SQIL: Imitation Learning via Reinforcement Learning with Sparse Rewards——hoho

论文试图解决什么问题？

为了学到reward function，目前的方法有如下问题：

1. 基于监督学习的行为克隆方法，没考虑到专家数据的分布偏移问题：agent总是贪婪的模仿专家的行为，由于误差的积累它会渐渐偏移专家数据中状态，而且此时（out-of-distribution states），它也无法知道如何回到专家数据中的状态。
2. 一些模仿学习方法，如GAIL，需要用到对抗训练，实践中难以实现与应用

这是否是一个新的问题？

否

这篇文章要验证一个什么科学假设？

不需要使用对抗训练方法或学习一个reward function，也可以解决分布偏移问题，而且性能还和对抗训练方法差不多，实现简单。

有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员？

hoho_todo

论文中提到的解决方案之关键是什么？

1. 对专家数据的s,a都给奖励+1，对新交互采样的s,a都给0
2. 初始化经验池为专家示例，维护经验池为50%专家示例，50%为采样数据

3. 用以上技巧，修改soft Q learning来进行off-policy学习
4. 由于是用off-policy学习，因此不需要真正采样到相应的状态才可以学习，因此可以扩展到高维和连续状态空间中去

算法过程如下：

Algorithm 1 Soft Q Imitation Learning (SQIL)

1: Require $\lambda_{\text{samp}} \in \mathbb{R}_{\geq 0}$
2: Initialize $\mathcal{D}_{\text{samp}} \leftarrow \emptyset$
3: **while** $Q_{\boldsymbol{\theta}}$ not converged **do**
4: $\quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} (\delta^2(\mathcal{D}_{\text{demo}}, 1) + \lambda_{\text{samp}} \delta^2(\mathcal{D}_{\text{samp}}, 0))$ {See Equation 1}
5: $\quad$ Sample transition (s, a, s') with imitation policy $\pi(a|s) \propto \exp(Q_{\boldsymbol{\theta}}(s, a))$
6: $\quad \mathcal{D}_{\text{samp}} \leftarrow \mathcal{D}_{\text{samp}} \cup \{(s, a, s')\}$
7: **end while**

soft贝尔曼误差的平方为：

$$\delta^2(\mathcal{D}, r) \triangleq \frac{1}{|\mathcal{D}|} \sum_{(s,a,s') \in \mathcal{D}} \left(Q_{\boldsymbol{\theta}}(s, a) - \left(r + \gamma \log \left(\sum_{a' \in \mathcal{A}} \exp\left(Q_{\boldsymbol{\theta}}(s', a')\right) \right) \right) \right)^2,$$

根据soft贝尔曼方程：

$$Q(s, a) \triangleq R(s, a) + \gamma \mathbb{E}_{s'} \left[\log \left(\sum_{a' \in \mathcal{A}} \exp\left(Q(s', a')\right) \right) \right].$$

可以得到奖励的计算公式：

$$R_q(s, a) \triangleq Q_{\boldsymbol{\theta}}(s, a) - \gamma \mathbb{E}_{s'} \left[\log \left(\sum_{a' \in \mathcal{A}} \exp\left(Q_{\boldsymbol{\theta}}(s', a')\right) \right) \right]$$

本文的理论分析是解释SQIL是一种regularized BC

$$\ell_{\mathrm{RBC}}(\boldsymbol{\theta}) \triangleq \ell_{\mathrm{BC}}(\boldsymbol{\theta}) + \lambda\delta^2(\mathcal{D}_{\mathrm{demo}} \cup \mathcal{D}_{\mathrm{samp}}, 0),$$

并且证明其梯度与SQIL成正比：

$$\nabla_{\boldsymbol{\theta}}\ell_{\mathrm{RBC}}(\boldsymbol{\theta}) \propto \nabla_{\boldsymbol{\theta}}\left(\delta^2(\mathcal{D}_{\mathrm{demo}}, 1) + \lambda_{\mathrm{samp}}\delta^2(\mathcal{D}_{\mathrm{samp}}, 0) + V(s_0)\right)$$

论文中的实验是如何设计的？

hoho_todo

用于定量评估的数据集是什么？代码有没有开源？

hoho_todo

论文中的实验及结果有没有很好地支持需要验证的科学假设？

hoho_todo

这篇论文到底有什么贡献？

hoho_todo

下一步呢？有什么工作可以继续深入？

hoho_todo