Regularized MCE IRL

The single-task IRL problem is to recover a reward function $R$ given demonstrations $t$ from an MDP $M_t = (S, A, T, \gamma, R)$ and access to the world model $(S, A, T, \gamma, \mu)$. Maximum causal entropy (MCE) IRL assumes the reward function is linear in features over state-action pairs:

$$R(s, a) = \langle \theta, F(s, a) \rangle$$

For convenience, we write $F(t)$ to mean the (discounted) sum of features over state-action pairs in the trajectory. Maximum causal entropy IRL is equivalent to maximum causal likelihood estimation of $\theta$ given $t$, i.e. finding $\theta$ that maximizes the log likelihood:

$$\ell(\theta; t) = \sum_{s,t} \log P(a_t | s_{t-1}, a_{t-1}) + \frac{1}{2} \| \theta - \hat{\theta} \|^2$$

We evaluate this multi-task IRL algorithm in a few-shot reward learning problem on the above environment.

In the fixed test case (left), both the single-task baseline and our meta-AIRL algorithm produce near-optimal solutions. We conjecture this is because the optimal policy is unimodal, making it simple to extrapolate from a single trajectory. In the variable test case (right), single-task IRL fails to find a good solution even after observing 100 trajectories. Reptile can only learn a good initialisation in the outer loop when progress is made in the AIRL inner loop, so unsurprisingly our meta-AIRL algorithm also fails. Note the variable test case has a bimodal expert policy.

Conclusion & Further Work

Sample efficient solutions to the multi-task IRL problem are critical for enabling real-world applications, where collecting human demonstrations is expensive and slow. The multi-task IRL problem has previously been studied exclusively from a Bayesian IRL perspective. In this paper we took the alternative approach of formulating the multi-task problem inside the maximum causal entropy IRL framework.

Our first contribution uses the original MCE IRL algorithm, by adding a regularisation term to the loss. Experiments find our regularized MCE IRL algorithm can perform one-shot imitation learning in an environment that otherwise requires hundreds of demonstrations to learn.

In preliminary work, we combined the Reptile meta-learner with adversarial IRL, a sample-based MCE IRL algorithm. Testing revealed that adversarial IRL can only learn from unimodal expert policies, seriously limiting the applicability of meta-AIRL. We conjecture this limitation in adversarial IRL is related to the mode collapse in generative adversarial networks (GAN). A fruitful research direction might be to apply recent innovations in GAN training, such as unrolling the optimisation of the discriminator or variational learning, to stabilise adversarial IRL training.

Further Information

Full paper: bit.ly/MultiTaskIRL
Source code: bit.ly/MultiTaskIRLCode