Reinforcement Learning

Reported by ALISURE

Date 2019/11/16

1. 基础知识

2. 求解强化学习

3. 求解强化学习进阶

4. 深度强化学习

基础知识

基本组成元素：智能体、环境、状态、动作和奖励

- 智能体通过状态、动作、奖励与环境进行交互。

两类环境模型：基于模型（Model-based）和免模型（Model-free）

- 基于模型：智能体已经对环境进行建模。在状态s下执行动作a转移到状态s’的概率P_sa是已知的，其状态转移所带来的奖励r同样是已知的。
- 免模型：环境的状态转移概率P、奖励函数R难以提前获取，也不一定知道环境中有多少个状态。

- 基于模型的任务：使用动态规划来求解
- 免模型的任务：使用蒙特卡洛法、时间差分法、值函数近似法、策略梯度法来求解

探索与利用和预测与控制：

- 探索：尝试还未尝试过的动作行为
- 利用：从已知动作中选择下一步的行动
- 预测：验证未来，给定一个策略，智能体需要去验证该策略能够到达的理想状态值，以确定该策略的好坏。
- 控制：优化未来，给出一个初始化策略，智能体希望基于该策略找到一个最优的策略。

MP：(S,P) 马尔可夫过程
MRP：(S,P,R) 马尔可夫奖励过程
MDP：(S,A,P,R)马尔可夫决策过程

S：状态空间集
A：动作空间集
P：状态转移概率
R：奖励函数

求解强化学习

- 策略
  - 确定性策略$a = \pi(s)$，策略根据状态s选择动作a
  - 随机性策略$\pi(s, a)$，策略在状态s下选择动作a的概率
- 奖励
  - 总奖励
  - 未来累积奖励
  - 折扣未来累积奖励
- 价值函数：评估当前智能体在该时间步状态的好坏程度
  - 状态值函数：v(s)，在状态s下，执行动作得到的奖励期望
  - 动作值函数：q(s,a)，在状态s下，执行动作a的好坏程度
- 贝尔曼方程：当前时刻状态的价值$v(s_t)$和下一时刻状态的价值$v(s_{t+1})$的关系

$$v = R + \gamma P v$$

- 最优值函数
  - 最优状态值函数：$v^*(s) = \max_\pi v(s)$
  - 最优动作值函数：$q^*(s, a) = \max_\pi q_\pi(s, a)$
- 最优策略：可通过最优值函数得到
- ε-贪婪算法：在探索和利用之间权衡
  - 探索：以概率ε随机选择一个动作
  - 利用：以概率1-ε选择奖励最高的动作

求解强化学习：方法

基于模型的任务

- 动态规划
  - 策略迭代
    - 策略评估
    - 策略改进
  - 值迭代

$$v(s_t) \leftarrow \sum \pi(a|s) \sum p(s'|s,a)\big(r(s'|s,a) + \gamma v(s')\big)$$

免模型的任务

- 蒙特卡洛法
  - 经验轨迹
  - 蒙特卡洛预测
  - 蒙特卡洛评估
  - 蒙特卡洛控制
- 时间差分法
  - $TD(\lambda)$
  - Sarsa算法
  - Q-learning算法

$$v(s_t) \leftarrow v(s_t) + \alpha\big(G_t - v(s_t)\big)$$

$$v(s_t) \leftarrow v(s_t) + \alpha\big(r_{t+1} + \gamma v(s_{t+1}) - v(s_t)\big)$$

求解强化学习进阶：近似求解法

表格求解方法

- 动态规划法、蒙特卡洛法和时间差分法属于表格求解方法。
- 所有的<状态-动作>对都可以存储在有限的表格上，并且可以通过特殊的方法进行枚举。

大规模强化学习

- 迭代次数较多也无法保证值函数v(s)或状态值函数q(s,a)在计算过程中能正确地收敛。
- 当<状态-动作>对趋于无限时，无法存储大量的状态值以及进行有效地索引枚举。

近似求解法

- 寻找优化目标的近似函数而非原函数
- 在保证求解结果有效性的前提下大大降低了计算的规模和复杂度。

求解强化学习进阶：近似求解法

- 值函数近似
  - 基于价值的强化学习任务求解方法：对价值函数进行近似
- 策略梯度法
  - 基于策略的强化学习任务求解方法：对策略函数进行近似并求解其梯度
- 学习与规划
  - 基于模型的强化学习任务求解方法：使用函数通过采样方法模拟环境模型

1. 值函数近似

通过寻找状态值V或动作值Q的近似替代函数$\hat{v}(s,w)$或$\hat{q}(s,a,w)$的方式来求解大规模强化学习任务。

通过寻找值函数的近似函数来解决大规模马尔可夫决策过程：

$$\hat{v}(s,w) \approx v_\pi(s) \text{ 或 } \hat{q}(s,a,w) \approx q_\pi(s,a)$$

所有与机器学习相关的算法都可以作为强化学习的近似函数，如线性组合、决策树、最近邻法、深度神经网络等。

2. 策略梯度法

策略梯度法将策略的学习从概率集合P(a|s)变成策略函数$\pi(a|s)$，并通过求解策略目标函数的极大值，得到最优策略π^*。

- 蒙特卡洛策略梯度法
- 演员-评论家 算法

3. 学习与规划

基于模型的强化学习方法通过环境模型进行规划以找到最优策略。

学习阶段

- 智能体从真实的经验轨迹数据中进行模型学习，获得模拟的环境模型，进而获得环境的准确描述。

规划阶段

- 基于获得的环境模型，智能体与模拟环境进行交互并获得大量的模拟经验轨迹数据。
- 在模拟经验轨迹集之上，智能体采用免模型的强化学习求解法对价值函数或策略函数进行更新。

规划：指智能体与模拟环境进行交互，而非与实际环境进行交互，并在与模拟环境的交互过程中进行经验轨迹数据的采集，采集到的经验轨迹数据成为模拟经验轨迹。

DQN

- Double DQN
  - 对动作的选择和动作状态值估计进行解耦，使用两个Q网络分别进行学习。
- Prioritized DQN
  - 优先级回放
- Dueling DQN
  - 动作状态值函数Q(s,a)分解成状态值函数V(s)和动作优势函数A(s,a)