National University of Singapore
School of Computing

Semester 1, AY2023-24

CS4246/CS5446 AI Planning and Decision Making

Tutorial Week 8: MDP and RL

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.

Problem 1: Online Search for Markov Decision Process

Consider an MDP where the state is described using M variables where each variable can take n values. The MDP has 2 actions and at each state each action can only lead to 2 possible next states.

a) What is the size of the state space of this MDP? Can this MDP be efficiently solvable with value iteration as M grows?

Solution:
States space size is n^M. Value iteration is not efficient as M grows as runtime will be exponential in M.

b) A search tree of depth D (number of actions from the root to any leaf is D) is constructed from an initial state s. What is the size of the search tree (the number of nodes and edges) as a function of M and D, in O-notation? Can online search be done efficiently as M grows if D is a fixed small constant?

Solution:
The search tree size is $O(2^{2D})$. If D is a small fixed constant, then online search is efficient as the size of the search tree is constant as M grows (although the computation at each node will still grow at least linearly with M for representing the state).

c) MCTS is used for solving this MDP. What is the size of the search tree if T trials of MTCS is performed up to a search depth of D, as a function of M, D and T in O-notation?

Solution:
Each trial contributes at most T nodes and edges to the search tree, so the size is $O(DT)$.

d) Consider a search tree where the reward is zero everywhere except at the leaves. When a MCTS trial goes through a node, we say that an action at the node wins if the trial ends in a leaf with reward 1. Consider an MCTS simulation where a node has been visited 16 times and has two actions, A and B. Action A has a won 2 out 4 times whereas action B has won 8 out of 12 times. Which action will the MCTS algorithm chose given the exploration parameter c is set to 1? Give the values of π_{UCT} for the node (consider log base 2 in UCT bound).

Solution:
Node A. $\pi_{UCT}(n) = \underset{a}{\text{argmax}}\Big(\hat{Q}(n,a) + c\sqrt{\frac{\log(N(n))}{N(n,a)}}\Big)$. UCT function value for action A is $\frac{2}{4} + \sqrt{\frac{\log 16}{4}} = 1.5$ and for action B is $\frac{8}{12} + \sqrt{\frac{\log 16}{12}} = 1.244$, so $\pi_{UCT}(n) = 1.5$.

Problem 2: 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 $R(A) = -0.1$.

a 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^*?

Solution:
$\pi^*(A) = L$.

b 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 $T(\cdot,\cdot,\cdot)$? Using ADP, what is the estimate of $U^{\pi^*}(A)$, assuming a discount factor of $\gamma = 0.5$?

Solution:
$T(A, L, A) = 3/7$ and $T(A, L, B) = 4/7$.
Note that $T(A, S, A)$ and $T(A, S, B)$ cannot be computed from the data given in the text and they are not needed since we assume that we follow the optimal policy.

$U^{\pi^*}(A) = R(A) + \gamma\,(T(A, L, A)\,U^{\pi^*}(A) + T(A, L, B)\,U^{\pi^*}(B))$
$U^{\pi^*}(A) = -0.1 + 0.5 \times (3/7 \times U^{\pi^*}(A) + 4/7 \times 1)$
$11/14 \times U^{\pi^*}(A) = -0.1 + 4/14$
$U^{\pi^*}(A) = 26/110 = 0.2364.$

c 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 $\alpha = 0.5$. Use the reward as the starting value of U^{π^*} in your calculation.

Solution:

Transition A to A:

$$\begin{aligned} U^{\pi^*}(A) \leftarrow \;& U^{\pi^*}(A) + \alpha(R(A) + \gamma U^{\pi^*}(A) - U^{\pi^*}(A)) \\ =\;& -0.1 + 0.5 \times (-0.1 + 0.5 \times -0.1 - (-0.1)) = -0.125 \end{aligned}$$

Transition A to B:

$$\begin{aligned} U^{\pi^*}(A) \leftarrow \;& U^{\pi^*}(A) + \alpha(R(A) + \gamma U^{\pi^*}(B) - U^{\pi^*}(A)) \\ =\;& -0.125 + 0.5 \times (-0.1 + 0.5 \times 1 - (-0.125)) = 0.1375 \end{aligned}$$

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.

Solution:

$Q(s_1, b)$ is the affected entry.

For SARSA,

$$\begin{aligned} Q(s_1, b) \leftarrow \;& Q(s_1, b) + \alpha(R(s_1) + \gamma Q(s_2, b) - Q(s_1, b)) \\ =\;& 2 + 0.2 \times (1 + 0.8 \times 2 - 2) = 2.12 \end{aligned}$$

For Q-learning,

$$\begin{aligned} Q(s_1, b) \leftarrow \;& Q(s_1, b) + \alpha(R(s_1) + \gamma \max_{u \in \{a,b\}} Q(s_2, u) - Q(s_1, b)) \\ =\;& 2 + 0.2 \times (1 + 0.8 \times 4 - 2) = 2.44 \end{aligned}$$