RL seminar #6: Inverse RL

Eugene Golikov

MIPT, Deep learning lab & iPavlov.ai

9 Dec 2017

Outline

Inverse Reinforcement Learning

- Problem statement
- Probabilistic inference
- MaxEnt IRL

Questions

Table of Contents

IRL: problem statement

What we have:

- State space $\mathcal{S}$ and action space $\mathcal{A}$
- Transition dynamics $p(s'|s,a)$ (sometimes)
- Samples from experts policy $\{\tau_i\}$, $\forall i\ \tau_i \sim \pi^*(\tau)$
- Parametric class of reward functions $r_\psi(s,a)$

IRL: problem statement

What we have:

- State space $\mathcal{S}$ and action space $\mathcal{A}$
- Transition dynamics $p(s'|s, a)$ (sometimes)
- Samples from experts policy $\{\tau_i\}$, $\forall i \ \tau_i \sim \pi^*(\tau)$
- Parametric class of reward functions $r_\psi(s, a)$

Goal:

Learn ψ such that experts policy "optimizes" $r_\psi(\tau)$

IRL: problem statement

What we have:

- State space $\mathcal{S}$ and action space $\mathcal{A}$
- Transition dynamics $p(s'|s, a)$ (sometimes)
- Samples from experts policy $\{\tau_i\}$, $\forall i \ \tau_i \sim \pi^*(\tau)$
- Parametric class of reward functions $r_\psi(s, a)$

Goal:
Learn ψ such that experts policy "optimizes" $r_\psi(\tau)$
(and additionally learn experts policy π^*)

Problem:

- Expert is *not* an optimal controller

Problem:

- Expert is *not* an optimal controller
- Experts policy is *not* (strictly) optimal to any r_ψ (since it is stochastic)

Problem:

- Expert is *not* an optimal controller
- Experts policy is *not* (strictly) optimal to any r_ψ (since it is stochastic)

Also, that's one of the reasons why imitation learning is not generally a good idea

Solution:

Let experts trajectories $\{\tau_i\}$ be the samples from some distribution of optimal trajectories — *soft-optimal policy*

Solution:

Let experts trajectories $\{\tau_i\}$ be the samples from some distribution of optimal trajectories — *soft-optimal policy*

In RL formalism optimal policy is always greedy:

$$\pi^*(a_t|s_t) = \arg\max_a Q(a|s_t)$$

Solution:

Let experts trajectories $\{\tau_i\}$ be the samples from some distribution of optimal trajectories — *soft-optimal policy*

In RL formalism optimal policy is always greedy:

$$\pi^*(a_t|s_t) = \arg\max_a Q(a|s_t)$$

RL formalism doesn't fit $\Rightarrow$ we need another formalism

A probabilistic graphical model of decision making

- Introduce dummy boolean variables $\mathcal{O}_{1:T}$ with distribution:

$$p(\mathcal{O}_t | s_t, a_t, \psi) \propto \exp(r_\psi(s_t, a_t))$$

A probabilistic graphical model of decision making

- Introduce dummy boolean variables $\mathcal{O}_{1:T}$ with distribution:

$$p(\mathcal{O}_t | s_t, a_t, \psi) \propto \exp(r_\psi(s_t, a_t))$$

- IRL problem is stated as follows:

$$\frac{1}{N} \sum_{i=1}^{N} \log p(\tau_i | \mathcal{O}_{1:T}, \psi) \to \max_\psi$$

A probabilistic graphical model of decision making

- Introduce dummy boolean variables $\mathcal{O}_{1:T}$ with distribution:

$$p(\mathcal{O}_t | s_t, a_t, \psi) \propto \exp(r_\psi(s_t, a_t))$$

- IRL problem is stated as follows:

$$\frac{1}{N} \sum_{i=1}^{N} \log p(\tau_i | \mathcal{O}_{1:T}, \psi) \to \max_\psi$$

- "Optimal" policy π^{r_ψ} is not stated explicitly, but can be *inferred*, if the transition dynamics is known:

$$\pi^{r_\psi}(a_t | s_t) = p(a_t | s_t, \mathcal{O}_{1:T}, \psi)$$

Probabilistic inference

Useful formulae:

- Backward messages:

$$\beta^{r_\psi}(s_t, a_t) = p(\mathcal{O}_{t:T}|s_t, a_t, \psi)$$
$$= p(\mathcal{O}_t|s_t, a_t, \psi)\mathbb{E}_{s_{t+1}\sim p(\cdot|s_t, a_t)}\beta^{r_\psi}(s_{t+1})$$

$$\beta^{r_\psi}(s_{t+1}) = p(\mathcal{O}_{t:T}|s_{t+1}, \psi) = \mathbb{E}_{a_{t+1}\sim p(\cdot)}\beta^{r_\psi}(s_{t+1}, a_{t+1})$$

Probabilistic inference

Useful formulae:

- Backward messages:

$$\begin{aligned} \beta^{r_\psi}(s_t, a_t) &= p(\mathcal{O}_{t:T} | s_t, a_t, \psi) \\ &= p(\mathcal{O}_t | s_t, a_t, \psi) \mathbb{E}_{s_{t+1} \sim p(\cdot | s_t, a_t)} \beta^{r_\psi}(s_{t+1}) \end{aligned}$$

$$\beta^{r_\psi}(s_{t+1}) = p(\mathcal{O}_{t:T} | s_{t+1}, \psi) = \mathbb{E}_{a_{t+1} \sim p(\cdot)} \beta^{r_\psi}(s_{t+1}, a_{t+1})$$

- Forward messages:

$$\begin{aligned} \alpha^{r_\psi}(s_t) &= p(s_t | \mathcal{O}_{1:t-1}, \psi) \\ &= \mathbb{E}_{a_{t-1} \sim p(\cdot)} \mathbb{E}_{s_t \sim p(\cdot | s_{t-1}, a_{t-1})} p(\mathcal{O}_{t-1} | s_{t-1}, a_{t-1}, \psi) \frac{\alpha^{r_\psi}(s_{t-1})}{\beta^{r_\psi}(s_{t-1})} \end{aligned}$$

Probabilistic inference

Useful formulae:

- Backward messages:

$$\begin{aligned}\beta^{r_\psi}(s_t, a_t) &= p(\mathcal{O}_{t:T}|s_t, a_t, \psi) \\ &= p(\mathcal{O}_t|s_t, a_t, \psi)\mathbb{E}_{s_{t+1}\sim p(\cdot|s_t,a_t)}\beta^{r_\psi}(s_{t+1})\end{aligned}$$

$$\beta^{r_\psi}(s_{t+1}) = p(\mathcal{O}_{t:T}|s_{t+1}, \psi) = \mathbb{E}_{a_{t+1}\sim p(\cdot)}\beta^{r_\psi}(s_{t+1}, a_{t+1})$$

- Forward messages:

$$\begin{aligned}\alpha^{r_\psi}(s_t) &= p(s_t|\mathcal{O}_{1:t-1}, \psi) \\ &= \mathbb{E}_{a_{t-1}\sim p(\cdot)}\mathbb{E}_{s_t\sim p(\cdot|s_{t-1},a_{t-1})}p(\mathcal{O}_{t-1}|s_{t-1}, a_{t-1}, \psi)\frac{\alpha^{r_\psi}(s_{t-1})}{\beta^{r_\psi}(s_{t-1})}\end{aligned}$$

- Policy:

$$\pi^{r_\psi}(a_t|s_t) = p(a_t|\mathcal{O}_{1:T}, s_t, \psi) = \frac{\beta^{r_\psi}(s_t, a_t)}{\beta^{r_\psi}(s_t)}p(a_t)$$

Connection with Q and V

Let

$$Q^{r_\psi}(a_t, s_t) := \log \beta^{r_\psi}(s_t, a_t), \qquad V^{r_\psi}(s_t) := \log \beta^{r_\psi}(s_t)$$

Connection with Q and V

Let

$$Q^{r_\psi}(a_t, s_t) := \log \beta^{r_\psi}(s_t, a_t), \qquad V^{r_\psi}(s_t) := \log \beta^{r_\psi}(s_t)$$

Then

$$V^{r_\psi}(s_t) = \log(\mathbb{E}_{a_t \sim p(\cdot)} \exp(Q^{r_\psi}(a_t, s_t))) \approx \max_a Q^{r_\psi}(a, s_t)$$

Connection with Q and V

Let

$$Q^{r_\psi}(a_t, s_t) := \log \beta^{r_\psi}(s_t, a_t), \qquad V^{r_\psi}(s_t) := \log \beta^{r_\psi}(s_t)$$

Then

$$V^{r_\psi}(s_t) = \log(\mathbb{E}_{a_t \sim p(\cdot)} \exp(Q^{r_\psi}(a_t, s_t))) \approx \max_a Q^{r_\psi}(a, s_t)$$

Policy:

$$\pi^{r_\psi}(a_t|s_t) = \exp(Q^{r_\psi}(a_t, s_t) - V^{r_\psi}(s_t)) = \exp(A^{r_\psi}(a_t, s_t))$$

MaxEnt IRL

Objective

Maximize:

$$\begin{aligned}\mathcal{L}(\psi) &= \mathbb{E}_{\tau\sim\pi^*} \log p(\tau|\mathcal{O}_{1:T}, \psi) \\ &= \mathbb{E}_{\tau\sim\pi^*}(\log p(\mathcal{O}_{1:T}|\tau, \psi) + \log p(\tau)) - \log p(\mathcal{O}_{1:T}|\psi)\end{aligned}$$

MaxEnt IRL

Objective

Maximize:

$$\begin{aligned}\mathcal{L}(\psi) &= \mathbb{E}_{\tau\sim\pi^*} \log p(\tau|\mathcal{O}_{1:T}, \psi) \\ &= \mathbb{E}_{\tau\sim\pi^*} (\log p(\mathcal{O}_{1:T}|\tau, \psi) + \log p(\tau)) - \log p(\mathcal{O}_{1:T}|\psi)\end{aligned}$$

Perform gradient ascent:

$$\begin{aligned}\nabla_\psi \mathcal{L}(\psi) &= \mathbb{E}_{\tau\sim\pi^*} \nabla_\psi \log p(\mathcal{O}_{1:T}|\tau, \psi) - \nabla_\psi \log p(\mathcal{O}_{1:T}|\psi) \\ &= \mathbb{E}_{\tau\sim\pi^*} \nabla_\psi r_\psi(\tau) - \mathbb{E}_{\tau\sim p(\tau|\mathcal{O}_{1:T},\psi)} \nabla_\psi r_\psi(\tau)\end{aligned}$$

$p(\tau|\mathcal{O}_{1:T}, \psi)$ — distribution on optimal (w.r.t. r_ψ) trajectories inferred from our probabilistic model, or *soft-optimal policy*

MaxEnt IRL

Gradient ascent:

$$\nabla_\psi \mathcal{L}(\psi) = \mathbb{E}_{\tau \sim \pi^*} \nabla_\psi r_\psi(\tau) - \mathbb{E}_{\tau \sim p(\tau | \mathcal{O}_{1:T}, \psi)} \nabla_\psi r_\psi(\tau)$$

The first expectation is easy; what about the second?

MaxEnt IRL

Gradient ascent:

$$\nabla_\psi \mathcal{L}(\psi) = \mathbb{E}_{\tau \sim \pi^*} \nabla_\psi r_\psi(\tau) - \mathbb{E}_{\tau \sim p(\tau | \mathcal{O}_{1:T}, \psi)} \nabla_\psi r_\psi(\tau)$$

The first expectation is easy; what about the second?

- If the transition dynamics is known, and both $\mathcal{A}$ and $\mathcal{S}$ are small, it can be computed explicitly (via prob. inference)

MaxEnt IRL

Gradient ascent:

$$\nabla_\psi \mathcal{L}(\psi) = \mathbb{E}_{\tau \sim \pi^*} \nabla_\psi r_\psi(\tau) - \mathbb{E}_{\tau \sim p(\tau | \mathcal{O}_{1:T}, \psi)} \nabla_\psi r_\psi(\tau)$$

The first expectation is easy; what about the second?

- If the transition dynamics is known, and both $\mathcal{A}$ and $\mathcal{S}$ are small, it can be computed explicitly (via prob. inference)
- Else, we can fit the soft-optimal policy using MaxEnt RL algorithm, and sample trajectories from it

MaxEnt RL

How to fit the soft-optimal policy?

MaxEnt RL

How to fit the soft-optimal policy?

Minimize $D_{KL}(q(\tau) \| p(\tau | \mathcal{O}_{1:T}, \psi))$ w.r.t. $q(\tau)$

MaxEnt RL

How to fit the soft-optimal policy?

Minimize $D_{KL}(q(\tau) \| p(\tau | \mathcal{O}_{1:T}, \psi))$ w.r.t. $q(\tau)$

We can show that:

$$
\begin{aligned}
D_{KL}(q(\tau) \| p(\tau | \mathcal{O}_{1:T}, \psi)) &= D_{KL}(q(\tau) \| p(\tau)) \\
&\quad + \log p(\mathcal{O}_{1:T} | \psi) - \mathbb{E}_{\tau \sim q(\cdot)} \log p(\mathcal{O}_{1:T} | \tau, \psi)
\end{aligned}
$$

MaxEnt RL

How to fit the soft-optimal policy?

Minimize $D_{KL}(q(\tau)\|p(\tau|\mathcal{O}_{1:T},\psi))$ w.r.t. $q(\tau)$

We can show that:

$$\begin{aligned} D_{KL}(q(\tau)\|p(\tau|\mathcal{O}_{1:T},\psi)) &= D_{KL}(q(\tau)\|p(\tau)) \\ &+ \log p(\mathcal{O}_{1:T}|\psi) - \mathbb{E}_{\tau\sim q(\cdot)} \log p(\mathcal{O}_{1:T}|\tau,\psi) \end{aligned}$$

If we choose $p(\tau)$ as uniform, then minimizing r.h.s. is equivalent to:

$$\mathbb{E}_{\tau\sim\pi} \sum_{t=1}^{T} (r_\psi(a_t, s_t) + \mathcal{H}(\pi(\cdot|s_t))) \to \max_\pi$$

MaxEnt IRL: algorithm

1. Perform MaxEnt RL for r_ψ:

$$\pi^{r_\psi} = \arg\max_\pi \left(\mathbb{E}_{\tau\sim\pi} \sum_{t=1}^{T} (r_\psi(a_t, s_t) + \mathcal{H}(\pi(\cdot|s_t))) \right)$$

2. Compute gradient of experts trajectories likelihood:

$$\nabla_\psi \mathcal{L}(\psi) = \mathbb{E}_{\tau\sim\pi^*} \nabla_\psi r_\psi(\tau) - \mathbb{E}_{\tau\sim\pi^{r_\psi}} \nabla_\psi r_\psi(\tau)$$

3. $\psi := \psi + \alpha \nabla_\psi \mathcal{L}(\psi)$

MaxEnt IRL: algorithm

1. Perform MaxEnt RL for r_ψ:

$$\pi^{r_\psi} = \arg\max_\pi \left(\mathbb{E}_{\tau\sim\pi} \sum_{t=1}^{T} (r_\psi(a_t, s_t) + \mathcal{H}(\pi(\cdot|s_t))) \right)$$

2. Compute gradient of experts trajectories likelihood:

$$\nabla_\psi \mathcal{L}(\psi) = \mathbb{E}_{\tau\sim\pi^*} \nabla_\psi r_\psi(\tau) - \mathbb{E}_{\tau\sim\pi^{r_\psi}} \nabla_\psi r_\psi(\tau)$$

3. $\psi := \psi + \alpha \nabla_\psi \mathcal{L}(\psi)$

Modifications:

- Perform only one step at a time for MaxEnt RL
- Use importance sampling to keep expectation estimate unbiased

Table of Contents

Questions?