Reinforcement Learning

Yu Xueqing

2022 年 8 月 4 日

1 Introduction of RL

不知道最佳输出 (label) 应该是什么，e.g. 下棋
actor 和 environment 互动，采取 action，environment 给予 reward, 目标是使 reward 总和最大

- function: policy network
  输入:observation 输出：action, 类似 classification
- define "loss": maximize total reward
- optimization 训练难点：输出有随机性; env 和 reward 是黑箱

2 Policy Gradient

希望 actor 在 observation s 时做出 action $\hat{a}$，计算 cross-entropy 使得 loss 最小，即 a 和 $\hat{a}$ 越接近越好；同理不希望做出的 action 取-loss

收集 traning data: $\{s_i, \hat{a}_i\}$, $L = \sum e_i (or - e_i)$
A_i 表示希望/不希望做出某个 action 的程度，$L = \sum A_n e_n$
A_i 的选取：

- 1. 即刻 reward-短视
- 2. t 时刻之后的累计 reward $G_t = \sum_{n=t}^{N} r_n$
- 3. 考虑时间越远影响越小，加入衰减系数 γ, $G'_t = \sum_{n=t}^{N} \gamma^{n-t} r_n$

on-policy: actor to train 和 actor for interact 相同, 每次更新后重新搜集数据
off-policy: actor to train 和 actor for interact 不同，不用每次更新后重新搜集
method：PPO(proximal policy optimization)

3 Actor-Critic

value function $V^\theta(s)$: actor θ 在观察到 s 后期望得到的累计 reward
计算 $V^\theta(s)$ 的两种方式：

- Monte-Carlo(MC) based approach: 重复模拟，取平均值
- Temporal-difference(TD) approach: 在 s_t 时执行 a_t,reward r_t, 环境变为 s_{t+1}
  $V^\theta(s_t) = \gamma V^\theta(s_{t+1}) + r_t$, 训练 V 使其满足递推关系

$A_t = r_t + V^\theta(s_{t+1}) - V^\theta(s_t)$,
$r_t + V^\theta(s_{t+1})$ 即在 s_t 时执行 a_t 后的期望 reward 之和，$V^\theta(s_t)$ 是 s_t 时执行不同 action 后得到 reward 之和的期望，于是 A_t 表示执行某 action 的好坏程度

4 Reward Shaping

现实中 reward 往往很稀疏，不容易训练-> 人为添加 reward

5 Inverse RL

demonstration of the expert, principle:the teacher is always the best
reward function 是学到的，使得 expert 的得分总是比 actor 高，然后用这个 reward function 训练 actor