Review of Inverse Reinforcement Learning

Guannan Hu

April 7, 2018

1 Inverse Reinforcement Learning

The inverse reinforcement learning (IRL) problem can be characterized informally as follow:

> **Given** 1) measurements of an agent's behaviour over time, in a variety of circumstances, 2) if needed, measurements of the sensory inputs to that agent; 3) if available, a model of the environment.
>
> **Determine** the reward function being optimized.

In examining animal and human behaviour we must consider the reward function as an unknown to be ascertained through empirical investigation.

The inverse reinforcement learning problem is to find a reward function can explain observed behaviour. We begin with the simple case where the state space is finite, the model is known, and the complete policy is observed. Most precisely, we are given a finite state space S, a set of k actions $A = a_1, ..., a_k$, transition probabilities $\{P_{sa}\}$, a discount factor γ, and a policy π; we then wish to find the set of possible reward functions R such that π is an optimal policy in the MDP $(S, A, \{P_{sa}\}, \gamma, R)$.

2 Inverse Reinforcement Learning Review

Acquiring a reward function is important (and challengeing)

Goal of Inverse RL : infer reward function underlying expert demonstrations

Evaluating the partition function:

- initial approaches solve the MDP in the inner loop and/or assume known dynamics
- with unknown dynamics, estimate Z using samples

Connection to generative adversarial networks:

- sampling-based MaxEnt IRL is a GAN with a special form of discriminator and uses RL to optimize the generator.

3 Markov Decision Processes

A (finite) MDP is a tuple $(S, A, P_{sa}, \gamma, R)$, where

- S is a finite set of N **states**.
- $A = \{a_1, ..., a_k\}$ is a set of k **actions**.
- $P_{sa}(\cdot)$ are the state **transition probabilities** upon taking action a in state s.
- $\gamma \in [0, 1)$ is the **discount factor**.
- $R : S \to \mathbb{R}$ is the **reinforcement function** bounded in absolute value by R_{max}.

A **policy** is defined as any map $\pi : S \to A$, and the **value function** for a policy π, evaluated at any state s_1 is given by

$$V^\pi(s_1) = \mathbb{E}[R(s_1) + \gamma R(s_2) + \gamma^2 R(s_3) + \cdots |\pi] \tag{1}$$

where the expectation is over the distribution of the state sequence $(s_1, s_2, ...)$,we pass through when we execute the policy π starting from s_1. We also defined the **Q-function** according to

$$Q^\pi(s, a) = R(s) + \gamma \mathbb{E}_{s' \sim P_{s,a}(\cdot)}[V^\pi(s^{'})] \tag{2}$$

The **optimal value function** is $V^*(s) = sup_\pi V^\pi(s)$, and the **optimal Q-function** is $Q^*(s, a) = sup_\pi Q^\pi(s, a)$.

4 Basic Properties of MDPs

Theorem 1(Bellman Equations) *Let an MDP $M = (S, A, \{P_{sa}\}, \gamma, R)$ and a policy $\pi : S \to A$ be given. Then, for all $s \in S, a \in A, V^\pi$ and Q^π satisfy*

$$V^\pi(s) = R(s) + \gamma \sum_{s^{'}} P_{s\underline{\pi(s)}}(s^{'})V^\pi(s^{'}) \tag{3}$$

$$Q^\pi(s, a) = R(s) + \gamma \sum_{s^{'}} P_{s\underline{a}}(s^{'})V^\pi(s^{'}) \tag{4}$$

Theorem 2 (Bellman Optimality) *Let an MDP $M = (S, A, \{P_{sa}\}, \gamma, R)$ and a policy $\pi : S \to A$ be given. Then π is an optimal policy for M if and only if, for all $s \in S$,*

$$\pi(s) \in \arg\max_{a \in A} Q^\pi(s, a) \tag{5}$$

5 Furthe Reading on Inverse RL

- **MaxEnt-based IRL**: Ziebart et al. AAAI'08, Wulfmeier et al. arXiv'16, Finn et al. ICML'16
- **Adversarial IRL**: Ho & Ermon NIPS'16, Finn*, Christiano* et al. arXiv'16, Baram et al. ICML'17
- **Handling multimodality**: Li et al. arXiv'17, Hausman et al.arXiv'17, Wang, Merel et al. arXiv'17
- **Handling domain shift**: Stadie et al. ICLR' 17