RLAI

Enigma Dimitri

October 29, 2024

1 Introduction

Exercise 1.1: Self-Play Suppose, instead of playing against a random opponent, the reinforcement learning algorithm described above played against itself, with both sides learning. What do you think would happen in this case? Would it learn a different policy for selecting moves?

Answer

It would of course learn a different policy for selecting movies because the behavior of the opponent would totally not be the same. Now what I think would happen is that both sides would learn how to counter the other and pretty much all runs would end up in ties. □

Exercise 1.2: Symmetries Many tic-tac-toe positions appear different but are really the same because of symmetries. How might we amend the learning process described above to take advantage of this? In what ways would this change improve the learning process? Now think again. Suppose the opponent did not take advantage of symmetries. In that case, should we? Is it true, then, that symmetrically equivalent positions should necessarily have the same value?

Answer

In theory we might think that if we code the symmetries to be the exact same state and change the actions accordingly then we would probably need way less examples to actually learn how to play the game. It would also make the algorithm slightly better because in average the best cases will also have more examples. Now in practice if the opponent do not use the symmetries then we should also not use them as his game may be rigged toward a particular axis. In this sense, if we abuse symmetries, we may miss which axis is the weakest for the opponent. Hence, in this case, symmetrically equivalent positions should not have the same value by default. □

Exercise 1.3: Greedy Play Suppose the reinforcement learning player was *greedy*, that is, it always played the move that brought it to the position that it rated the best. Might it learn to play better, or worse, than a nongreedy player? What problems might occur?

Answer

It would be bad for the algorithm. Indeed, convinced that it's moves are always the best, it would never know if some others may actually lead to a better reward. The algorithm would be stuck in a confirmation bias and actually never truly see what is the objective best maneuver. The nongreedy player would at the opposite explore different paths and sometimes find new solutions that are actually more efficient. □

Exercise 1.4: Learning from Exploration Suppose learning updates occurred after *all* moves, including exploratory moves. If the step-size parameter is appropriately reduced over time (but not the tendency to explore), then the state values would converge to a different set of probabilities. What (conceptually) are the two sets of probabilities computed when we do, and when we do not, learn from exploratory moves? Assuming that we do continue to make exploratory moves, which set of probabilities might be better to learn? Which would result in more wins?

Answer

I am going to wildly guess on this one. If we learn from every moves then in theory I guess we are going to learn the set of probabilities that are best because the influence of exploratory moves will fade over time (as the step-size is reduced). This will still be less optimal than the standard case because we'll take into account bad moves done by exploration. If we keep making exploratory moves, probably the classic probabilities will be better because we'll not use exploratory moves (which are bad) to influence the actual important decisions. □

Exercise 1.5: Other Improvements Can you think of other ways to improve the reinforcement learning player? Can you think of any better way to solve the tic-tac-toe problem as posed?

Answer

I think of two things right now, the first one is to decrease the exploration probability over time because if the opponent doesn't change his strategy, at some point we will not need to explore anything anymore because he will not adapt to our learning. The second thing is to use intermediate rewards to not only reward the whole game sequence but specific states of the game leading toward the win. □
