MDL - Assignment-3

ROMICA RAISINGHANI
2021101053

Implementing Value Iteration for MDP

According to value iteration algorithm,

Step 1:
We initialize the utilities of each state to zero.

i.e, $U_0(I) = 0$

$U_0(I)$ → Utility of state I at time $t=0$.

Step 2: Iterate:

According to Bellman Update Equation:

$$U_{t+1}(I) = \max_A \left[R(I,A) + \gamma \sum_J P(J|I,A) \times U_t(J) \right]$$

where,

$R(I, A)$ → immediate reward
OR
Reward on taking action A on state I (step cost)

$P(J|I,A)$ → transition probability on reaching state J from state I upon action A.

$U_t(J)$ → Utility value of state J in the previous iteration

γ → discount factor

We will continue iterating until the values $\{U_t, U_{t+1}\}$ converge.

At the end of iteration, we calculate optimal policy:

$$\text{Policy}(I) = \underset{A}{\text{argmax}} \left[R(I,A) + \gamma \sum_J P(J|I,A) \times U_t(J) \right]$$

On the question given:

Utility of goal state = +1
Utility of penalty state = -1

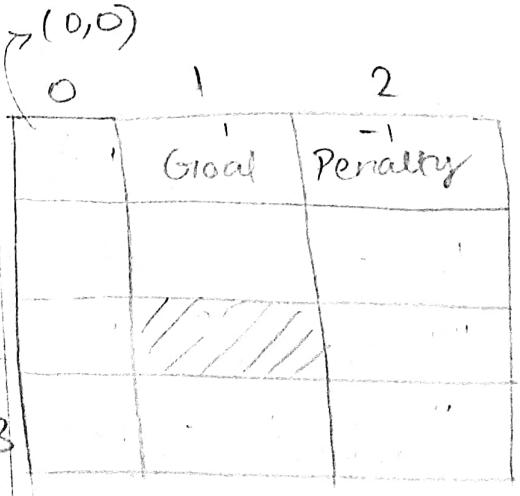
Step cost = -0.04

Discount factor = $\gamma = 0.95$

Prob. of going in the direction of action = 0.7

Prob. of going perpendicular to direction of action = 0.15

Prob. of going opposite to the direction of action = 0



According to algorithm:

Initial utilities of all states

$= U[i][j] = 0$

~~for~~

For (0,0):

North →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

South →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

West/East →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

max of all actions $= -0.04$

$U[0][0] = -0.04$

For (0,1): Goal state

$U[0][1] = 1$

For (0,2): Penalty state

$U[0][2] = -1$

For (1,0):

North →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

South →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

West/East →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

max of all directions $= -0.04$

$U[1][0] = -0.04$

For (1,1):

North →

$-0.04 + 0.95[0.7 \times 1 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04 + 0.665$

$= 0.625$

South →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

West/East →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 1 + 0.15 \times 0]$

$= -0.04 + 0.95 \times 0.15$

$= 0.1025$

$\Rightarrow U[1][1] = 0.625$

For (1,2):

North:

$-0.04 + 0.95[0.7 \times -1 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04 - 0.95 \times 0.7$

$= -0.705$

South:

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

West/East:

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times (-1) + 0.15 \times 0]$

$= -0.04 - 0.95 \times 0.15$

$= -0.1825$

max of all directions $= -0.04$

$U[1,2] = -0.04$

For (2,0):

North →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

South →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

West/East →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

max of all directions $= -0.04$

$U[2][0] = -0.04$

For (2,1): wall

$U[2][1] = 0$

For (2,2):

North →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

South →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

West/East →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

max of all directions $= -0.04$

$U[2][2] = -0.04$

For (3,0):

North →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

South →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

East/west →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

max of all directions $= -0.04$

$U[3][0] = -0.04$

For (3,1):

North →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

South →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

West/East →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15 \times 0]$

$= -0.04$

max of all dir^{n}s $= -0.04$

$U[3][1] = -0.04$

For (3,2):

North →

$-0.04+0.95[0.7\times0+0.15\times0+0.15\times0]$

$= -0.04$

South →

$-0.04+0.95[0.7\times0+0.15\times0+0.15\times0]$

$= -0.04$

West / East →

$-0.04+0.95[0.7\times0+0.15\times0+0.15\times0]$

$= -0.04$

max of all dir^{n}s $= -0.04$

$\boxed{U[3][2] = -0.04}$

ITERATION 2:

For (0,0):

North →

$-0.04+0.95[0.7\times0+0.15\times0+0.15\times1]$

$= -0.04+0.95\times0.15$

$= 0.1025$

South →

$-0.04+0.95[0.7\times(-0.04)+0.15\times0+0.15\times1]$

$= -0.0666$

West →

$-0.04+0.95[0.7\times0+0.15\times0+0.15(-0.04)]$

$= -0.0457$

East →

$-0.04+0.95[0.7\times1+0.15\times0+0.15(-0.04)]$

$= -0.04+0.95[0.7-0.006]$

$= 0.6193$

max of all directions $= 0.6193$

$U[0][0] = 0.6193$

For (0,1): Goal state

$U[0][1] = 1$

For (0,2): Penalty state

$U[0][2] = -1$

For (1,0):

North →

$-0.04+0.95[0.7(-0.04)+0.15\times0+0.15\times0.625]$

$= -0.04+0.95[-0.028+0.09375]$

$= -0.04+0.95[0.6575]$

$= 0.0224625$

South →

$-0.04+0.95[0.7(-0.04)+0.15\times0+0.15\times0.625]$

$= 0.0224625$

West →

$-0.04+0.95[0.7\times0+0.15(-0.04)+0.15(-0.04)]$

$= -0.04-0.0114$

$= -0.0514$

East →

$-0.04 + 0.95[0.7 \times 0.625 + 0.15(-0.04) + 0.15(-0.04)]$

$= -0.04 + 0.95[0.4375 - 0.012]$

$= 0.364225$

max of all directions = 0.364225

$\boxed{U[1][0] = 0.364225}$

FOR (1,1):

North →

$-0.04 + 0.95[0.7 \times 1 + 0.15(-0.04) + 0.15(-0.04)]$

$= 0.62386$

South →

$-0.04 + 0.95[0.7 \times 0 + 0.15(-0.04) + 0.15(-0.04)]$

$= -0.0514$

West →

$-0.04 + 0.95[0.7(-0.04) + 0.15 \times 1 + 0.15 \times 0]$

$= 0.0759$

East →

$-0.04 + 0.95[0.7(-0.04) + 0.15 \times 1 + 0.15 \times 0]$

$= 0.0759$

max of all directions =

$\boxed{U[1][1] = 0.62386}$

For (1,2):

North →

$-0.04 + 0.95[0.7(-1) + 0.15 \times 0 + 0.15(0.625)]$

$= -0.6159375$

South →

$-0.04 + 0.95[0.7(-0.04) + 0.15 \times 0 + 0.15 \times 0.625]$

$= 0.0224625$

West →

$-0.04 + 0.95[0.7(0.625) + 0.15(-1) + 0.15(-0.04)]$

$= 0.227425$

East →

$-0.04 + 0.95[0.7 \times 0 + 0.15(-1) + 0.15(-0.04)]$

$= -0.1882$

max of all directions = 0.227425

$\boxed{U[1][2] = 0.227425}$

FOR (2,0):

North →

$-0.04 + 0.95[0.7(-0.04) + 0]$

$= -0.0666$

South →

$-0.04 + 0.95[0.7 \times (-0.04) + 0]$

$= -0.0666$

West/East →

$-0.04 + 0.95[0.7 \times 0 + 0.15(-0.04) + 0.15(-0.04)]$

$= -0.0514$

max of all directions $= -0.0514$

$U[2][0] = -0.0514$

For (2,1): wall

$\boxed{U[2][1] = 0}$

For (2,2):

North →

$-0.04 + 0.95[0.7 \times (-0.04) + 0]$

$= -0.0666$

South →

$-0.04 + 0.95[0.7(-0.04) + 0]$

$= -0.0666$

West/East →

$-0.04 + 0.95[0.7 \times 0 + 0.15(-0.04) + 0.15(-0.04)]$

$= -0.0514$

max of all directions $= -0.0514$

$\boxed{U[2][2] = -0.0514}$

For (3,0):

North →

$-0.04 + 0.95[0.7(-0.04) + 0 + 0.15(-0.04)]$

$= -0.0723$

South →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15(-0.04)]$

$= -0.0457$

West →

$-0.04 + 0.95[0.7 \times 0 + 0.15 \times 0 + 0.15(-0.04)]$

$= -0.0457$

East →

$-0.04 + 0.95[0.7(-0.04) + 0.15(-0.04) + 0.15 \times 0]$

$= -0.0723$

max of all directions $= -0.0457$

$\boxed{U[3][0] = -0.0457}$

For (3,1):

North →

$-0.04 + 0.95[0.7 \times 0 + 0.15(-0.04) + 0.15(-0.04)]$

$= -0.0514$

South →

$-0.04 + 0.95[0.7 \times 0 + 0.15(-0.04) + 0.15(-0.04)]$

$= -0.0514$

West/East →

$-0.04 + 0.95[0.7(-0.04) + 0 + 0]$

$= -0.0666$

max of all directions $= -0.0514$

$\boxed{U[3][1] = -0.0514}$

For (3,2):

North →

$-0.04 + 0.95[0.7(-0.04) + 0.15(-0.04) + 0]$

$= -0.0723$

South →

$-0.04 + 0.95[0.7 \times 0 + 0.15(-0.04) + 0]$

$= -0.0457$

~~East~~ West →

$-0.04 + 0.95[0.7(-0.04) + 0.15(-0.04) + 0]$

$= -0.0457$

$= -0.0723$

East →

$-0.04 + 0.95[0.7 \times 0 + 0.15(-0.04) + 0]$

$= -0.0457$

max of all directions $= -0.0457$

$u[3][2] = -0.0457$

Matrix after Iteration 1 :

$$\begin{bmatrix} 0.62499 & 1 & -1 \\ -0.04 & 0.6249.. & -0.04 \\ -0.04 & 0 & -0.04 \\ -0.04 & -0.04 & -0.04 \end{bmatrix}$$

Matrix after Iteration 2 :

$$\begin{bmatrix} 0.70836 & 1 & -1 \\ 0.4589 & 0.6135 & 0.2274 \\ -0.0779 & 0 & -0.0779 \\ -0.0779 & -0.0779 & -0.0779 \end{bmatrix}$$

Both iteration values matches with the code in python.

Hence, the algorithm works by manual calculation as well as on code.

There are 22 iterations taken for the algorithm to converge.