Inverse reinforcement learning and reinforcement learning from human feedback

Justin

April 18, 2023

1 Introduction

Is reinforcement learning from human feedback (RLHF) related to inverse reinforcement learning (IRL)? Let's write out the math of both and see if they line up.

2 Inverse reinforcement learning (IRL)

IRL is a specific imitation learning method. Imitation learning methods aim to learn a good policy from expert demonstrations. The goal of IRL is to first learn a reward function from human demonstrations, then use that reward function to determine a policy.

We will focus on learning a reward function, since distilling a reward function into a policy is straightforward (albeit expensive) using policy gradient and variance reduction techniques.

2.1 Maxent IRL

Given a dataset of expert demonstrations consisting of contexts x_i and trajectories y_i, the goal is to learn a reward function $r(x, y)$. The maximum entropy IRL hypothesis is that the expert takes a trajectory with probability proportional to the exponentiated reward:

$$\log P(x \mid y) = R(x, y) - \log Z(R, y), \tag{1}$$

where the partition function $Z(R, y) = \sum_x e^{(R(x,y))}$ is intractable. This is a conditional model over trajectories given a context, globally normalized over all trajectories.

To train this model, we must find the reward function R that best explains the expert demonstrations. We would like to use maximum likelihood estimation,

$$\underset{R}{\operatorname{argmax}} \sum_i \log P(x_i \mid y_i),$$

by performing gradient descent. The main challenge is computing the derivative of the partition function $Z(R, y)$. We can approximate the derivative with importance sampling:

$$\nabla_R \log P(x \mid y) \approx \nabla_R R(x, y) - \sum_{x'} \nabla_R R(x', y),$$

where we sample non-expert trajectories from the model, $x' \sim P(x \mid y)$. Note that sampling from a globally normalized distribution is also difficult, so this would require an easy-to-sample from approximation (e.g a policy).

3 Reinforcement learning from human feedback (RLHF)