Machine Learning and Artificial Intelligence - COS 402

Written Homework Assignment 5

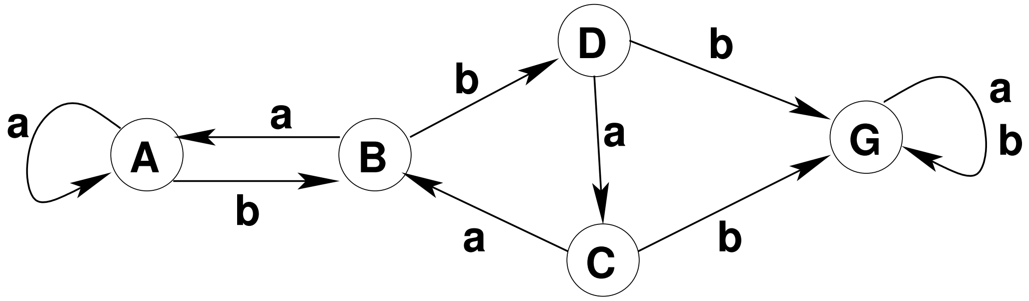
Due Date: two classes from the announcement in class, due in class

(a) **Consulting other students from this course is allowed. In this case - clearly state whom you consulted with for each problem separately.**

(b) **Searching the internet or literature for solutions is NOT allowed.**

(c) **Submit your homework in separate pages for the different questions, each including your name and email address (this is to help the graders). Typing solutions up is strongly advised.**

1. [15] Consider the following MDP:



There are five states: A, B, C, D and G. The reward at every state is -1, except at G where the reward is 0. There are two actions, a and b, and the effect of each action is deterministic as indicated in the figure. For instance, executing a in state B leads to state A. Assume $\gamma = 1$ in this problem.

[Note: If you understand the algorithms, this problem can (and should) be solved without a lot of tedious calculations, and without the use of a computer or even a calculator. You do not need to show easy calculations in detail, but should nevertheless justify your reasoning.]

(a) Show the sequence of utility estimates U_i that would result from executing value iteration on this MDP. Also show the optimal policy that is computed using the final utility estimate.

(b) Show the sequence of policies π_i and corresponding utility functions U^{π_i} that would result from executing policy iteration on this MDP. Assume that you start with a policy that assigns action a to every state. The utility functions U^{π_i} should be computed exactly; note that these utilities may be infinite for some states. Also, assume that all ties between the actions a and b in the policy improvement step are always broken in favor of a.

(c) Generalizing this example, suppose we are given a graph with a distinguished node (i.e., state) G, and k edges emanating from every node corresponding to k (deterministic) actions. As in this example, all of the edges emanating from G are self-loops, the node G is assigned reward 0, and all other nodes are assigned reward -1. In terms of properties of the graph, what is the optimal utility function U^*, and what is the optimal policy π^*? If value iteration is applied to this graph (viewed as an MDP), exactly how many iterations will be needed until the algorithm converges? How about for policy iteration?

2. [10] Sometimes MDP's are formulated with a reward function $R(s, a)$ that depends on the action taken (so that reward $R(s, a)$ is received when action a is executed from state s). For each of these formulations, show how to appropriately modify each of the following:

- the Bellman equation:

$$v(s) = R(s) + \gamma \max_{a \in A} \sum_{s'} P_{ss'a} v(s')$$

- the formula for converting the optimal utility v^* into an optimal policy π^*

$$\pi^*(s) = \arg\max_{a \in A} \sum_{s'} P_{ss'a} v^*(s')$$

- the value iteration algorithm;
- the policy iteration algorithm.

3. [15] Let $T(v)$ and $||\cdot||_\infty$ be as defined in class. The purpose of this exercise is to complete the proof, whose sketch we have seen in class, that T is a *contraction*, i.e., that $||T(v) - T(v')||_\infty \leq \gamma ||v - v'||_\infty$. As discussed in class, this is the key step in showing that value iteration converges to the right answer.

We will begin by proving some basic facts. Be sure to give genuine mathematical proofs for each part of this problem. Also, your proofs should use elementary facts — in other words, do not give proofs that rely on mathematical sledge-hammers like the Cauchy-Schwartz inequality.

(a) Let $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ be any sequences of real numbers. Prove that if $u_i \leq v_i$ for all i then

$$\max_i u_i \leq \max_i v_i.$$

(b) Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be any sequences of real numbers. Prove that

$$\left(\max_i x_i\right) - \left(\max_i y_i\right) \leq \max_i (x_i - y_i),$$

and also that

$$\max_i (x_i - y_i) \leq \max_i |x_i - y_i|.$$

(Hint: both of these inequalities can be proved using part (a) for an appropriate choice of u_i and v_i.)

Finally, use these facts to prove that

$$\left|\left(\max_i x_i\right) - \left(\max_i y_i\right)\right| \leq \max_i |x_i - y_i|.$$

(c) Let $x_1, \ldots, x_n$ be any real numbers, and suppose that $p_1, \ldots, p_n$ are nonnegative real numbers such that $\sum_i p_i = 1$. Use the fact that $|a + b| \leq |a| + |b|$ for any real numbers a and b to prove that

$$\left|\sum_i p_i x_i\right| \leq \max_i |x_i|.$$

(d) Now let s be any state, and let $(T(v))(s)$ denote the value of $T(v)$ at state s. By plugging in the definition of T, and using the properties proved above, prove that

$$|(T(v))(s) - (T(v'))(s)| \leq \gamma \|v - v'\|_\infty.$$

Conclude that

$$\|T(v) - T(v')\|_\infty \leq \gamma \|v - v'\|_\infty.$$

4. [15] This exercise asks you to prove the *policy improvement theorem* which, as discussed in class, is the basis for proving that policy iteration is an effective method for finding an optimal policy. (As a side note, and as mentioned in class as a bonus question, the theorem can also be used to prove the *existence* of an optimal policy π^*, that is, a policy that is optimal for all states simultaneously.)

Let π be any policy, and let π' be the result of applying the policy improvement step of policy iteration. That is, for all states s,

$$\pi'(s) = \arg\max_a \sum_{s'} P(s'|s,a)\ v^\pi(s'),$$

where, as usual, the "$\arg\max$" returns any action a that realizes the maximum of the value on the right.

We make the usual assumptions that the number of states and number of actions are both finite, that $\gamma < 1$, etc.

Let us define the following functions $v_k(s)$ defined over states s. The first of these v_0 is identical to v^π so that $v_0(s) = v^\pi(s)$ for all s. And for $k \geq 1$, and for all s, we define

$$v_k(s) = R(s) + \gamma \sum_{s'} P(s'|s,\pi'(s))\ v_{k-1}(s').$$

(a) Prove by induction on k that $v_k(s) \geq v^\pi(s)$ for all states s and for all $k \geq 0$.

(b) Prove that $\|v_k - v^{\pi'}\|_\infty \to 0$ as $k \to \infty$.

(c) Combine parts (a) and (b) to prove that $v^{\pi'}(s) \geq v^\pi(s)$ for all states s. This shows that policy iteration can only produce policies that are at least as good as the preceding policy at every state.

(d) Prove that π is an optimal policy if and only if $v^{\pi'}(s) = v^\pi(s)$ for all states s. This implies that if π is not already optimal, then each policy improvement step will lead to a new policy that is strictly better than the last one for at least one state.