Apprenticeship Learning via Inverse Reinforcement Learning

Reminders

- γ-discounted state visitation distribution of a policy is defined as:

$$\rho_\pi(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{p_0,\pi}[s_t = s],$$

  or

$$\rho_\pi(s) = \sum_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p(s'_t) p(s'_t \to s_{t+1}|\pi)$$

Overview

- Seminal paper on inverse reinforcement learning
- Goal is to try to learn a reward from the expert, and use that to solve the MDP
- The initial MDP denoted as MDP\R
- Two methods presented for learning the expert behavior
- Presented evaluations on two simple scenarios: a gridworld 8x8 and a 2D car simulator

Key ingredients

- Using binary coding to create features from the states: $\phi : \mathcal{S} \to [0,1]^k$
- Assume "true" reward function: $R^*(s) = \omega^* \phi(s)$, with $||\omega||_1 \leq 1$ (R is linear in features)
- This yields:

$$\mathbb{E}_{s_0 \sim D}[V^\pi(s_0)] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \big| \pi\right] = \omega \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) \big| \pi\right]$$

- Now define the **feature expectation** as: $\mu(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) \big| \pi\right]$
  - This looks similar to the discounted state visitation distribution, which is what the algorithm tries to learn in a sense
  - "Pick the policy that best describes expert's discounted state visitation distribution"
- The goal is then to learn the experts feature expectation and to obtain ω as the max-margin

Comments

- Overall interesting paper full of fresh ideas
- It took me three passes to fully understand it (hopefully only three)
- Interested to explore what the benefits of this approach would be compared to simple imitation learning