EE932 Assignment-2 Solution

---

eMasters in Communication Systems, IITK
EE932: Introduction to Reinforcement Learning
Instructor: Prof. Subrahmanya Swamy Peruru
Student Name: Venkateswar Reddy Melachervu
Roll No: 23156022

Question 11:

		+5
S		-5

Figure 1: Grid-World

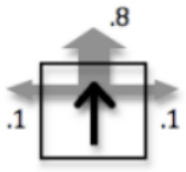

Figure 2: Tranistion probabilities for an 'UP' action

Consider the grid shown in Figure 1. The states are grid squares, identified by their row and column numbers. The agent always starts in the bottom left state (1,1), marked with the letter S. (Note that the bottom row is denoted by number 1, the top row by number 2). There are two terminal goal states, (2,3) with reward +5 and (1,3) with reward -5. Rewards are 0 in non-terminal states. (The reward for a state is received as the agent moves into the state.) The transition function is such that the intended agent movement (UP, Down, Left, or Right) happens with probability 0.8. With probability 0.1 each, the agent ends up in one of the states perpendicular to the intended direction. Please refer Figure 2.

Assume $\gamma = 1$, and make an intelligent guess for the optimal policy for this grid

Solution

Intuitively, the optimal policy for the agent to maximize the reward/return is:

- From the starting state (1,1), the agent should move up to the state (2,1) or right to (1,2).
  - This is because the intended movement up has a higher probability 0.8 of reaching the state (1,2) or (2,1) which are closer to the positive reward state (2,3).
- From the state (1,2), the agent should move up to the state (2,2)
  - This is because this intended movement has a higher probability 0.8 of reaching the state (2,2) which is still closer to the positive reward state (2,3).
- From the state (2,1), the agent should move right to the state (2,2).
  - This is because this intended movement has a higher probability 0.8 of reaching the state (2,2) which is still closer to the positive reward state (2,3).
- From the state (2,2), the agent should move right to the state (2,3), the positive reward terminal state with a reward of +5.

This policy maximizes the expected cumulative reward for the agent, as it takes the path with the highest probability of reaching the positive reward state (2,3) while avoiding the negative reward state (1,3)

---------- End of the Document ----------

C-501, Salarpuria Serenity, 5th Main, Sector 7, HSR Layout, Bengaluru 560102 KA India
Mobile: +91 97012 22130, Email: vmelachervu@gmail.com, vmela23@iitk.ac.in, Website: www.linkedin.com/in/vmelachervu