CSC 448/548 Advanced Topics in Artificial Intelligence
Deep Reinforcement Learning
Fall, 2022

Programming Assignment 1

Your task is to implement General Policy Iteration on a Gridworld problem. The Gridworld that you are to implement is a variation of the one described in the textbook and lectures. There are 16 states, and the transition probability function $\mathrm{p}(s'|sa)$ is deterministic. You may have seen $\mathrm{p}(s'|s,a)$ given as $\mathrm{p}(s,a,s')$. These notations are, for our purposes, equivalent. The following figure shows the state numbers:

0	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

The dynamics of this variation of the Gridworld are described as follows:

- States 8 and 15 are special.
  - State 8 is a "magic teleporter", but using it is expensive. In state 8, the LEFT action takes you to state 15 with probability 1 and immediate reward -2.
  - In state 15, the RIGHT and DOWN actions take you back to state 15 with probability 1 and immediate reward 0.
- In all states other than states 8 and 15, choosing an action that would move you off the grid will move you back into the original state with a reward of -1.
- In all remaining cases, choosing an action will move you to an adjacent state within the grid with probability 1 and an immediate reward of -1. For example, choosing the DOWN action in state 5 will take you to state 9 with probability 1 and immediate reward -1.

Feel free to ask if you need any further clarification.

Undergraduate Students

Write a program that calculates and prints out one optimal deterministic policy for this variation of the Gridworld, and the value function $v_*(s)$ for that policy.

Graduate Students

Write a program that

- calculates and prints out one optimal deterministic policy for this variation of the Gridworld and the value function $v_*(s)$ for that policy,
- calculates and prints out an optimal stochastic policy (if one exists) and the value function $v_*(s)$ for that policy, and
- reports the number of optimal deterministic policies that exist for this problem.

Hints

In the slides, the reward function is given as r(s), which is the expected immediate reward for being in state s. His notation is sloppy. I think that he should have used r(s'), which would be the reward for being in state s'. In practice, it may be better to use the reward function r(s, a), which is the expected immediate reward for taking action a in state s. For example, the reward for taking the LEFT action in state 8 is -2, and the reward for taking the RIGHT or DOWN actions in state 15 is zero. The reward for all other actions in any state is -1. That is pretty easy to code.

You may use any language that you prefer. but COBOL is probably not a good choice.