Imitation Learning with a Value-Based Prior

Umar Syed, Robert E. Schapire
Princeton University

Abstract: In imitation learning, an intelligent agent in a stochastic environment learns a policy by observing a teacher executing the correct policy. By contrast, in a Markov Decision Process, the agent's "learning" consists of computing the policy that maximizes expected cumulative reward. This approach does not require any example trajectories from a teacher, but it does require specifying an appropriate reward function, which can sometimes be just as challenging.

Recently, Abbeel and Ng (2004) and Ratliff et al (2006) have shown that imitation learning can be sped up just by assuming that the teacher's policy is optimal with respect to some unknown reward function. In this work, we assume that we have access to a reward function that is *similar* to the one being optimized by the teacher. The reward function acts as a prior over the space of policies, assigning greater weight to those policies that have a higher value with respect to the given reward function. We give an efficient algorithm that locally maximizes a posterior that is based on this 'value-based' prior. A particularly useful feature of our algorithm is that it can be used in a setting where the example trajectories are not directly visible, but have been corrupted by noise. We show empirically that the rate at which our algorithm learns the teacher's policy increases with the value of that policy with respect to the given reward function.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.

Nathan D. Ratliff, J. Andrew Bagnell, and Martin A. Zinkevich. Maximum margin planning. In *Proceedings of the Twenty-Third International Conference on Machine Learning*, 2006.