学号：20232241391 姓名：刘洛松

报告五：强化学习

# 强化学习概述

近年来，深度学习(Deep Learning，DL)作为机器学习领域一个重要的研究热点，已经在图像分析、语音识别、自然语言处理、视频分类等领域取得了令人瞩目的成功[1]。

# 强化学习的构成和原理

受到生物学习规律的启发，强化学习以试错机制与环境进行交互，通过最大化累积奖赏的方式来学习和优化，最终达到最优策略.在强化学习中，定义决策者或学习者为“学习机”，将学习机之外的事物定义为“环境”，系统与环境相融。学习机和环境之间的交互过程可以由三个要素来描述，分别是:状态s、动作a、奖励r。学习机根据初始状态S0，执行动作a0并与环境进行交互,得到奖励r1并获得更新的状态s1。在时间步t,根据当前状态St和奖励rt，学习机提供当前动作at。接着，系统状态由st转变为st+1,与环境交互反馈奖励rt+1[2]。

强化学习的要素简单来说就是：感知，决策，奖励。

# 强化学习与一般机器学习的区别

一般机器学习可以被称为有监督学习，它的输入输出都是固定的，找出输入输出两者间的关系便是一般机器学习算法的目标。

## 与环境的交互

有监督学习的任务建立在从给定的数据分布中采样得到的训练数据集上，通过优化在训练数据集中设定的目标函数（如最小化预测误差）来找到模型的最优参数。这里，训练数据集背后的数据分布是完全不变的。

在强化学习中，数据是在智能体与环境交互的过程中得到的。如果智能体不采取某个决策动作，那么该动作对应的数据就永远无法被观测到，所以当前智能体的训练数据来自之前智能体的决策结果[3]。

## 应用场景与解决的问题

强化学习所能解决的问题是智能决策问题。更确切地说是序贯决策问题。什么是序贯决策问题呢？就是需要连续不断地做出决策,才能实现最终目标的问题[4]。

# 强化学习的策略

强化学习有许多决策方法，如策略梯度方法，深度强化学习方法，演化强化学习方法等。

策略梯度算法广泛应用在连续空间中的RL问题中，其主要思想是将策略Π参数化表示为Πθ，并计算出关于动作的策略梯度，然后沿着梯度的方向，不断地调整动作，逐渐得到最优策略。常见的策略梯度算法有REINFORCE算法和行动者-评论家 (Actor—Critic，AC)算法等。

策略梯度中，策略分为随机性策略和确定性策略。随机性策略对应的含义为，当前状态为S时，动作a满足参数为θ的某个概率分布，因此相同状态也会对应不同的动作。而确定性策略和随机性策略不同的是，对于一个确定性的策略μθ来说，每个状态对应唯一的动作[5]。

# 参考文献：

1. 刘全,翟建伟,章宗长,等. 深度强化学习综述[J]. 计算机学报,2018,41(1):1-27. DOI:10.11897/SP.J.1016.2018.00001.
2. 孙长银,穆朝絮. 多智能体深度强化学习的若干关键科学问题[J]. 自动化学报,2020,46(7):1301-1312. DOI:10.16383/j.aas.c200159.
3. 张伟楠，沈键，俞勇著. 动手学强化学习. 北京: 人民邮电出版社, 2022. Web.
4. 郭宪，方勇纯编著. 深入浅出强化学习 原理入门. 北京: 电子工业出版社, 2018. 博文视点AI系列. Web.
5. 刘建伟, 高峰, and 罗雄麟. "基于值函数和策略梯度的深度强化学习综述." Jísuánjī Xuébào 42.6 (2019): 1406-438. Web.