强化学习

维基百科，自由的百科全书

强化学习是机器学习中的一个领域，强调如何基于环境而行动，以取得最大化的预期利益。其灵感来源于心理学中的行为主义理论，即有机体如何在环境给予的奖励或惩罚的刺激下，逐步形成对刺激的预期，产生能获得最大利益的习惯性行为。这个方法具有普适性，因此在其他许多领域都有研究，例如博弈论、控制论、运筹学、信息论、仿真优化、多主体系统学习、群体智能、统计学以及遗传算法。在运筹学和控制理论研究的语境下，强化学习被称作“近似动态规划”（approximate dynamic programming，ADP）。在最优控制理论中也有研究这个问题，虽然大部分的研究是关于最优解的存在和特性，并非是学习或者近似方面。在经济学和博弈论中，强化学习被用来解释在有限理性的条件下如何出现平衡。

在机器学习问题中，环境通常被规范为马可夫决策过程（MDP），所以许多强化学习算法在这种情况下使用动态规划技巧。传统的技术和强化学习算法的主要区别是，后者不需要关于MDP的知识，而且针对无法找到确切方法的大规模MDP。

强化学习和标准的监督式学习之间的区别在于，它并不需要出现正确的输入/输出对，也不需要精确校正次优化的行为。强化学习更加专注于在线规划，需要在探索（在未知的领域）和遵从（现有知识）之间找到平衡。强化学习中的“探索-遵从”的交换，在多臂老虎机问题和有限MDP中研究得最多。

导论

基本的强化学习模型包括：

1. 环境状态的集合S;
2. 动作的集合A;
3. 在状态之间转换的规则；
4. 规定转换后“即时奖励”的规则；
5. 描述主体能够观察到什么的规则。

规则通常是随机的。主体通常可以观察即时奖励和最后一次转换。在许多模型中，主体被假设为可以观察现有的环境状态，这种情况称为“完全可观测”（*full observability*），反之则称为“部分可观测”（*partial observability*）。有时，主体被允许的动作是有限的（例如，你使用的钱不能多于你所拥有的）。

强化学习的主体与环境基于离散的时间步长相作用。在每一个时间t，主体接收到一个观测o_t，通常其中包含奖励r_t。然后，它从允许的集合中选择一个动作a_t，然后送出到环境中去。环境则变化到一个新的状态s_{t+1}，然后决定了和这个变化(s_t, a_t, s_{t+1})相关联的奖励r_{t+1}。强化学习主体的目标，是得到尽可能多的奖励。主体选择的动作是其历史的函数，它也可以选择随机的动作。

将这个主体的表现和自始自终以最优方式行动的主体相比较，它们之间的行动差异产生了“悔过”的概念。如果要接近最优的方案来行动，主体必须根据它的长时间行动序列进行推理：例如，要最大化我的未来收入，我最好现在去上学，虽然这样行动的即时货币奖励为负值。

因此，强化学习对于包含长期反馈的问题比短期反馈的表现更好。它在许多问题上得到应用，包括机器人控制、电梯调度、电信通讯、双陆棋和西洋跳棋。[1]

强化学习的强大能来源于两个方面：使用样本来优化行为，使用函数近似来描述复杂的环境。它们使得强化学习可以使用在以下的复杂环境中：

- 模型的环境已知，且解析解不存在；
- 仅仅给出环境的模拟模型（模拟优化方法的问题）[2]

- 从环境中获取信息的唯一办法是和它互动。前两个问题可以被考虑为规划问题，而最后一个问题可以被认为是genuine learning问题。使用强化学习的方法，这两种规划问题都可以被转化为机器学习问题。

注释

1. Sutton1998|Sutton and Barto 1998 Chapter 11
2. Gosavi, Abhijit. Simulation-based Optimization: Parametric Optimization Techniques and Reinforcement. Springer. 2003. ISBN 1-4020-7454-9.