论文来源：
https://papers.nips.cc/paper/2016/file/cc7e2b878868cbae992d1fb743995d8f-Paper.pdf
NIPS 2016

Learning And Mining from DatA
http://lamda.nju.edu.cn

Generative Adversarial Imitation Learning

Presented by Zhaojin Wen

2020.11.10

Background

- 行为克隆（BC）：

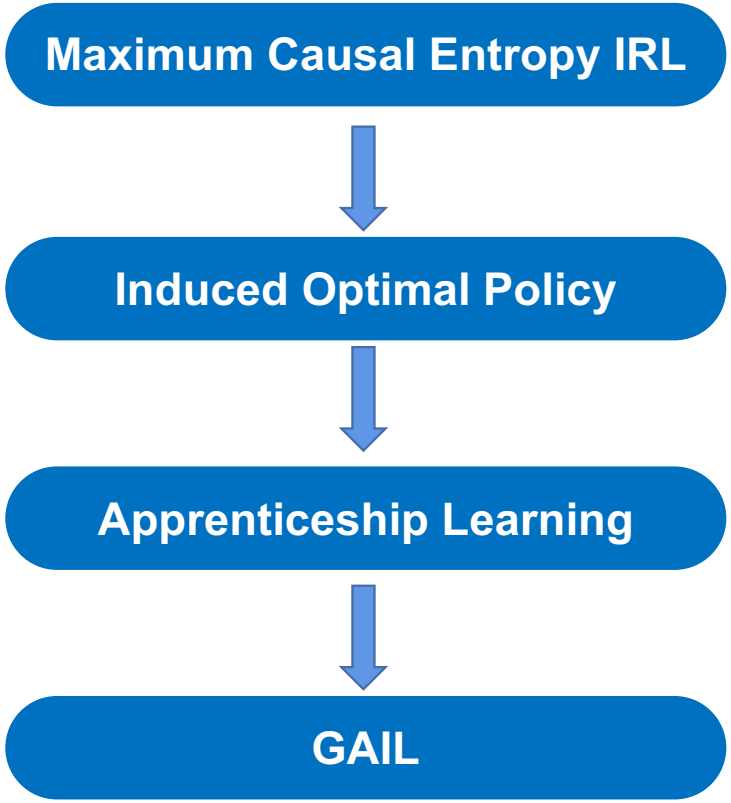
方法简单且需要大量数据，且由于只学习单步策略会产生极联误差

- 逆强化学习（IRL）：

不会产生极联误差，但学习cost function需要嵌套RL过程，复杂度高

因此希望设计一个算法，能够直接从示例数据中学习策略，避免复杂计算

Structure


Maximum Causal Entropy IRL
Induced Optimal Policy
Apprenticeship Learning
GAIL

Notations

- $c(s,a)$: $cost\ function$ （即负reward）
- $E_\pi[c(s,a)] \triangleq E[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)]$
- π_E: $expert\ policy$
- occupancy measure（占用度）: $\rho_\pi(s,a) = \pi(a|s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \pi)$

占用度可看作一个非归一化的动作-状态对的分布

$$E_\pi[c(s,a)] = \sum_{s,a} \rho_\pi(s,a) c(s,a)$$

- convex conjugate（凸共轭）: $f^*(x) = sup_{y \epsilon R^{S \times \mathcal{A}}} x^T y - f(y)$

Background: Maximum Causal Entropy IRL

- Optimization function:

$$\max_{c \epsilon C}\left(\min_{\pi \epsilon \Pi} -H(\pi) + E_{\pi}[c(s,a)]\right) - E_{\pi_E}[c(s,a)] \qquad (1)$$

找到的cost function在专家策略上具有较低的cost,对于其他策略都具有较高的cost

- reinforcement learning procedure:

$$RL(c) = argmin_{\pi \epsilon \Pi} - H(\pi) + E_{\pi}[c(s,a)]$$

$$H(\pi) \triangleq E_{\pi}[-log\pi(a|s)], \gamma - discounted\ causal\ entropy$$

Induced Optimal Policy

- (1)中函数族C为所有$R^{S\times A} = \{c: \mathcal{S}\times\mathcal{A} \longrightarrow R\}$ ——容易过拟合
- 加入凸的正则项: ψ : $R^{S\times A} \longrightarrow R$ 这里凸性要求不严但必须定义在$R^{S\times A}$上

IRL原始过程：

$$IRL_{\psi}(\pi_E) = argmax_{c\epsilon R^{\mathcal{S}\times\mathcal{A}}} -\psi(c) + \left(\min_{\pi\epsilon\Pi} -H(\pi) + E_{\pi}[c(s,a)]\right) - E_{\pi_E}[c(s,a)]$$

令$\tilde{c} \in IRL_{\psi}(\pi_E)$， 关注$RL(\tilde{c})$，即利用RL，基于IRL得出的cost function学习策略

Induced Optimal Policy

Proposition 3.1:

$$\mathrm{RL} \circ \mathrm{IRL}_{\psi}(\pi_E) = \ argmin_{\pi \epsilon \Pi} - \mathrm{H}(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E})$$

- occupancy measure : $\rho_\pi(s,a) = \ \pi(a|s) \sum_{t=0}^{\infty} P(s_t = s \mid \pi)$
- $E_\pi[c(s,a)] = \sum_{s,a} \rho_\pi(s,a) c(s,a)$
- convex conjugate : $f^*(x) = sup_{y \epsilon R^{S \times \mathcal{A}}} x^T y - f(y)$

Proposition 3.2:

$Suppose\ \rho_{\pi_E} > 0. If\ \psi\ is\ a\ constant\ function, \tilde{c} \in IRL_{\psi}(\pi_E), and\ \tilde{\pi} \in RL(\tilde{c}), then\ \rho_{\tilde{\pi}} = \rho_{\pi_E}.$

- recovered policy exactly matches the expert's occupancy measure
- one-to-one correspondence between policy and occupancy measure

Induced Optimal Policy

Conclusion:

- IRL是如下occupancy measure问题的对偶问题:

$$\min_{\rho \in \mathcal{D}} -\bar{H}(\rho) \quad subject\ to \quad \rho(s,a) = \rho_E(s,a) \ \ \forall s \in \mathcal{S}, a \in \mathcal{A}$$

- 恢复得到的induced optimal policy就是对偶问题和原问题的最优解

因此可以将IRL看成试图induce得到可以匹配专家策略的occupancy measure

Apprenticeship Learning

上述证明了当ψ是常数时，可以求得与专家策略的occupancy measure完全匹配的策略，但并不适用于实际问题

实际上，专家数据都是有限的样本集合，因此高维环境中学得的策略基本不会访问那些数据集中没有出现过的状态

用函数逼近的方式来学习策略

- Motivated by proposition 3.1，将IRL的对偶问题转化为如下形式:

$$\underset{\pi}{\text{minimize}}\ d_{\psi}(\rho_{\pi}, \rho_{E}) - H(\pi) \qquad (2)$$

其中， $d_{\psi}(\rho_{\pi}, \rho_{E}) \triangleq \psi^{*}(\rho_{\pi} - \rho_{\pi_{E}})$，可以通过修改ψ平滑惩罚不同occupancy measure之间的扰动

在特定的ψ下，上式可转化为学徒学习的算法求解，并拓展到高维环境

Apprenticeship Learning

- 学徒学习的优化目标：

$$\underset{\pi}{\text{minimize}} \max_{c \in \mathcal{C}} E_{\pi}[c(s,a)] - E_{\pi_E}[c(s,a)] \qquad (3)$$

常用的$\mathcal{C} \in R^{\mathcal{S}\times\mathcal{A}}$是基础函数的线性组合：

(3) is a special case of (2)

$$\mathcal{C}_{linear} = \left\{\sum_i \omega_i f_i : ||\omega||_2 \le 1\right\} \; and \; \mathcal{C}_{convex} = \left\{\sum_i \omega_i f_i : \sum_i \omega_i = 1, \omega_i \ge 0, \forall i\right\}$$

- Proof: 相当于基于$\psi = \delta_c$实现IRL和RL的嵌套过程

define indicator function δ_c: $R^{S\times A} \to R$ (示性函数):

$$\delta_c = 0 \; if \; c \in \mathrm{C} \; and \; +\infty \text{ otherwise}$$

$$\max_{c \in \mathcal{C}} E_{\pi}[c(s,a)] - E_{\pi_E}[c(s,a)] = \max_{c \in R^{\mathcal{S}\times\mathcal{A}}} -\delta_c(c) + \sum_{s,a} \left(\rho_{\pi}(s,a) - \rho_{\pi_E}(s,a)\right) c(s,a) = \delta_c^*(\rho_{\pi} - \rho_{\pi_E})$$

Conclusion Apprenticeship Learning

- Cons: too restrictive

  当真实的cost function 不包含在C中时, 学徒学习无法学到比专家更好的策略

- Pros:

  可以拓展到高维动作和状态空间

$$\nabla_\theta \max_{c \epsilon C} E_{\pi_\theta}[c(s,a)] - E_{\pi_E}[c(s,a)] = \nabla_\theta\ E_{\pi_\theta}[c^*(s,a)] = E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q_{c^*}(s,a)]$$

where $c^* = \underset{c \epsilon \mathcal{C}}{\operatorname{argmax}} E_{\pi_\theta}[c(s,a)] - E_{\pi_E}[c(s,a)]$, $Q_{c^*}(\bar{s},\bar{a}) = E_{\pi_\theta}[c^*(\bar{s},\bar{a})|s_0 = \bar{s}, a_0 = \bar{a}]$

fit a cost function c_i^* using π_{θ_i} $\rightleftarrows$ take TRPO step to produce $\pi_{\theta_{i+1}}$ using c_i^*

GAIL

常数正则项可完全匹配专家的occupancy measure，但无法在高维空间中实现

示性函数正则项适用于高维空间，但cost function过于局限

$$\psi_{\mathrm{GA}}(c) \triangleq \begin{cases} \mathbb{E}_{\pi_E}[g(c(s,a))] & \text{if } c < 0 \\ +\infty & \text{otherwise} \end{cases} \quad \text{where } g(x) = \begin{cases} -x - \log(1 - e^x) & \text{if } x < 0 \\ +\infty & \text{otherwise} \end{cases}$$

$$\psi^*_{\mathrm{GA}}(\rho_\pi - \rho_{\pi_E}) = \sup_{D \in (0,1)^{\mathcal{S} \times \mathcal{A}}} \mathbb{E}_\pi[\log(D(s,a))] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))]$$

GAIL

Goal: 最小化其与专家策略的occupancy measure的JS散度

$$\min_{\theta}\max_{\omega}\mathbb{E}_{\pi_\theta}[\log(D_\omega(s,a))]+\mathbb{E}_{\pi_E}[\log(1-D_\omega(s,a))]-\lambda H(\pi_\theta)$$

Algorithm 1 Generative adversarial imitation learning

1: **Input:** Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters θ_0, w_0
2: **for** $i = 0, 1, 2, \ldots$ **do**
3: Sample trajectories $\tau_i \sim \pi_{\theta_i}$
4: Update the discriminator parameters from w_i to w_{i+1} with the gradient

寻找鞍点(π, D)

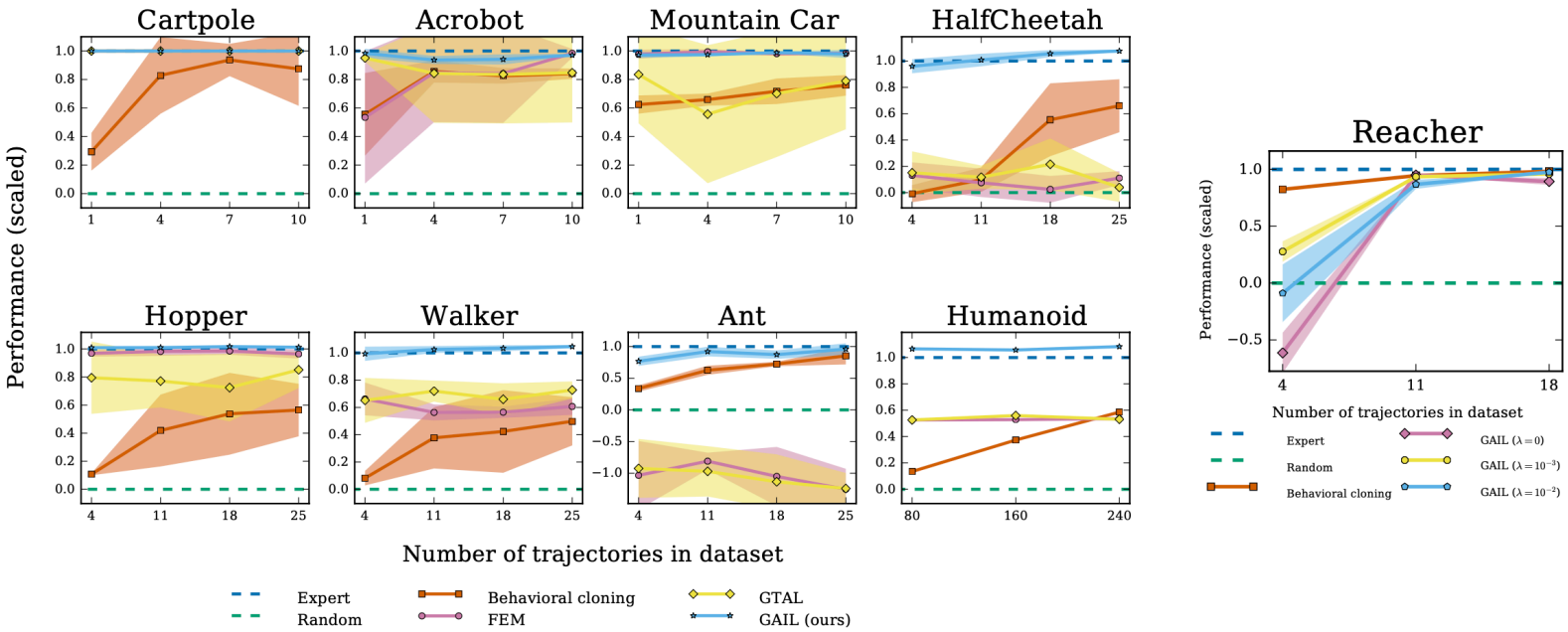
$$\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\nabla_w \log(1 - D_w(s,a))] \tag{17}$$

5: Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $\log(D_{w_{i+1}}(s,a))$. Specifically, take a KL-constrained natural gradient step with

$$\begin{aligned}&\hat{\mathbb{E}}_{\tau_i}[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)] - \lambda \nabla_\theta H(\pi_\theta),\\ &\text{where } Q(\bar{s},\bar{a}) = \hat{\mathbb{E}}_{\tau_i}[\log(D_{w_{i+1}}(s,a)) \mid s_0 = \bar{s}, a_0 = \bar{a}]\end{aligned} \tag{18}$$

6: **end for**

Experiments



(a) (b)

Figure 1: (a) Performance of learned policies. The y-axis is negative cost, scaled so that the expert achieves 1 and a random policy achieves 0. (b) Causal entropy regularization λ on Reacher. Except for Humanoid, shading indicates standard deviation over 5-7 reruns.

Thanks