Reinforcement Learning

Quiz 1

11/10/2022

Sanjit K. Kaul

Instructions: You have an hour and twenty minutes to work on the questions. Answers with no supporting steps will receive no credit. No resources, other than a pen/pencil, are allowed. In case you believe that required information is unavailable, make a suitable assumption.

Question 1. 50 **marks** We have three lanes 1, 2, and 3. In any lane, an agent may choose to stay in it or change its lane. In lane 1, a lane change by the agent results in the agent entering lane 2 with probability 0.6 and lane 3 with probability 0.4. The agent receives a reward of -2 for a lane change to 3 and a reward of -1 for a lane change to 2. In case an agent in lane 1 chooses to stay in 1, the agent gets a reward of 1. In lane 2, a lane change by the agent, results in the agent entering lane 1 or lane 3 with equal probability and obtaining a reward of -1. In case an agent chooses to stay in 2, the agent gets a reward of 1. In lane 3, a lane change by an agent results in the agent entering lane 2 with probability 0.6 and lane 1 with probability 0.4. For the former, the agent gets a reward of -1 and for the latter it gets a reward of -2. In case an agent chooses to stay in lane 3, the agent gets a reward of 2. Answer the following questions.

1) Draw the MDP.
2) Begin with a policy that chooses actions with equal probability in every state. Carry out policy iteration. Show four iterations or fewer in case you arrive at the optimal policy sooner. State your obtained policy and its value function on completion of the iterations. Explain whether your policy is the optimal policy. Assume a discount factor of 0.5.

Question 2. 30 **marks** Write down five different episodes, each ten time steps long, which are valid given the MDP in Question 1. Your episodes must include information of states, actions, and rewards. Also, each episode must have all states visited. Assume every visit Monte Carlo. After every episode, calculate your estimates of all visited state-action pairs and calculate the corresponding improved policy. Assume $\gamma = 1$ and $\alpha = 0.5$. Choose all initial state and action values to be 0.

Question 3. 20 **marks** Suppose we have a 100 armed bandit testbed with arms $1, 2, \ldots, 100$. Further let arm i have a reward distribution that is Gaussian with mean 1 and variance i^2. Which arms maximize the expected reward? Suppose you had 10 attempts and you could choose any arm during each of the 10 attempts. Assume you know the reward distributions of the arms. Which arms would you choose and why?