CS419m Practice Questions RL

November 14, 2018

1 Questions

1. **Modelling a problem as an MDP**

   Consider an agent in a 5x5 grid. Let the cells of the grid be denoted by c_{ij}, with $i, j \in \{0, 1, 2, 3, 4\}$. From any position in the grid, the agent must reach the target cell c_{44} in as few steps as possible. The agent can take only one step at a time either to the cell on its immediate left, right, up or down position. Suppose that the grid is slippery, so even if the agent chooses to go immediate left, then with probability 0.99 it will move to the cell on its left but with probability 0.01 it can end up going immediate right, up or down. If the agent tries to move to a non-existent cell, then it will remain in its current cell with probability 0.99 and go to a neighbouring cell with probability 0.01.

   Consider modelling this problem as a Markov Decision Problem (MDP), with discount factor $\gamma \in (0, 1)$.

   (a) Would you model this as a continuing or an episodic task? (0.5)

   (b) How many states will this MDP have? (1)

   (c) How many actions can the agent perform from a given state? (0.5)

   (d) What is the reward you will give when the agent reaches the target cell? What about any other cell? (1)

   Now suppose we set the discount factor γ to exactly 1.

   (e) What changes, if any, will you need to make to your reward scheme to encourage the agent to get to the target cell in as few steps as possible? (1)

   Suppose a policy π is a mapping from a state to an action.

   (f) How many possible policies are there for this MDP? (1)

   (g) How many optimal policies are there for this MDP? (2)

   (h) Why is the slippery assumption needed? What happens without this assumption? (1)

2 Solutions

1. Solutions to Question 1

   (a) Episodic Task. The episode ends on reaching the target cell.

   (b) Need to keep track of the position of the agent on the grid. So $5 \times 5 = 25$

   (c) 4 actions - left, right, up, down

   (d) Reward 1 on reaching target. 0 on reaching any other state.

   (e) Reward 10 on reaching target, -1 on reaching any other state. Any **negative** reward for non-goal state will do.

   (f) 4 possible actions for every state and 25 states, so 4^{25} policies.

   (g) Either go down or right from any state which is not at the bottom or to the far right. From the bottom row you must only go right, and from the right column, you must only go down. Hence 2 optimal actions for 16 states and only 1 optimal action for the 8 states on the bottom or right. At the goal state the episode ends, so does not matter which action - hence all 4 actions are optimal for that state. Hence total optimal policies $= 2^{16}.4$

   (h) Without this assumption, we will have policies which can make the agent loop in the grid. With this assumption the agent is guaranteed to eventually reach the target cell with probability 1.