1 强化学习与多臂老虎机

1.1 强化学习引言

人类是通过与环境互动来学习的, 而**强化学习 (Reinforcement Learning, RL)**，就是通过计算来实现从互动中学习. 它是一种目标导向的互动学习, *互动*性是它与监督学习的重要区别, 而有无明确的目标则是它与无监督学习的主要区别. 强化学习需要学习的主要内容, 就是怎样做才能使回报最大化. 它有两个显著特征: *试错*与*延迟回报*.

强化学习的主体通常称作**智能体 (Agent)**, 也通常称作代理. 我们将在下一节介绍智能体的数学框架, 即感知 (Sensation, S), 行动 (Action, A) 和目标 (Goal ,G).

1.2 多臂老虎机: 探索与利用的平衡

1.2.1 多臂老虎机问题的定义

有关多臂老虎机的现实背景不再赘述. 这里用数学语言描述多臂老虎机问题:

智能体在每一步都需要从 K 个可能的行动 $\{a_k\}_{k=1}^K$ 中选择一个行动, 当选择行动 a_k 后将得到回报 R_k, R_k 的静态分布 $p(R_k)$ 仅取决于行动 a_k. 智能体应当怎样逐步选择 N 个行动 $A_t(t=1,\cdots,N)$ 使得获得的总回报 $\sum_{t=1}^{N} R_t$ 最大?

定义行动 a 的价值函数为

$$q_*(a) := \mathbb{E}[R_t|A_t = a]$$

如果 $q_*(a)$ 已知, 那么上述问题就很简单, 只需在每步选择使得 $q_*(a)$ 最大的 a 即可. 然而, 智能体实际上不能获知 $q_*(a)$, 只能在第 t 步做选择时根据之前的结果计算 $q_*(a)$ 的估计值 $Q_t(a)$, 并据此选择合适的行动.

如果选择 $Q_t(a)$ 更大的行动, 那么主要是利用当前对行动价值函数的了解进行回报最大化, 但是容易陷入局部最优中; 如果选择其它行动, 那么更有利于精确地估计行动价值函数 $Q_t(a)$, 但是容易损失回报. 这两者经常是矛盾的, 倾向一者就会远离另一者, 这就是多臂老虎机中探索-利用权衡的难题.

针对上述问题, 人们提出了许多算法.

1.2.2 多臂老虎机的若干算法

多臂老虎机的算法大都基于对行动价值函数的估计, 即估计行动 a 的价值 $Q_t(a)$ 并在此基础上采取行动. 行动价值函数的一种简单的估计方法是统计此前所有采取 a 行动后得到回报的均值, 即

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i I(A_i = a)}{\sum_{i=1}^{t-1} I(A_i = a)}$$

其中 I 是示性函数. 当然, 采用贝叶斯方法通过后验概率估计也是可行的.

贪心算法 在智能体采取行动时可以采用**贪心算法 (Greedy Algorithm)**, 即选取使得 $Q_t(a)$ 取最大值的行动, 即

$$A_t = \arg\max_a Q_t(a)$$

为了防止贪心算法陷入局部最优, 可以强制加入探索的成分, 典型的是 **ε-贪心算法**, 即有小概率 ε 随机选择所有可能的行动, 大概率 $1-\varepsilon$ 遵循贪心算法选择回报期望最高的行动.

然而, 在 K 个选择的 ε-贪心算法中, 在足够长的时间后, 尽管 $Q_t(a)$ 已经足够接近 $q_*(a)$, 算法仍然有 $\dfrac{K-1}{K}\varepsilon$ 的概率选择非最优的步骤. 这启示我们也许可以在不同的阶段采取不同的策略.

UCB 算法 多臂老虎机问题的著名算法, **置信区间上界算法 (Upper Confidence Bound Algorithm, UCB)** 在一定程度上解决了上述问题, 即随着行动数目 t 的增大而逐步减小探索的概率.

UCB 算法分析了 $Q_t(a)$ 估计 $q_*(a)$ 的误差. 通常, 对随机变量 x 进行 n 次测量, 结果为 $X_1, \cdots, X_n$, 那么通常把结果写成

$$x = \bar{X} \pm \frac{\sigma}{\sqrt{n}}, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

根据这一分析结果, 我们认为 $q_*(a)$ 有很大概率落在以下区间 (即置信区间) 内:

$$\left[Q_t(a) - \frac{\sigma_t(a)}{\sqrt{N_t(a)}}, Q_t(a) + \frac{\sigma_t(a)}{\sqrt{N_t(a)}}\right]$$

其中 $N_t(a) = \sum_{i=1}^{t-1} I(A_i = a)$ 是行动 a 被采纳的次数. UCB 算法采用了一种面对不确定性时的乐观想法, 将上述置信区间的上界作为 $q_*(a)$ 的估计, 给出如下的行动结果:

$$A_t = \arg\max_a \left[Q_t(a) + c\sigma_t(a)\sqrt{\frac{\ln t}{N_t(a)}}\right]$$

引入因子 $\ln t$ 和常数 c 可以更好地控制误差. 通过选择具有最高置信上界的行动 a, 就会倾向于选择那些既有较高期望奖励又较少被探索的行动, 从而获得更好的探索-利用平衡效果.