5 (6 + (1 \* 0)) = 3

C2 = (1 – 0.5) 3 + 0.5 (5 + (1 \* 0)) = 4

C3 = (1 – 0.5) 4 + 0.5 (4 + (1 \* 0)) = 4

* What are the learned Q-values from Q-learning after all four observations?

B = (1 – 0.5) 0 + 0.5 (2 + (1 \* 0)) = 1

C1 = (1 – 0.5) 0 + 0.5 (6 + (1 \* 0)) = 3

C2 = (1 – 0.5) 0 + 0.5 (5 + (1 \* 0)) = 2.5

C3 = (1 – 0.5) 0 + 0.5 (4 + (1 \* 0)) = 2

1. Programming Assignment I:
2. Run the code:

Text

Description automatically generated

1. Explain the relationship between the class MDP and the class GridMDP in mdp.py.

The GridMDP class transforms objects created in the MDP class into two-dimensional grids for easy understanding.

1. Explain the following python functions from the sample code:
   * Vector add: It adds the values of two vectors together to form one vector.
   * Turn right: It updates the state value of the vector when it turns right.
   * Turn left: It updates the state value of the vector when it turns left.
   * Value iteration: It calculates the reward values with the given discount value and α.
   * Best policy: Returns the best path from the start variable to the goal value.
   * Expected utility: Returns the expected utility values for all states in the transition model.
2. What happens with the found policy if the reward for the entries that have -0.04 are changed to -1, -0.3, or -0.02? Explain the impact of changing rewards on the policy.

The policy for each iteration would be different since the reward values for all states are different, due to the equation having a different value than -0.04. Therefore, the best policy would be changed each time the reward is changed.

1. Text

   Description automatically generatedText

   Description automatically generatedProgramming Assignment II:

The implementations in both the book and the python implementation are similar, the only parameters that change are the number of iterations some agents do in the book compared to the python ones.

1. Reinforcement Learning Applications in Healthcare

This type of implementation of Reinforcement Learning is one of the most important from the website in my opinion. This process could greatly decrease the number of fatalities from misdiagnosis or problems alike. It also could reduce the amount of time spent deducing the sickness for every patient. The states for this example would be the symptoms a patient might experience. The actions the algorithm will take are deciding on what type of treatment a patient might need depending on his/her symptoms severity. The transition function will depend on what type of disease it is trying to treat, and the reward functions would be whether the patient’s symptoms are improving or worsening. One of the algorithms that have been used according to the website is Q-Learning.

1. Pacman
2. What is P(+w), the marginal probability that Pacman will win?

56.6%

1. What is the conditional probability P(+p | + w) that the food he ate was a mini-pellet, given that he won?

(.60 + .80)/2 = 0.7

1. What is the probability of the event (-m, +p, +w, -b), where Pacman eats a mini-pellet and has fresh breath before winning a fight against a nice ghost?

0.80