Max Ryoo (hr2ee)
CS 4710 (A.I.)
October 21, 2019
Professor Lu Feng

Paper Review

Apprenticeship Learning via Inverse Reinforcement Learning dives into the concept of using inverse reinforcement learning to extract an unknown reward function given an expert's performance.

One case that the authors gave to make an analogy of using inverse reinforcement learning was through the learning of driving. They said "When teaching a young adult to drive, rather than telling them what the reward function is, it is much easier and more natural to demonstrate driving to them, and have them learn from the demonstration.". They will then thus use this logic to extract a reward function from the actions of experts.

Before exploring the question of using inverse reinforcement learning dives there were a few cases and assumptions made by the researchers. The assumption that was made that in a Markov Decision Process (MDP) there is some vector of features over the states, which holds the true reward function. Also, the true reward function should be existing in the real number space. Along with this assumption the assumption of demonstrations from experts holding observable. This means that the experts will have their own reward functions that can be measured for the inverse reinforcing learning.

The paper goes into depth for two main algorithms that were distinguished by one algorithm requiring access to a QP (or SVM) solver, while in the other algorithm no QP solver was needed. The paper calls the QP-based algorithm the max-margin method and the simpler algorithm (no QP solver) the projection method. Both algorithms predicted on the assumption that the algorithm terminates with t <= epsilon, and the paper states that there is no case that the algorithm will take a large number of iterations to terminat nor will it ever not terminate.

Experiments were done in order to determine the convergence of these two algorithms, which showed that the two algorithms had similar rates of convergence with the projection version doing slightly better than the max-margin method. The second part of the experiment was to find a reward function expressed as a linear combination of known features based on demonstrations by an expert. The algorithms were formulated assuming the reward function is expressible as a linear function of known features. The limitation to this experiment is that if there are non-linear functions of the features that make up the reward functions it will not be caught in the algorithm. This experiment has a basis in assuming that the expert has a linear combination of features for their reward function.

This may give a fairly strong prediction and reward function that imitates the expert, but in some cases personally I think there may be limitations to extracting the correct reward function. In the case of driving, there may not be an exact expert. People have different ways of driving and different styles of driving. Some people may want to drive in an empty line and some people may want to drive the following the speed limit. These rules would lead to different reward function and never a true reward function converging from the reverse reinforcement learning.