WDS

DataScience@Web

模仿学习--逆向强化学习

INVERSE REINFORCEMENT LEARNING (IRL)

汇报人：耿飚

2021.8.3

2021 WDS暑期讨论班

目录

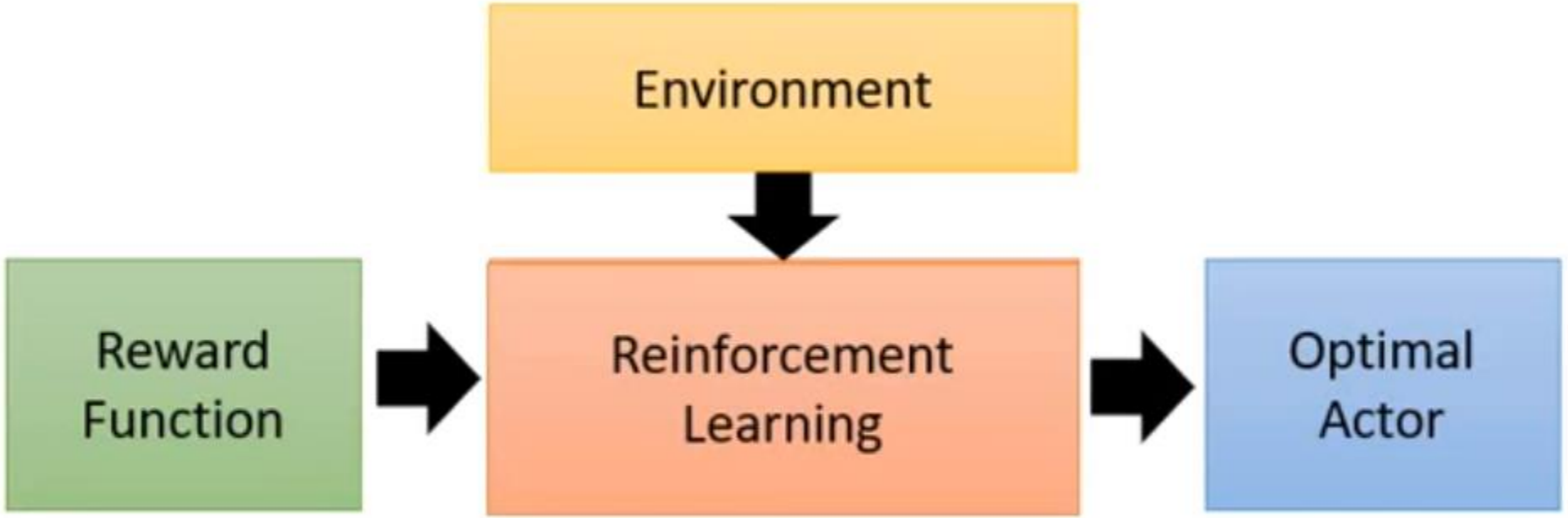
- 模仿学习(Imitation learning)
- 逆向强化学习(Inverse Reinforcement Learning)
- 总结

模仿学习(Imitation learning)

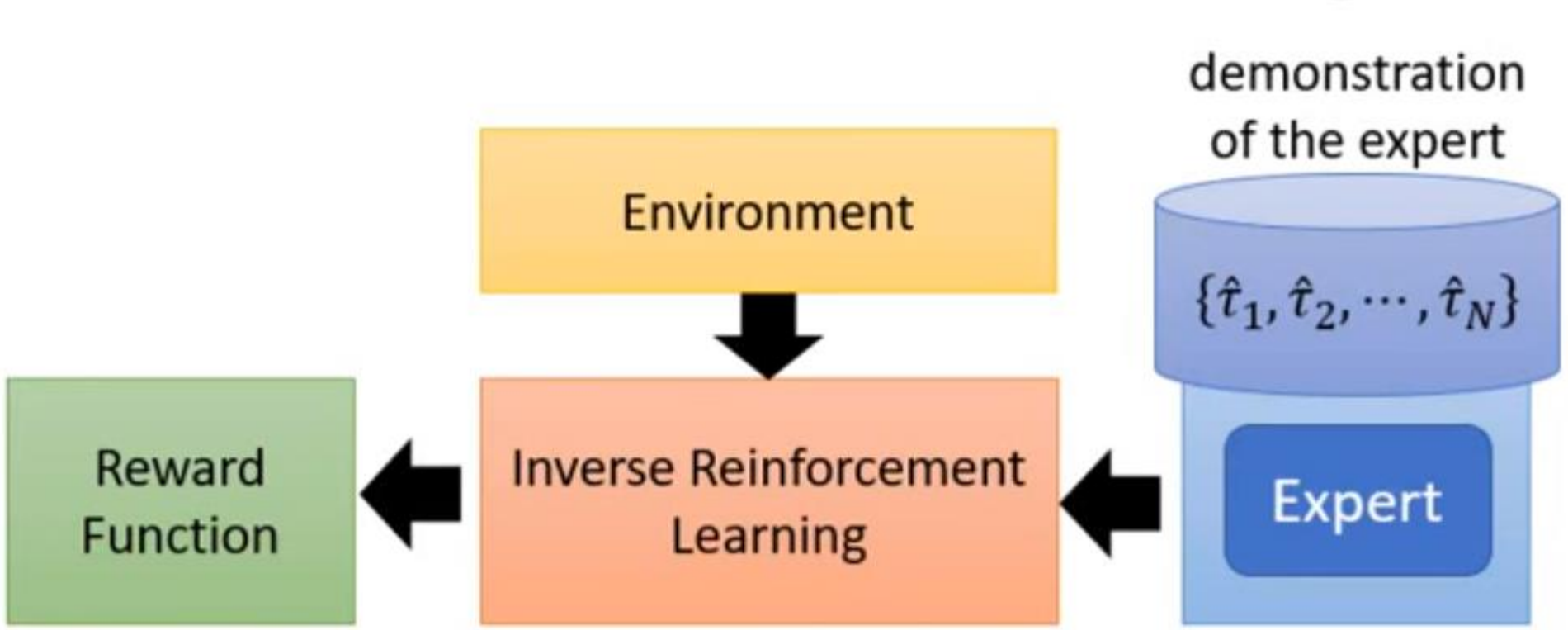
- Imitation learning
  - Also known as learning by demonstration，apprenticeship learning
- An expert demonstration how to solve the task
  - Machine can also interact with the environment，but cannot explicitly obtain reward
  - It is hard to define reward in some tasks
  - Hand-crafted rewards can lead to uncontrolled behavior
- Two approaches
  - Behavior cloning
  - Inverse reinforcement learning

目录

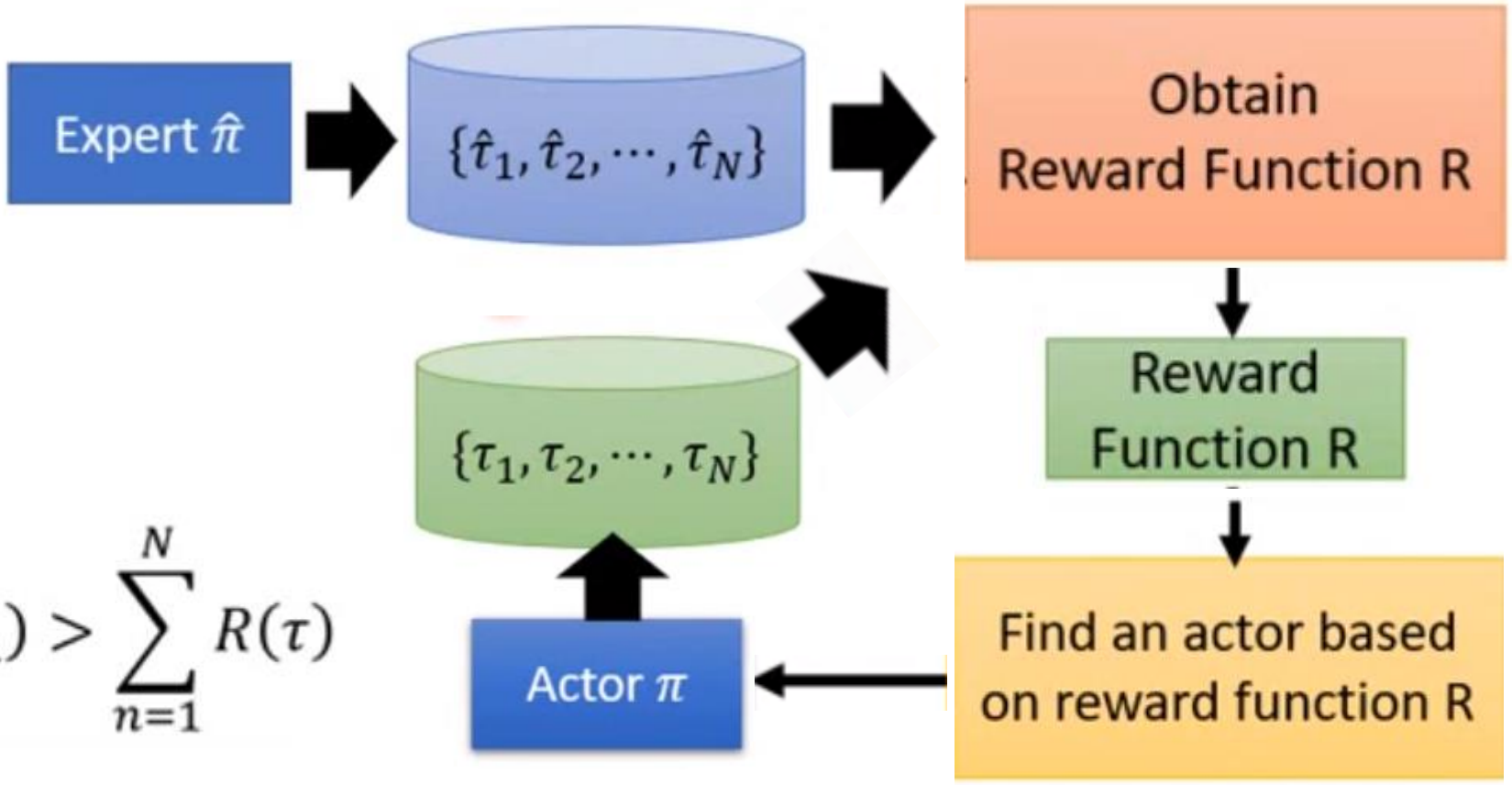
逆向强化学习(Inverse Reinforcement Learning)



逆向强化学习(Inverse Reinforcement Learning)



逆向强化学习(Inverse Reinforcement Learning)



$$\sum_{n=1}^{N} R(\hat{\tau}_n) > \sum_{n=1}^{N} R(\tau)$$

逆向强化学习(Inverse Reinforcement Learning)

- 最大边际形式化方法:
  - 学徒学习、MMP方法、结构化分类、神经逆向强化学习
- 基于概率模型的形式化方法:
  - 最大熵IRL、相对熵IRL、深度逆向强化学习

逆向强化学习(Inverse Reinforcement Learning)

最大熵逆向强化学习

τ：表示收集的某一条轨迹

D：表示轨迹数据集

$R_\psi(\tau)$ ：表示回报函数

$Z = \int exp(R_\psi(\tau))d\tau$ ： 是partition function

也叫做划分函数

Maximum Entropy Inverse RL

(Ziebart et al. '08)

handle ambiguity using probabilistic model of behavior

Notation:

$\tau = \{s_1, a_1, ..., s_t, a_t, ..., s_T\}$ trajectory

$R_\psi(\tau) = \sum_t r_\psi(s_t, a_t)$ learned reward

$\mathcal{D} : \{\tau_i\} \sim \pi^*$ expert demonstrations

MaxEnt formulation:

$$p(\tau) = \frac{1}{Z}\exp(R_\psi(\tau))$$

$$\max_\psi \sum_{\tau \in \mathcal{D}} \log p_{r_\psi}(\tau)$$

知乎 @郑思座

逆向强化学习(Inverse Reinforcement Learning)

最大熵逆向强化学习

展开所要优化的目标函数

并对参数 ψ 求导

Maximum Entropy IRL Optimization

$$\nabla_\psi \mathcal{L}(\psi) = \sum_{\tau \in \mathcal{D}} \frac{dR_\psi(\tau)}{d\psi} - M \frac{1}{\sum_\tau \exp(R_\psi(\tau))} \sum_\tau \exp(R_\psi(\tau)) \frac{dR_\psi(\tau)}{d\psi}$$

$$\sum_\tau p(\tau \mid \psi) \frac{dR_\psi(\tau)}{d\psi}$$

$$\sum_{\mathbf{s}} p(\mathbf{s} \mid \psi) \frac{dr_\psi(\mathbf{s})}{d\psi}$$

$p(s|\psi)$

逆向强化学习(Inverse Reinforcement Learning)

最大熵逆向强化学习

Maximum Entropy Inverse RL

(Ziebart et al. '08)

Algorithm:

0. Initialize ψ, gather demonstrations $\mathcal{D}$
1. Solve for optimal policy $\pi(\mathbf{a}|\mathbf{s})$ w.r.t. reward r_ψ
2. Solve for state visitation frequencies $p(\mathbf{s}|\psi)$
3. Compute gradient $\nabla_\psi \mathcal{L} = -\frac{1}{|\mathcal{D}|}\sum_{\tau_d \in \mathcal{D}} \frac{dr_\psi}{d\psi}(\tau_d) - \sum_s p(s|\psi)\frac{dr_\psi}{d\psi}(s)$
4. Update ψ with one gradient step using $\nabla_\psi \mathcal{L}$

目录

- 模仿学习(Imitation learning)
- 逆向强化学习(Inverse Reinforcement Learning)
- 总结

总结

- Advantages：
  - It solves the difficulty of artificially designing the reward function
  - The reward function is more realistic
- Disadvantages：
  - Learning efficiency is not high
  - The quality of the reward function is difficult to assess

Thanks!