National University of Singapore
School of Computing

Semester 1, AY2021-22

CS4246/CS5446 AI Planning and Decision Making

Issued: 1 Oct, 2021

Tutorial Week 9: Reinforcement Learning

Guidelines

You may discuss the content of the questions with your classmates. But everyone should work on and be ready to present ALL the solutions. Note: Materials in this Tutorial will not be covered in the Mid-Term Exam on 8 October 2021

Problem 1: ADP and TD Learning

Consider an agent starting in a room A in which it can take two possible actions: to leave the room (action 'L') or to stay (action 'S'). If it leaves A, the agent moves to room B, which is a terminal state (no more actions can be taken). The outcomes of the actions are uncertain, so that when executing action L (or action S), there is some probability that the agent will leave A (or stay in A). We assume that the reward in entering state B is $R(B) = 1$ and the reward for being in state A is $R(A) = -0.1$.

1. Assume that actions L is more likely to succeed than not, and similarly action S is also more likely to succeed than not. What is the optimal policy π^*?

2. Assume that the agent knows neither the transition function nor the utilities of the states. Assume that the agent, for some reason, happens to follow the optimal policy π^*. The rewards received at states A and B are the same as described above. In the process of executing this policy, the agent executes four trials and, in each trial, it stops after reaching state B. The following state sequences are recorded during the trials: $AAAB$, AAB, AB, AB. What is the estimate of $T(.,.,.)$? Using ADP, what is the estimate of $U^{\pi^*}(A)$, assuming a discount factor of $\gamma = 0.5$?

3. Assume now that the agent is executing only one trial yielding the sequence of states AAB. Compute the estimate of the utility $U^{\pi^*}(A)$ using TD learning. Use discount $\gamma = 0.5$ and learning rate $\alpha = 0.5$. To know the starting values for U^{π^*}, refer to the TD learning algorithm in the lecture notes.

Problem 2: Q-Learning

Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

1. s_1, $R(s_1) = -10$, a_1, s_1
2. s_1, $R(s_1) = -10$, a_2, s_2
3. s_2, $R(s_2) = 20$, a_1, s_1
4. s_1, $R(s_1) = -10$, a_2, s_2

What is the policy derived from the Q-function at this point?

Problem 3: SARSA and Q-Learning

Consider using SARSA and Q-learning to learn a policy in an MDP with two states s_1 and s_2 and two actions a and b. Assume that $\gamma = 0.8$ and $\alpha = 0.2$, and that the current values of Q are:

Q	s_1	s_2
a	2	4
b	2	2

Suppose that, when we were in state s_1, we took action b, received reward 1 and moved to state s_2 and take action b there. Which item of the Q-table will change and what is the new value? Compute for both SARSA and Q-learning.