Planing, Learning and Intelligent Decision Making - Homework 2

99326 - Sebastião Carvalho, 99331 - Tiago Antunes

March 5, 2024

Contents

1 Question 1

1 a)

Considering $\mathcal{X}$ as our state space, $\mathcal{X} = \{1P, 2P, 2NP, 3P, 3NP, 4P, 4NP\}$, where each state represents the corresponding cell and if they have the passanger or not (P or NP). Since when we are on the first cell we always have the passanger, state we dont consider $1NP$.

Considering $\mathcal{A}$ as our action space, $\mathcal{A} = \{U, D, L, R\}$, where each action represents the movement of the agent, up, down, left and right, respectively.

1 b)

The transition matrix for the action down, P_D, is given by $\begin{bmatrix} 0.2 & 0 & 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0.2 & 0 & 0 & 0 & 0.8 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$

, where each row represents the state we are in and each column represents the state we are going to.

Since we want a simple cost function, we'll consider that the movement to each state costs 1, and the movement to the goal state ($2P$) costs 0. So, the cost matrix for all actions is given by $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$

, where each row represents the state we are in and each column represents the action we are taking.

Basically, $c(x,a) = \begin{cases} 0 & \text{if } (x,a) = (4P,U),(2P,U),(2P,R),(2P,L) \\ 1 & \text{else} \end{cases}$

1 c)

Let policy, π, be the policy in which the taxi driver always goes down. In this case there's only one action the taxi can do, action 'D', therefor the policy matrix can be represented by:

$$\pi = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

The policy is Markov and stationary, since the distribution over action given the history depends only on the last state and doesn't change through time.

Because the policy is stationary, the cost-to-go function associated with it, verifies:

$$J^\pi = c_\pi + \gamma P_\pi J^\pi$$

Solving the system, we get:

$$J^\pi = (I - \gamma P_\pi)^{-1} c_\pi$$

The cost matrix for the policy π can be obtained by multiplying the policy matrix with the immediate cost matrix value by value, and then summing the values in each row. Since here all columns in π are zeros, except the column corresponding to action 'D', we only need to calculate for that action, since all else will simply be 0. Here we put the column corresponding to action 'D' in the policy matrix in a diagonal and multiply both matrices to get the result.

$$c_\pi = diag(\pi(D|\cdot))C(D|\cdot) \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Now we will calculate the matrix P_π. This matrix is obtained by multiplying each value of a column by each row of the transition probability matrix of the action the column corresponds to. In other words, the probability of doing an action in a state by the transition probabilities of that state, and then summing the matrices corresponding results for each action. Since here all columns in π are zeros, except the column corresponding to action 'D', we only need to calculate for that action, since all else will simply be 0. Here, once again, we put the column corresponding to action 'D' in the policy matrix in a diagonal and multiply both matrices to get the result.

$$P_\pi = diag(\pi(D|\cdot))P_D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.2 & 0 & 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0.2 & 0 & 0 & 0 & 0.8 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.2 & 0 & 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0.2 & 0 & 0 & 0 & 0.8 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Now, we have everything we need to calculate J_π. Considering $\gamma = 0.9$ we have:

$$J^\pi = \left(\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.2 & 0 & 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0.2 & 0 & 0 & 0 & 0.8 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1.22 & 0 & 0 & 8.78 & 0 & 0 & 0 \\ 0 & 1.22 & 0 & 0 & 0 & 8.78 & 0 \\ 0 & 0 & 1.22 & 0 & 0 & 0 & 8.78 \\ 0 & 0 & 0 & 10 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 10 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 10 & 0 \\ 0 & 0. & 0 & 0 & 0 & 0 & 10 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \end{bmatrix}$$

So, the cost-to-go function associated with policy π, using $\gamma = 0.9$, is:

$$J^{\pi} = \begin{bmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \end{bmatrix}$$