强化学习基础篇（三）动态规划之基础介绍

强化学习从动物学习行为中的试错方式和优化控制理论两个领域独立发展，最终经贝尔曼方程抽象为马尔可夫决策过程，从而奠定了强化学习的数学理论基础。在贝尔曼之后，经过了众多科学家的深入研究和补充，形成了相对完备的强化学习体系。
正由于涉及的数学理论众多、公式繁杂，强化学习常常被看作机器学习领域中较为深奥的范式之一。可以对强化学习有了较为全面而深入的认识，其事实上所涵盖的元素就清晰了：基于马尔可夫决策过程的4个重要元素（状态s、动作a、奖励和状态动作转换概率P），以及策略π、状态值函数$v(s)$和动作状态值函数$q(s,a)$。
而强化学习任务的求解实际上就是寻找最优策略π^*，基于贝尔曼方程可以有3种求解方法：动态规划法（Dynamic Programming）、蒙特卡洛法(Monte Carlo Method)和时间差分法（Temporal Difference）。本文将会着重介绍如何利用动态规划法来完成强化学习中基于模型的任务，并通过价值函数或者策略函数获得最优策略。

1、什么是动态规划

动态规划法(Dynamic Programming)将原问题分解为子问题，并通过对子问题的求解而解决较难的原问题。这与基于马尔可夫决策过程的强化学习任务具有天然的关联性。

> 基本定义：
>
> 动态规划（英语：Dynamic programming，简称DP）是一种在数学、管理科学、计算机科学、经济学和生物信息学中使用的，通过把原问题分解为相对简单的子问题的方式求解复杂问题的方法。
>
> 动态规划常常适用于有重叠子问题和最优子结构性质的问题，动态规划方法所耗时间往往远少于朴素解法。
>
> 动态规划背后的基本思想非常简单。大致上，若要解一个给定问题，我们需要解其不同部分（即子问题），再根据子问题的解以得出原问题的解。
>
> 通常许多子问题非常相似，为此动态规划法试图仅仅解决每个子问题一次，从而减少计算量：一旦某个给定子问题的解已经算出，则将其记忆化存储，以便下次需要同一个子问题解之时直接查表。这种做法在重复子问题的数目关于输入的规模呈指数增长时特别有用。

换言之，动态规划通过把复杂问题划分为子问题，并逐个求解子问题，最后把子问题的解进行结合，进而解决较难的原问题。其中，“动态”指问题由一系列的状态组成，而且能随时间变化而逐步发生改变；“规划”(Programming)即优化每一个子问题。

2、动态规划与贝尔曼方程

根据前面的定义可知，动态规划是一个非常通用的方法。当问题具有下列特性时，通常可以考虑使用动态规划来求解：

- 第一个特性是一个复杂问题的最优解由数个小问题的最优解构成，可以通过寻找子问题的最优解来得到复杂问题的最优解；
- 子问题在复杂问题内重复出现，使得子问题的解可以被存储起来重复利用。

马尔科夫决定过程（MDP）具有上述两个属性：贝尔曼方程（Bellman equation)把问题递归为求解子问题，价值函数就相当于存储了一些子问题的解，可以复用。因此可以使用动态规划来求解MDP。

在求解贝尔曼方程中，首先使用的就是动态规划法，主要原因在于：

- 贝尔曼等人在研究多阶段决策过程优化问题时，提出了使用动态规划来求解多阶段决策过程；

- 由马尔可夫决策过程的马尔可夫特性（即某一时刻的子问题仅取决于上一时刻子问题的状态和动作）所决定的。贝尔曼方程可以递归地切分子问题，因此非常适合采用动态规划法来求解贝尔曼方程。

强化学习的核心思想是使用值函数$v(s)$或者动作值函数$q(s,a)$，找到更优的策略给智能体进行决策使用。由上一篇文章的内容可知，当找到最优的状态值函数$v(s)$或者最优的动作值函数$q(s,a)$，就可以找到最优策略π，公式如下:

$$\begin{aligned} v^*(s) &= \max_{a\in A}[R_{t+1}+\gamma v_\pi(S_{t+1})|S_t=s,A_t=a] \\ &= \max_{a\in A}\sum_{a\in A}p(s',r|s,a)[r+\gamma v^*(s')] \end{aligned}$$

$$\begin{aligned} q^*(S,A) &= \mathbb{E}(R_{t+1}+\gamma\max_{a'}q^*(S_{t+1},a')|S_t=s,A_t=a) \\ &= \sum_{s',a}p(s',r|s,a)[r+\gamma\max_{a'}q^*(s',a')] \end{aligned}$$

其中状态$s\in S$、动作$a\in A$、新的状态$s'\in S^+$。上面两式中，最优价值为环境中的每一个状态s和动作a对应的动作转换概率$p(s',r|s,a)$乘以未来折扣奖励中最大的价值$[r+\gamma\max_{a'}value^*(s',a')]$。其中$value^*(s',a')$为价值函数，可以为$v^*(s')$或者为$q^*(s',a')$。

动态规划法主要是将上式中的贝尔曼方程转换为赋值操作，通过更新价值来模拟价值更新函数。

需要注意的是，使用动态规划法求解强化学习时，由于涉及对强化学习中的策略进行评估与改进，于是引入了评估策略π优劣程度的策略评估方法，并通过策略改进和策略迭代算法，寻找最优v^*。除此之外，还可以通过值迭代算法代替策略迭代来求解最优。

3、动态规划解决规划(Planning)问题

我们用动态规划算法来求解一类称为“规划”的问题。“规划”指的是在了解整个MDP的基础上求解最优策略，也就是清楚模型结构的基础上：包括状态行为空间、转换矩阵、奖励等。这类问题不是典型的强化学习问题，我们可以用规划来进行预测和控制。
对于预测问题，具体的数学描述是这样：

> 预测（prediction）：
>
> 输入： MDP $<S,A,P,R,\gamma>$以及策略π。或者MRP $<S,P^\pi,R^\pi,\gamma>$。
>
> 输出：值函数v_π

对于控制问题，具体的数学描述是这样：

> 控制（Control）：
>
> 输入： MDP $<S,A,P,R,\gamma>$
>
> 输出：最优值函数v_*以及最优策略π_*

4、动态规划的其他应用

动态规划在很多领域都有着广泛的应用，比如：

- 用于比较两个文件的Unix diff。
- 隐马尔可夫模型的维特比算法（Viterbi algorithm）。
- 评价样条曲线的德布尔算法（De Boor's algorithm）。
- 用于基因序列比对的史密斯-沃特曼算法 (Smith–Waterman algorithm)。
- 网络中求最短路径路由的Bellman-Ford算法。
- 用于解析 context-free 语法的Cocke-Kasami。