强化学习作业三

周舒航
zhoush@mail.ustc.edu.cn

2024 年 12 月 24 日

离线强化学习 (Offline Reinforcement Learning)

定义:

- 离线强化学习是在一个固定的、预先收集的数据集上训练策略，而无需与环境进行交互。

关键特点:

- 数据是静态的，通常由其他策略或专家生成。
- 无需在线探索，降低了采样成本和潜在的风险。

挑战:

- 数据分布偏移 (Distribution Shift): 训练数据与策略评估的数据可能不同。
- 数据质量问题: 数据可能包含噪声或次优行为。

应用场景:

- 医疗决策、机器人学习、推荐系统等需要高安全性和低探索成本的领域。

Offline Reinforcement Learning

Definition:

- Offline reinforcement learning trains policies using a fixed, pre-collected dataset without interacting with the environment.

Key Characteristics:

- The data is static and typically generated by other policies or experts.
- No online exploration is required, reducing sampling costs and potential risks.

Challenges:

- Distribution Shift: The training data may differ from the data used during policy evaluation.
- Data Quality Issues: The dataset may contain noise or suboptimal behavior.

Application Scenarios:

- Fields requiring high safety and low exploration costs, such as medical decision-making, robotic learning, and recommendation systems.

CQL：保守 Q-learning 算法

核心思路

- 在离线场景下，数据集通常由有限的策略生成。若直接使用常规 Q-learning 方法，容易对数据集中较少出现或分布外（Out-of-distribution, OOD）的状态-动作对产生过高估计。
- 为避免策略在这些分布外动作上“捡漏”而导致过拟合，CQL 在目标函数中额外加入了一项“保守项”，在更新 Q 函数时尽量降低对分布外动作的 Q 值估计。

优势

- 通过抑制对稀有或分布外动作的过度乐观估计，使策略更保守，在无法继续采集新数据的前提下，保证学到策略不会依赖数据外动作获取虚高回报。

CQL: Conservative Q-Learning Algorithm

Core Idea

- In offline scenarios, the dataset is usually generated by a limited set of policies. Using standard Q-learning methods directly may lead to overestimation of state-action pairs that appear infrequently or are out-of-distribution (OOD).
- To prevent the policy from "overfitting" by exploiting these OOD actions, CQL introduces a "conservative term" in the objective function. This term reduces the Q-value estimates for OOD actions during Q function updates.

Advantages

- By suppressing overly optimistic estimates of rare or OOD actions, the policy becomes more conservative, ensuring that the learned policy does not rely on OOD actions to achieve unrealistically high rewards, especially when new data cannot be collected.

CQL 中的损失函数与优化目标

Q 函数整体结构：

$$
\begin{aligned}
\min_{Q} \max_{\mu} \alpha \Big(& \mathbb{E}_{s\sim\mathcal{D},\, a\sim\mu(a|s)} \big[Q(s,a) \big] \;-\; \mathbb{E}_{s\sim\mathcal{D},\, a\sim\hat{\pi}_{\beta}(a|s)} \big[Q(s,a) \big] \Big) \\
& + \frac{1}{2} \mathbb{E}_{s,a,s'\sim\mathcal{D}} \Big[\big(Q(s,a) \;-\; \hat{B}^{\pi^k} Q^k(s,a) \big)^2 \Big] \\
& + R(\mu).
\end{aligned}
$$

- $\mathcal{D}$ 表示离线数据集，$\hat{\pi}_{\beta}$ 可看作“行为策略”（行为策略的估计或先验），μ 表示当前在优化的学习策略（你希望最终得到的策略）。
- 从常规 Q-learning 的 Bellman 误差出发，**为避免分布外动作的过高估计**，引入了保守项，并使用 α 控制保守程度。
- 在保证 Bellman 误差最小化的同时，施加正则项 $R(\mu)$，使得最终得到的策略不会过度偏离可行的数据分布。

Loss Function and Optimization Objective in CQL

Overall Structure of the Q Function:

- $\mathcal{D}$ represents the offline dataset, $\hat{\pi}_\beta$ can be considered the "behavior policy" (an estimate or prior of the behavior policy), and μ represents the learning policy currently being optimized (the policy you aim to obtain).
- Starting from the standard Bellman error in Q-learning, to **avoid overestimation of out-of-distribution actions**, a conservative term is introduced, with α controlling the degree of conservativeness.
- While minimizing the Bellman error, a regularization term $R(\mu)$ is applied to ensure that the resulting policy does not deviate excessively from the feasible data distribution.

CartPole-v1 环境简介

CartPole-v1 是 OpenAI Gym 提供的经典强化学习环境，特点如下：

- **任务目标**：控制一个小车（Cart），通过施加左右方向的力，使得杆（Pole）保持直立并不倒下。
- **状态空间**：由 4 个连续变量组成：小车的位置 x 小车的速度 $\dot{x}$；杆的角度 θ；杆的角速度 $\dot{\theta}$。
- **动作空间**：离散动作空间 $\{0, 1\}$。0：向左施加力；1：向右施加力。
- **奖励函数**：每一步杆保持直立（$|\theta| < 12^\circ$ 且 $|x| < 2.4$），奖励为 $+1$，否则环境结束。
- **结束条件**：
  - 杆的角度超出 12°；
  - 小车位置超出轨道边界（$|x| > 2.4$）；
  - 累积时间步达到最大值 500。

Introduction to the CartPole-v1 Environment

CartPole-v1 is a classic reinforcement learning environment provided by OpenAI Gym, with the following characteristics:

- **Task Objective**: Control a cart by applying forces to the left or right, keeping the pole upright and preventing it from falling.
- **State Space**: Consists of 4 continuous variables: the cart's position x, the cart's velocity $\dot{x}$, the pole's angle θ, and the pole's angular velocity $\dot{\theta}$.
- **Action Space**: A discrete action space $\{0, 1\}$. 0: Apply force to the left; 1: Apply force to the right.
- **Reward Function**: A reward of $+1$ is given for every timestep the pole remains upright ($|\theta| < 12^\circ$ and $|x| < 2.4$). Otherwise, the episode ends.
- **Termination Conditions**:
  - The pole's angle exceeds 12°;
  - The cart's position exceeds the track boundaries ($|x| > 2.4$);
  - The cumulative timestep reaches the maximum value of 500.

作业要求（100 分，加权后算入课程总分）

要求如下：

1. **算法实现与训练（40 分）：**
   - 实现 CQL 算法；
   - 调整训练参数，记录训练结果。
2. **代码阐明（20 分）：**
   - 在实验报告中详细阐明代码实现的关键步骤与逻辑。
3. **实验报告分析（40 分）：**
   - 对比 CQL 算法与普通 Q-learning 算法的训练结果；
   - 评估不同数据集对训练结果的影响（通过采样降低数据集大小，或使用表现较差的数据集进行实验）；
   - 分析 Q 函数公式中不同参数对 CQL 算法训练结果的影响。
4. **额外加分（Bonus，二选一即可）：**
   - 实现行为克隆（Behavior Cloning, BC）算法；
   - 或实现离线强化学习中的 BCQ 算法；

   请在 `dataset_episode_350.npz` 数据集上记录至少一个实验结果。

Assignment Requirements (100 points, weighted into final course grade)

Requirements are as follows:

1. **Algorithm Implementation and Training (40 points)**:
   - Implement the CQL algorithm;
   - Tune training parameters and record the training results.
2. **Code Explanation (20 points)**:
   - Provide a detailed explanation of the key steps and logic in the code implementation in the experiment report.
3. **Experiment Report Analysis (40 points)**:
   - Compare the training results of the CQL algorithm and standard Q-learning algorithm;
   - Evaluate the impact of different datasets on training results (e.g., by reducing the dataset size through sampling or using suboptimal datasets for experiments);
   - Analyze the effect of different parameters in the Q-function formulation on the training results of the CQL algorithm.
4. **Extra Credit (Bonus, choose one)**:
   - Implement the Behavior Cloning (BC) algorithm;
   - Or implement the BCQ algorithm in offline reinforcement learning.

作业提交要求

提交截止日期：2025 年 1 月 17 日 23:59:59（UTC+8）

提交方式： 通过 **BB（Blackboard）平台提交。**

提交要求：

- 文件需以 .zip 格式提交；
- 压缩包内容需包含：
  - 至少一个代码文件；
  - 实验报告（必须为 **PDF** 格式，编写方式不限）。
- 压缩包命名格式：<student_id>_<name>_exp3.zip，学号务必大写。示例：SA23011281_周舒航_exp3.zip

注意： 请确保提交内容完整并符合命名要求，以免影响评分。

Assignment Submission Requirements

Submission Deadline: January 17, 2025, 23:59:59 (UTC+8)

Submission Method: Submit through the **BB (Blackboard) platform**.

Submission Requirements:

- Files must be submitted in **.zip** format;
- The compressed file must include:
  - At least one code file;
  - An experiment report (must be in **PDF** format, writing style is flexible).
- File naming format: <student_id>_<name>_exp3.zip, with the student ID in uppercase. Example: SA23011281_ZhouShuhang_e x p.zip

Note: Ensure that the submission is complete and follows the naming conventions to avoid affecting your grade.