Classic and Risk-Sensitive Reinforcement Learning

Tanner Fiez
University of Washington

Overview

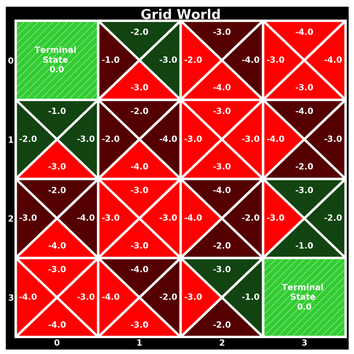
Reinforcement learning (RL) is a subfield of machine learning that focuses on learning from interactions. The core idea of RL is that an agent learns through the consequences of actions instead of being explicitly taught how to act. Actions are chosen based on past experience and new choices, which can be thought of as learning through trial and error. Inherently this forces a tradeoff between exploitation and exploration that must be solved.

Contributions

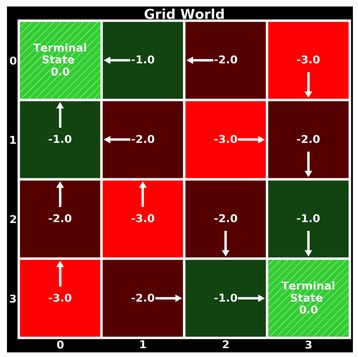
- Implementation of many classic RL algorithms in a flexible, object-oriented framework.
- Development of a grid world environment that allows for unique problem specifications, rapid testing and comparison of algorithms, and visualization of results.
- Compatibility of our library with OpenAI Gym [1]. OpenAI Gym is a recently developed python toolkit containing a wide variety of RL environments for evaluation purposes. Despite this advancement, there is still no universally used python package containing RL algorithms, providing motivation for this portion of the work as we seek to begin to create our own scikit-learn like package to make RL more accessible and ubiquitous.
- Implementation and analysis of risk-sensitive RL developed in [2] coupled with behavioral models of human decision making.
- Formulation of the decision process of taxi drivers in terms of where to go to search for new rides as an MDP. Application of RL methods to find the optimal policy based off information extracted from data in New York City [3] along with comparison to the empirical policy.

RL in Grid World

- In our grid world environment each grid square is considered as a state and actions can be configured to be the cardinal directions {N, E, S, W} or the compass directions {N, E, S, W, NE, SE, SW, NW}.
- In the example problem below transitions are deterministic with the exception that actions taking the agent off the grid keep the agent in the same state with probability 1, and when a terminal state is reached an agent stays is stuck in the state. The agent receives a reward of -1 for each action in every state with the exception that in a terminal state the agent receives a reward of 0. The discount factor on rewards is set to 1. Thus the optimal policy is to take the Manhattan Distance path from a state to a terminal state. We see below this is what is learned as expected.



Optimal State-Action Values



Optimal State Values and Policy

Risk-Sensitive Reinforcement Learning

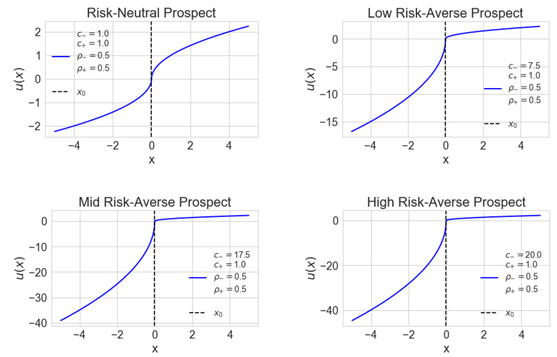
- Risk-Sensitive RL uses value functions that have been shown to capture true human decision making in algorithms to determine optimal policies for a decision maker with specific decision making characteristics. This is in contrast to classic RL methods which consider agents as expected utility maximizers.

Risk-Sensitive Q-Learning Update:

$$Q(s,a) = Q(s,a) + \alpha(u(r + \gamma \max_a q(s',a) - Q(s,a)))$$

Prospect Theory Value Function

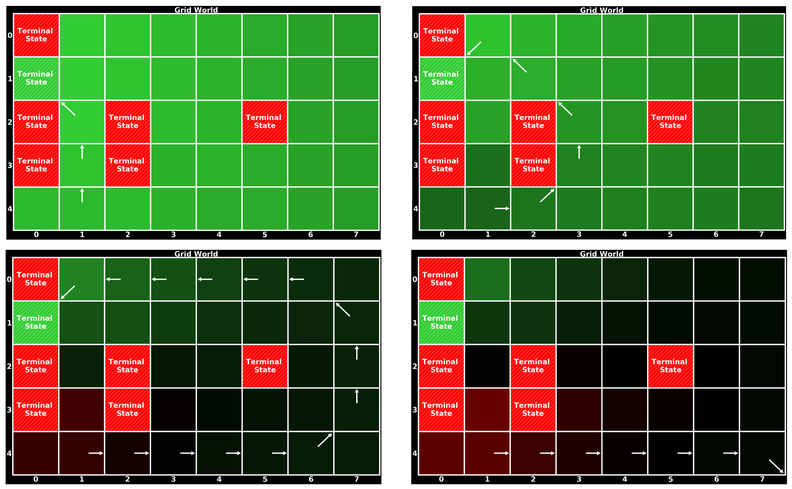
- Prospect theory provides a model of human decision making [4].



Prospect Theory Value Functions with Varying Degrees of Risk-Aversion

Risk-Sensitive RL in Grid World

- This example highlights how risk-preferences alter an optimal policy that is learned for an agent. The figures correspond to the value functions from above.
- In this example actions are the compass directions, the transition probabilities are such that the agent moves to the desired state with probability .93 and goes to a random state with probability .01 with the exception that when in a terminal state the agent stays in the terminal state with probability 1, an agent is given reward of .1 for each action with the exception that an agent gets reward -1 for entering and when in a bad terminal state and a reward of 1 for entering and when in a good terminal state. The discount factor is set to .95.



Optimal Policies Learned From a Specific Starting State with Different Risk Preferences

New York City Taxi Data

- The New York Taxi Dataset covers taxi operations from 2010-2013. We analyze a subset of the drivers and over only a few months in 2010. Each row in a file contains the information for a trip record. The key information that is contained is the hack license (driver ID), pickup date-time, dropoff date-time, trip time in seconds, trip distance in miles, GPS coordinates at the starting location, GPS coordinates at the ending location, total fare including tip, and the total cost of tolls.

Visualization of trip pickups. The majority of trips start in Manhattan or at one of the airports.

MDP Formulation

We model a taxi driver as acting according to a finite MDP where an episode corresponds to a single days work. They salient features of the formulation are:

- Each state is a tuple containing the node the driver is in–we discretized location into a grid using district boundaries for New York City–an indicator of whether the taxi currently is full (just picked up passengers) or empty (just dropped off passengers), and the cumulative reward interval the driver is in.
- Actions are moving between nodes in the location grid we created.
- The reward functions use values derived from the data, such as earning rate, the expected time searching for a passenger, and the fare between grid nodes. The transition probabilities use empirical transition probabilities as well as expected earning rates.
- The policy of a driver are the decisions they make of where to pick up new riders.

Conclusion and Future Work

In this work we extensively studied finite MDPs. Significant effort and time was devoted to implementing these algorithms and developing them in such a way that they were easy to evaluate in a wide array of settings. We also explored a lesser known RL paradigm of risk-sensitive RL that more effectively captures human decisions. By formulating the decision process of a taxi driver we were able to determine for a driver what the optimal policy would be to maximize earning rates and find the drivers are not acting optimally. In future work we seek to apply risk-sensitive RL to this problem to determine the risk preferences of a driver that explain their decisions.

References

[1] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. CoRR, abs/1606.01540, 2016.
[2] Yun Shen, Michael J. Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. CoRR, abs/1311.2097, 2013.
[3] B Donovan and DB Work. New york city taxi trip data (2010–2013), 2014.
[4] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and uncertainty, 5(4):297–323, 1992.
[5] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction, volume 1. MIT press Cambridge, 1998.