Machine Learning - CS 7641 Assignment 4

Anthony Menninger
Georgia Tech OMSCS Program
amenninger3
tmenninger@gatech.edu

Abstract

This paper explores Markov Decision Processes through the use of Value Iteration, Policy Iteration and Q Learning. It looks at a simple "Forest" MDP and a larger Frozen Lake "Grid World".

Problem Introduction

A key element in both Markov Decision Processes (MDPs) is the idea of a discount, which is often referred to as ***Gamma*** (γ). This discount factor, between 1 and 0, allows solutions for infinite MDPs by discounting future value by ***Gamma*** for each step into the future. In an infinite activity, this ends up valuing future value at $\frac{1}{1-\gamma}$. Gamma close to 0 creates a very close horizon where only immediate rewards are maximized, while Gamma close to 1 creates a long horizon, where maximizing long term reward is emphasized. As well be seen, MDPs are very sensitive to this and small changes can create very different solutions.

Forest MDP

Forest MDP is a simple MDP that models the value of a forest with respect to two actions that can be performed each year: ***wait*** or ***cut***. There is a stochastic element of forest fires that occur with probability ***p***. Each year is a state, with a max of ***S*** years / states. Cutting deterministically transitions to the initial state, ***state=0*** and provides a ***cutting reward***. Waiting transitions to the next year, or if in the max state, remains there. Forest fires provide a stochastic element, with a fire transitioning back to the initial state. A ***waiting reward*** only occurs in the max state, ***state=S***. This is a continual MDP, with no terminal or absorbing states.

The base setup for ***Forest*** is seven years (***S=7***), a 10% chance of forest fire (***p=0.1***), a cutting reward of 2 (***cutting reward=2***) and a waiting reward of 4 (***waiting reward = 4***).

In the context of MDP's, each state has an action that maximizes expected value and the set of maximizing actions for all states is called a ***policy***. This policy can be examined to determine what rational actors are likely going to do.

There are several very interesting aspects to this problem. ***Forest MDP*** models a key choice being made today around ecology and the environment. What are the rewards needed to keep forests around? How do maximizing actions (***policy***) change as the chance of forest fires increase? How does the length of the considered horizon (***Gamma discount***) influence expected outcomes.

This also highlights a key challenge in MDP's and Reinforcement Learning, which is understanding rewards and setting them appropriately. ***Forest MDP*** allows modeling of different situations to inform the construction of public policy for governments, NGO's and corporations. The ***cutting reward*** might be clearly modeled using expected timber value, but other types of reward may want to be considered in the ***waiting reward***, such as carbon reduction and maintenance of ecological diversity. This MDP can be used to model economic rewards and penalties to compel desired outcomes, such as a carbon tax that would reduce the value of the ***cutting reward***.

Forest MDP Value Iteration and Policy Iteration

Lake MDP Value Iteration and Policy Iteration

Forest MDP Q Learning

Lake MDP Q Learning

Summary

References

1 Markov Decision Processes (MDP) Toolbox, https://miat.inrae.fr/MDPtoolbox/ Accessed: 11/1/2022.

2 Markov Decision Process (MDP) Toolbox for Python, https://github.com/sawcordwell/pymdptoolbox: Accessed: 11/1/2022.

3 hiive Fork: Markov Decision Process (MDP) Toolbox for Python, https://github.com/hiive/hiivemdptoolbox: Accessed: 11/1/2022.

4 Brockman, G. et al., 2016. Openai gym.