R9/Ishank Juneja/16D070012

February 18, 2020

Under the Markov Decision Process (MDP) model of agent-environment interaction, there is a reward function $R(s, a, s')$ that tells us the one-step return associated with each state transition $s \to s'$ under a certain action a. In problems where we wish to understand animal and human learning, the optimal (desirable) behaviour exhibited by the agent is known (in the form of an optimal policy π^*) but the underlying reward function being optimized for is unknown. Solving such problems comes under the ambit of *inverse reinforcement learning* (IRL). The seminal paper by Ng and Russel introduces the IRL problem and provides algorithms to solve IRL under three situations, which are -

- A tabular representation is available and the policy π^* is known
- The MDPs state space is continuous and the policy π^* is known
- The policy π^* is available only through a finite set of observed trajectories

First the authors characterize possible solutions for reward functions under the case when a tabular representation is available. Given an MDP with a finite state space S, a set of k actions $A = \{a_1, \ldots, a_k\}$, transition probabilities $\{P_{sa}\}$, a discount factor γ and with the optimal policy π as known. We wish to characterize the possible reward functions R. For notational simplicity, the paper takes $\pi(s) \equiv a_1$. The criteria derived in the paper is

$$(P_{a_1} - P_a)(I - \gamma P_{a1})^{-1} R \succeq 0$$

Here P_{a1} and P_a are state transition probabilities under actions a_1 and $a \neq a_1$ respectively and the reward function R has been written as a vector for notational compactness. The authors show that this characterization of R captures all the solutions to the IRL problem, however many of the solutions characterized by the above criteria are uninformative. In particular, R being any constant vector is a solution. To filter out these *degenerate* answers, the authors add an additional heuristic that favours solutions which maximize the single step deviations from the optimal policy π. The motivation being that π become the clear choice for the optimal policy under reward scheme R. At the end of the section, the paper provides a Linear Programming (LP) formulation to determine R.

Next, the authors move onto to the IRL problem under infinite state spaces. In particular they consider the case of $S = \mathbb{R}^n$. In this situation reward functions are mappings $\mathbb{R}^n \to \mathbb{R}$ and the most general solution to the IRL problem would require variational calculus. The paper points out that this approach is algorithmically problematic. In view of this, the authors assumes R to be a linear combination of some basis functions ϕ_i's derived from the family of basis functions $\{\phi_i\}_1^d$. That is, they assume $R(s) = \alpha_1\phi_1(s) + \ldots + \alpha_d\phi_d(s)$. Under this scheme, the authors provide a criteria on R and an LP formulation to obtain desirable R's analogous to the tabular case.

Lastly, the authors provide a solution to the IRL problem when only sampled trajectories are available. This situation is quite a practical one since it is often the case that we do not have access to an explicit environment model in the form of an MDP, and that policies are not available as state-action mappings but merely as trajectories. Just like the previous case of continuous state-space representation, even here $R(s)$ is assumed to be a linear combination of basis functions $\{\phi_i\}_1^d$. The algorithm assumes the ability to simulate trajectories and compute their returns, and proceeds in two phases. First, the return of each basis function ϕ_i is estimated empirically using roll-outs. In the next stage of the algorithm, appropriate weights α_i for the basis functions are computed by solving an LP. To do this, an objective analogous to the last two cases is formulated and solved. The objective uses a set of trajectories $\{\pi_1 \ldots \pi_k\}$, and on every iteration expands the set of reference policies to include the policy optimal under the most recently computed reward scheme R. The process of solving Linear Programs to determine new reward schemes continues until we are satisfied by the optimal policy being close enough to the desired π.

Some questions I have are -

- It is not clear to me how the optimization problems presented in the paper is an LP formulation sine the objective involves cascaded max-min operators.
- I could not understand the transformation of the necessary conditions on R from the tabular to the infinite state space case.