强化学习2022
第10节

涉及知识点：
模仿学习、行为克隆、逆强化学习、生成对抗模仿学习

模仿学习

张伟楠 – 上海交通大学

课程大纲

强化学习基础部分

1. 强化学习、探索与利用
2. MDP和动态规划
3. 值函数估计
4. 无模型控制方法
5. 规划与学习
6. 参数化的值函数和策略
7. 深度强化学习价值方法
8. 深度强化学习策略方法

强化学习前沿部分

9. 基于模型的深度强化学习
10. 模仿学习
11. 离线强化学习
12. 参数化动作空间
13. 目标导向的强化学习
14. 多智能体强化学习
15. 强化学习大模型
16. 技术交流与回顾

模仿学习

张伟楠 – 上海交通大学

什么是模仿学习

- 模仿学习：从专家范例中学习出一个好的策略
  - 仅给出专家轨迹
  - 奖励函数未知
- 示例
  - 自动驾驶
  - 机器人控制
- 为什么？（难点）
  - 在某些任务中难以定义奖励函数
  - 人为设置的奖励函数可能会导致不合理的行为

模仿学习与监督学习相比较

- 需要求解的问题可能要求其解具有重要的结构性特征：
  - 包括约束（例如，机器人的关节限制），动态平滑性和稳定性，或要求能得到连贯的多步计划

- 需要考虑决策者所做出的决策和其接受的输入分布之间的关系（是在线策略或是离线策略）

- 收集数据通常需要较高的成本，模仿学习能够体现最小化这一成本的必要性

S. Shalev-Shwartz and S. Ben-David. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.

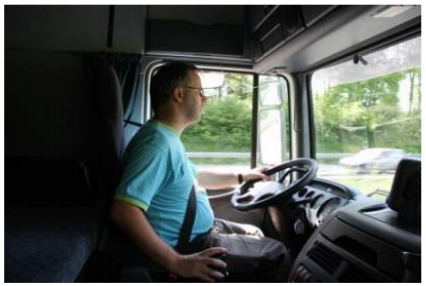
模仿学习算法

- 行为克隆（Behavior Cloning）
  - 在无需奖励函数的情况下，学习从状态/上下文关系到轨迹/动作的直接映射
- 逆强化学习（Inverse Reinforcement Learning）
  - 找到能够使专家策略比任何其他策略更优的奖励函数
- 生成对抗模仿学习（Generative Adversarial Imitation Learning）
  - 通过生成对抗网络的方式自动学习一个好的奖励函数

行为克隆

张伟楠 – 上海交通大学

行为克隆



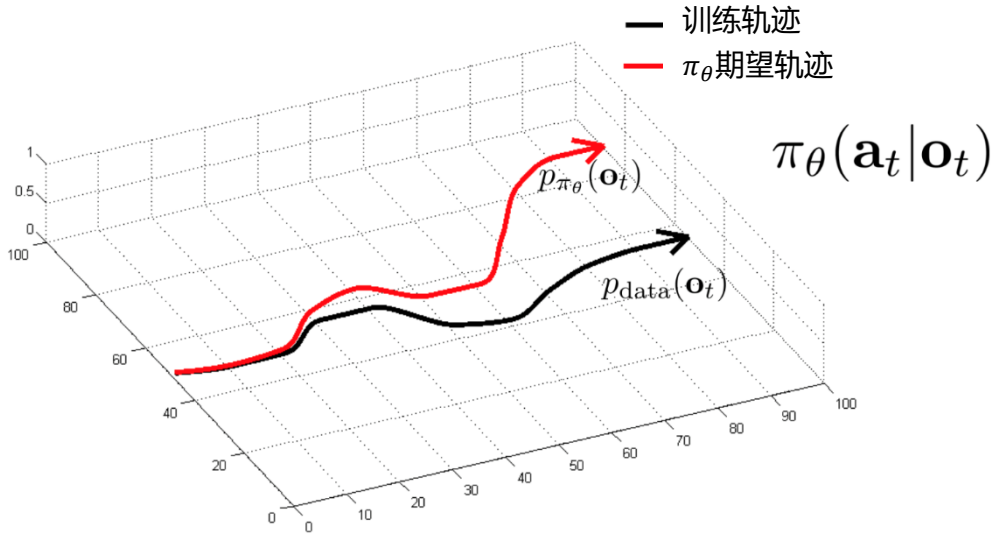
$\mathbf{o}_t$

$\mathbf{a}_t$

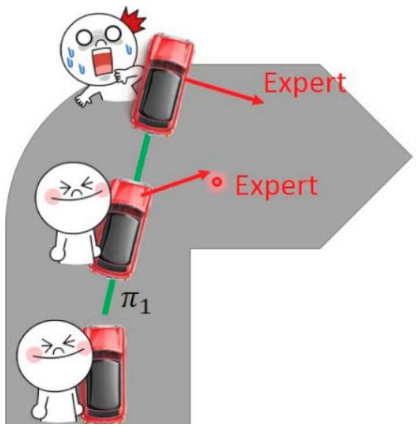
训练数据

监督学习

$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$



分布上的偏移



行为克隆

- 从一个稳定的轨迹分布中采样
  - 从稳定的控制器学习策略（有噪声）
- 加入更多的在线策略数据
  - 例如：使用DAgger技术

DAgger：数据融合（Dataset Aggregation）

① 使用人类数据训练$\pi_\theta(a_t, o_t)$

$$\mathcal{D} = \{o_1, a_1, \dots, o_N, a_N\}$$

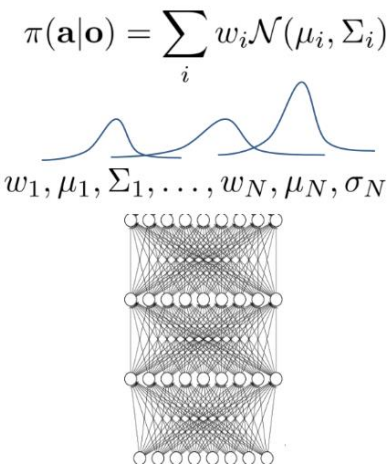
② 执行策略$\pi_\theta(a_t|o_t)$获取数据集$\mathcal{D}_\pi = \{o_1, a_1, \dots, o_N, a_N\}$

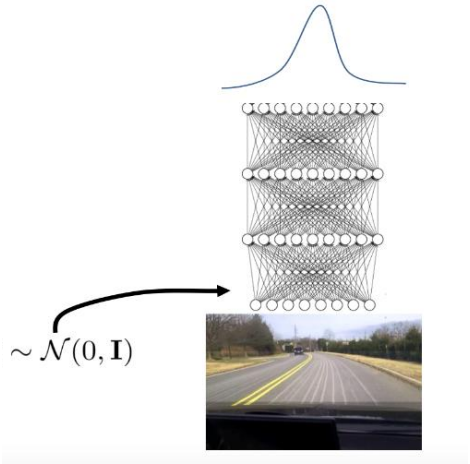
③ 让人类选择动作a_t标注数据集$\mathcal{D}_\pi$

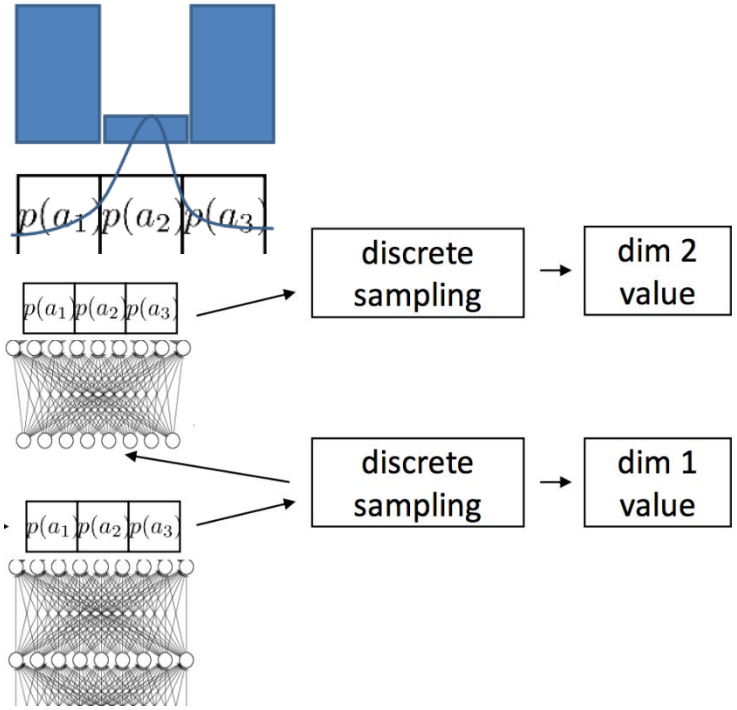
④ 融合数据集$\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$

行为克隆

- 除了分布上的偏移，可能面临的挑战
  - 不具有马尔科夫性质的行为
    - $\pi_\theta(a_t|o_t)$中的行为只取决于当前的观测
    - $\pi_\theta(a_t|o_1, o_2, \ldots, o_t)$中的行为取决于所有过去的观测
  - 多模态的行为
    - 输出混合高斯分布
    - 隐变量模型
    - 自回归的离散化







逆强化学习

张伟楠 – 上海交通大学

逆强化学习

正向强化学习

- 给定
  - 状态 $s \in S$
  - 动作 $a \in A$
  - 奖励函数 $r(s,a)$
  - 可能给出
    - 状态转移 $p(s'|s,a)$
- 学习最优策略 $\pi^*(a|s)$

逆强化学习

- 给定
  - 状态 $s \in S$
  - 动作 $a \in A$
  - 可能给出
    - 状态转移 $p(s'|s,a)$
  - 按照策略$\pi^*(a|s)$采样得到的轨迹$\{\tau_i\}$
- 学习$r_\psi(s,a)$并使用它学习$\pi^*(a|s)$

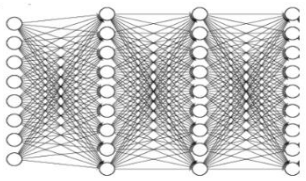
线性奖励函数

$$r_\psi(s,a) = \sum_i \psi_i f_i = \psi^T f(s,a)$$

神经网络表示的奖励函数

s

a



$r_\psi(s,a)$ 其中参数为ψ

逆强化学习

- 最大因果熵逆强化学习（Maximum casual entropy IRL）
  - 首先，找到一个代价函数$\tilde{c}$使得专家策略产生较低代价，而其他策略会产生较高的代价

$$\begin{aligned}\tilde{c} &= \mathrm{IRL}(\pi_E) \\ &= \arg\max_{c \in C} \left(\min_{\pi} -H(\pi) + \mathbb{E}_{\pi}[c(s,a)] \right) - \mathbb{E}_{\pi_E}[c(s,a)]\end{aligned}$$

  - 然后，通过使用代价函数$\tilde{c}$进行正向强化学习，尝试恢复专家策略

$$\tilde{\pi} = \mathrm{RL}(\tilde{c}) = \arg\min_{\pi} -H(\pi) + \mathbb{E}_{\pi}[\tilde{c}(s,a)]$$

- 优势
  - 使用神经网络计算成本函数
- 局限性
  - 仍需要反复求解马尔可夫决策过程
  - 假设环境已知

Ziebart B D et al. Maximum Entropy Inverse Reinforcement Learning. AAAI. 2008.

生成对抗模仿学习

张伟楠 – 上海交通大学

生成对抗模仿学习

- 生成对抗模仿学习（GAIL）

$$\min_{\pi}\max_{D}\mathbb{E}_{(s,a)\sim\pi_E}[\log D(s,a)]+\mathbb{E}_{(s,a)\sim\pi}\left[\log\left(1-D(s,a)\right)\right]-\lambda H(\pi)$$

- 生成对抗网络（GAN）

$$\min_{G}\max_{D}\mathbb{E}_{x\in p_{data}(x)}[\log D(x)]+\mathbb{E}_{z\sim p_z(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right]$$

- 判别器D判别数据分布是由生成器G（即生成对抗模仿学习中的π）还是真实数据分布（即生成对抗学习中的π_E）产生

Ho J, Ermon S. Generative adversarial imitation learning. NIPS 2016.

生成对抗网络与逆强化学习之间的联系

- 生成对抗网络（GAN）
  1. 训练出一个良好的判别器
  2. 训练出一个良好的生成器以欺骗上述判别器

$$\min_G \max_D \mathbb{E}_{x \in p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right]$$

- 逆强化学习（IRL）
  1. 训练出一个良好的代价函数
  2. 训练出一个良好的基于上述代价函数的策略以接近专家策略

$$\tilde{c} = \mathrm{IRL}(\pi_E) = \arg\max_{c \in C}\left(\min_{\pi} -H(\pi) + \mathbb{E}_{\pi}[c(s,a)]\right) - \mathbb{E}_{\pi_E}[c(s,a)]$$

$$\tilde{\pi} = \mathrm{RL}(\tilde{c}) = \arg\min_{\pi} -H(\pi) + \mathbb{E}_{\pi}[\tilde{c}(s,a)]$$

http://tsong.me/blog/gan-irl-ebm/

模仿学习的开放式问题

- 与算法相关的问题
  - 如何泛化复杂条件下的策略?
  - 如何有保证地求得解?
  - 如何根据维度的数量进行扩展?
  - 如何在高维空间中找到全局最优解? 如何做最容易?
  - 如何实现多智能体的模仿学习?
  - 如何在逆强化学习中实现增量/主动学习?
- 表现评估
  - 如何建立模仿学习的评估基准问题?
  - 应该使用什么指标评估模仿学习的表现?

THANK YOU