Maximum Entropy Inverse Reinforcement Learning

Brian D. Ziebart
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
bziebart@cs.cmu.edu

J. Andrew Bagnell
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
dbagnell@ri.cmu.edu

Anind K. Dey
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213
anind@cs.cmu.edu

1 Extended Abstract

In many domains, demonstrating good behavior is easier than tuning parameters of an agent so that it behaves in a desirable way. A powerful recent idea to approach problems of imitation learning is to structure the space of learned policies to be solutions to search, planning, or, more generally, Markov Decision Problems. The imitation learning problem then is reduced to recovering a reward function that induces the demonstrated behavior.

Ratliff et al. [2] cast this problem as one of structured maximum margin prediction (MMP). These authors consider a class of loss functions that directly measure disagreement between an expert and a learned policy, and then efficiently learn a reward function using a convex relaxation of the loss function using the structured margin method using only oracle access to an MDP solver. However, this method suffers from some significant drawbacks when a single policy is not significantly better than all other policies, which can occur frequently in the presence of noise.

Abbeel and Ng [1] provide an alternate approach based on Inverse Reinforcement Learning (IRL). They propose a strategy of matching feature expectations between an expert's policy and a learner's behavior. Unfortunately, both the IRL concept and the matching feature counts are ambiguous. Each policy can be optimal for many reward functions (e.g., all zeros) and many policies or distributions over state/action pairs can lead to the the same feature counts. No method is proposed to resolve the ambiguity.

In this work, we treat uncertainty about expert behavior in a thoroughly probabilistic way. As in [1] and [2], we require our policies to match feature expectations. However, we attempt to estimate the probability of an expert taking trajectories as $p(\xi)$ using the principle of maximum entropy, which suggests that the natural distribution on trajectories is the unique distribution that maximizes the entropy of the distribution of trajectories subject to meeting the expectation constraints. In the absence of additional knowledge, the problem then is clear: $\max H(p(\xi))$ subject to $E_p[f_i(\xi)] = c_i$ where c_i are feature counts experienced by the expert we wish to imitate.

Solving this optimization problem yields a distribution on trajectories of the form $p(\xi) \propto e^{-w^T f}$, for feature counts f. The most likely trajectory then is, naturally, the one which minimizes $w^T f$ (where f is the feature count over the trajectory). We show that MaxEntIRL is more robust to noise than MMP and removes the ambiguity about reward functions that occurs with methods based on IRL, while providing the same key guarantee. MaxEntIRL produces policies that achieve (nearly) the same reward as the expert demonstrating the policy on the expert's unknown reward function – even when that exact reward function is unrecoverable (without ambiguity) from the available data.

For problems where actions have deterministic outcomes, the gradient of this convex optimization is then the difference between our learned policy's feature expectations and those of the demonstrated policy (Equation 1), for which we provide efficient algorithms for fixed time thresholds.

$$\frac{\delta}{\delta w_k} P(\tilde{\xi}|w) = c_i - E_p[f_i(\xi)|w] \qquad (1)$$

We apply this method to learn context-sensitive driving route preferences. The features of each road segment (e.g., road type, number of lanes) are mapped to a negative reward by parameters we optimize using over 100,000 miles of GPS trace data collected from Yellow Cab Pittsburgh taxi drivers. The resulting model is employed for recommending routes that use some of the traffic-avoding tricks that cab drivers employ.

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proc. ICML*, 2004.

[2] N. Ratliff, J. A. Bagnell, and M. Zinkevich. Maximum margin planning. In *Proc. ICML*, 2006.