Homework #4
Q-Learning

Problem

Description

In this homework you will have the complete RL experience. You will work towards implementing and evaluating the Q-learning algorithm on a simple domain. Q-learning is a fundamental RL algorithm and has been successfully used to solve a variety of decision-making problems. For this homework, you will have to think carefully about algorithm implementation, specially exploration parameters.

The domain you will be tackling is called Taxi (Taxi-v2). It is a discrete MDP which has been used for RL research in the past. This will also be your first opportunity to become familiar with the OpenAI Gym environment (https://github.com/openai/gym). This is a cool and unique platform where users can test their RL algorithms over a selection of domains.

The Taxi problem was introduced in Dietterich(2000). It is a grid-based domain where the goal of the agent is to pick up a passenger at one location and drop them off in another. There are 4 fixed locations, each assigned a different letter. The agent has 6 actions; 4 for movement, 1 for pickup, and 1 for dropoff. The domain has a discrete state space and deterministic transitions.

Procedure

Implement a basic version of the Q-learning algorithm and use it to solve the taxi domain. The agent should explore the MDP, collect data to the learn the optimal policy and the optimal Q-value function. (Be mindful of how you handle terminal states, typically if St is a terminal state, V(St+1) = 0). Use γ = 0.90. Also, you will see how Epsilon-Greedy strategy can find the optimal policy despite of finding sub-optimal q-values. Because we are looking for optimal q-values, you will have to try different exploration strategies.

Evaluate you agent using the OpenAI gym environment. Your TAs will provide you more information about setting up the environment and its basic usage.

Examples

Below are the optimal Q values for 5 (state, action) pairs of the Taxi domain.

- Q(462, 4) = -11.374402515
- Q(398, 3) = 4.348907
- Q(253, 0) = -0.5856821173
- Q(377, 1) = 9.683
- Q(83, 5) = -12.8232660372

Resources

The concepts explored in this homework are covered by:

- Lectures
  - Convergence
  - Exploring Exploration
- Readings
  - Asmuth-Littman-Zinkov-2008.pdf
  - littman-1996.pdf (chapters 1-2)
- Documentation
  - https://gym.openai.com/docs

Submission Details

The due date is indicated on the Canvas page for this assignment.

Make sure you have set your timezone in Canvas to ensure the deadline is accurate.

You will be evaluated based on optimality of results. This will be assessed by your algorithm's optimal Q-values for 10 specific state-action pairs (remember to use $\gamma = 0.90$). You will submit your results to 10 problems selected for you on the rldm website. The values will be graded on a 0.01 precision threshold.

To complete assignment, submit your Q-values to:

https://rldm.herokuapp.com

Optionally, you might want to, *with the same implementation*, solve the environment under OpenAI's criteria. If you accomplish that, you definitely learned something about exploration vs exploitation, the difference between optimal policy and optimal q-values, and should be proud about that.

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. Journal of Artificial Intelligence Research, 13, 227–303.