Alfredo Reina Corona

FINAL GRADE: 75%

COMMENTS:
“task 2: both of these answers are off. why do non terminal rewards get assigned exactly? What does the discount factor do? -10 points task 3: part a) the right direction but wrong answer, left and right are walls but they are not disallowed as choices. part b) this is wrong -15 points”

Task 1:

value_iteration.py 'environment2.txt' -0.04 1 20

0.812, 0.868, 0.918, 1.000

0.762, 0.000, 0.660,-1.000

0.705, 0.655, 0.611, 0.387

>,>,>,o

^,X,^,o

^,<,<,<

value_iteration.py 'environment2.txt' -0.04 0.9 20

0.509, 0.650, 0.795, 1.000

0.399, 0.000, 0.486,-1.000

0.296, 0.254, 0.345, 0.130

>,>,>,o

^,X,^,o

^,>,^,<

Task 2:

a) A small negative value could be used throughout the learning process. This value would discourage any sub-optimal moves and guide the model to only make moves that minimize the negative effect.

b) If you want to have a model that performs as good as possible you need to have it consider both short term and long term terminal states, including the states in which it loses and wins. To achieve this you would need a discount factor closer to 1 so that it prioritizes winning, but not so much so, that it completely disregards immediate dangers. Assuming you can change it mid run, a high discount factor at the start that progressively get lower as the game goes on would be best.

Task 3:

a) “Succeeds with probability 0.8. Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.”

o = open block

o,1,o

x,o,x

o , -1 , o

80% → goes up

20% → goes to blocked state

Assuming it goes up ; U(h)((2,2) ,(3,2)) = −0.04 + 0.9 ∗ 1 = 0.86

Assuming it goes left/right ; left = 0+.9*0 = 0 = right

Assuming it wants to get to the ‘1’ terminal state, it cannot go left or right, so the value is .86

b)

up is optimal when U(2,2) = .86.

U(2,2)>=.86

if U(2,2) goes up → −0.04 + 0.9 ∗ 1 = 0.86

substitute it → U(2,2) >= r+.9 * 1 → .86 >= r+.9 → -.04>=r

knowing that -.04 make the optimal move up, r has to be less than -.04 for ‘up’ to not be optimal