马尔科夫决策过程

原创 冯晓妍 LOA算法学习笔记 2021-02-24 17:23

这个学期我们在卜老师的课堂上深入学习了动态规划的思想，知道了许多看似复杂难解的问题都可以用动态规划巧妙地找到答案。有意思的是，在数学学院一门运筹学的课程中，老师将动态规划作为马尔科夫决策过程（Markov Decision Process, MDP）的引子，为我们介绍了有限时间步的离散状态的马尔科夫决策过程。

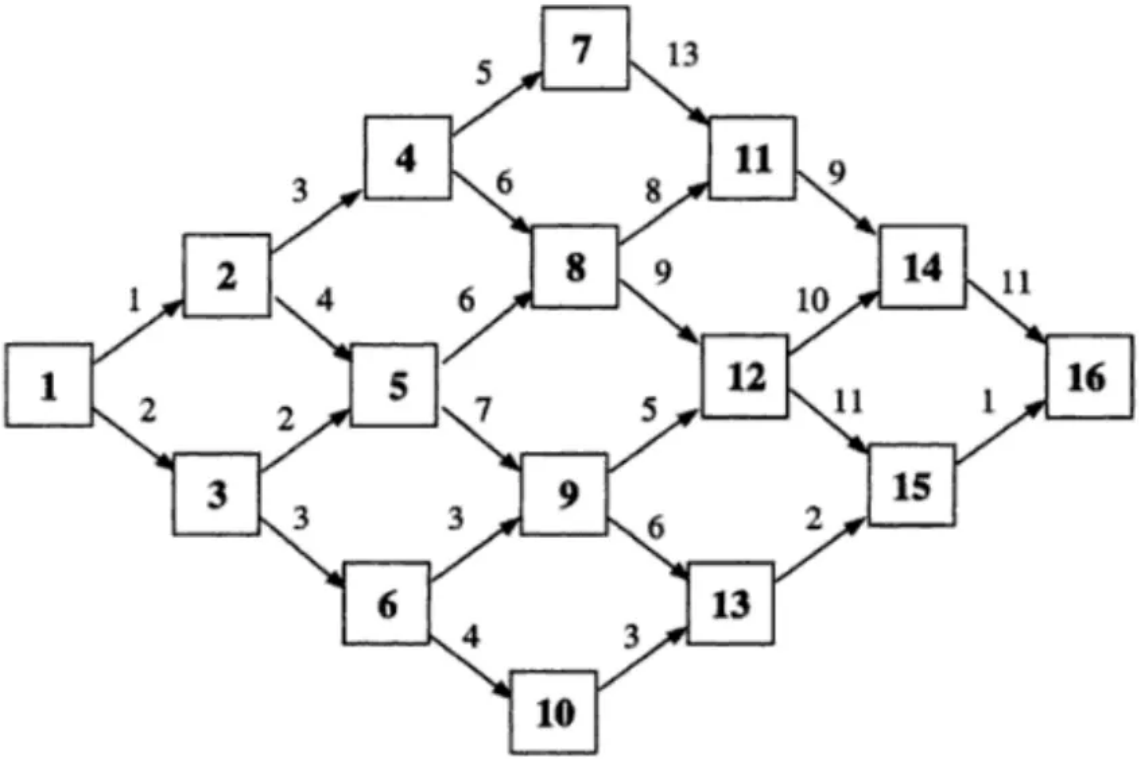
后来我在与其他同学的讨论中，发现马尔科夫决策过程也是强化学习的一个基本模型。强化学习中的MDP基于一组交互对象，即智能体和环境进行构建，具有状态、动作、策略和奖励等要素。在MDP的模拟中，智能体会感知当前的系统状态，按策略对环境实施动作，从而改变环境的状态并得到奖励，奖励随时间的积累被称为回报。

本文从动态规划出发，希望能够与大家分享有限时间步的离散状态的马尔科夫决策过程的思想和简单的例子。

01 离散状态的马尔科夫决策过程

（一）含机会节点和决策节点的最短路径问题

首先回顾最短路径问题，以下图为例，



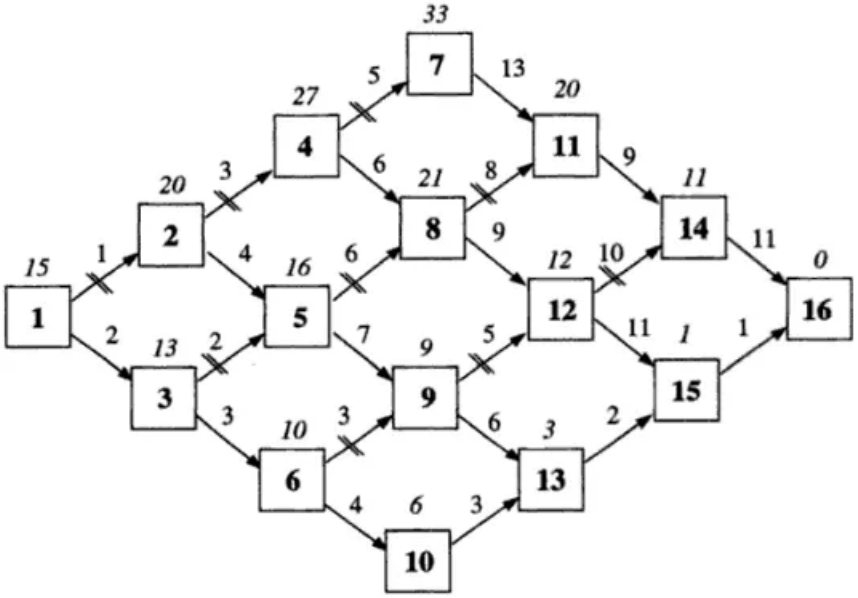
定义 f(i)表示从节点i到节点16的最短路径长度，i=1,2,...,16；d_{ij}表示从节点i到节点j的距离。最短路问题的递归求解过程为：

$$for\ i = 1,2,\dots,15, do:$$

$$f(i) = \min_{j}\{d_{ij} + f(j)\}$$

$$\text{边界条件}\ f(16) := 0$$

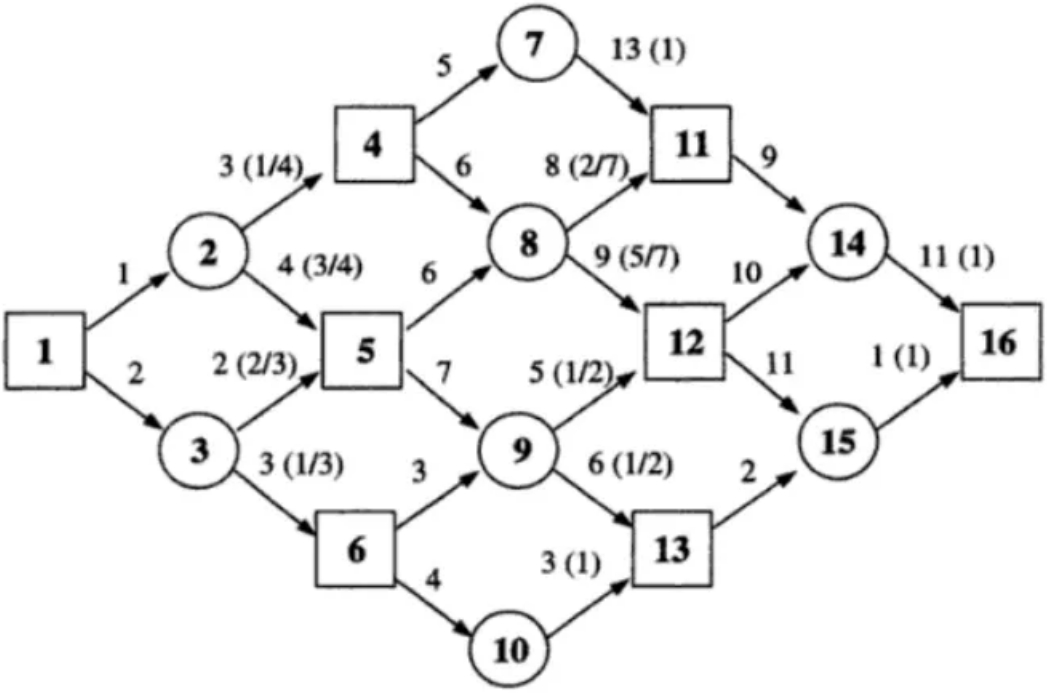
在最短路径问题中，动态规划的思想体现在计算f(i)的过程中用到了两个重要信息：（1）与它相邻的所有节点的到终点的最短路径距离，（2）它自身和与它相邻的节点的距离。这样求得的后向求得的每一个f(i)都是最优子问题的答案，f(16)是考察了全局信息的最优解。



如图所示，最短路径是1-3-6-10-13-15-16，最短路径长度是15。

在最短路径问题中，两个点之间的有向边是**确定的**，也可以理解为边总是存在的。

下面在最短路径问题中引入机会节点和决策节点。**决策节点**（decision nodes），由正方形节点表示；**机会节点**（chance nodes），由圆形节点表示。目标是最小化从节点1到节点16的**期望路径长度**（expected distance）。



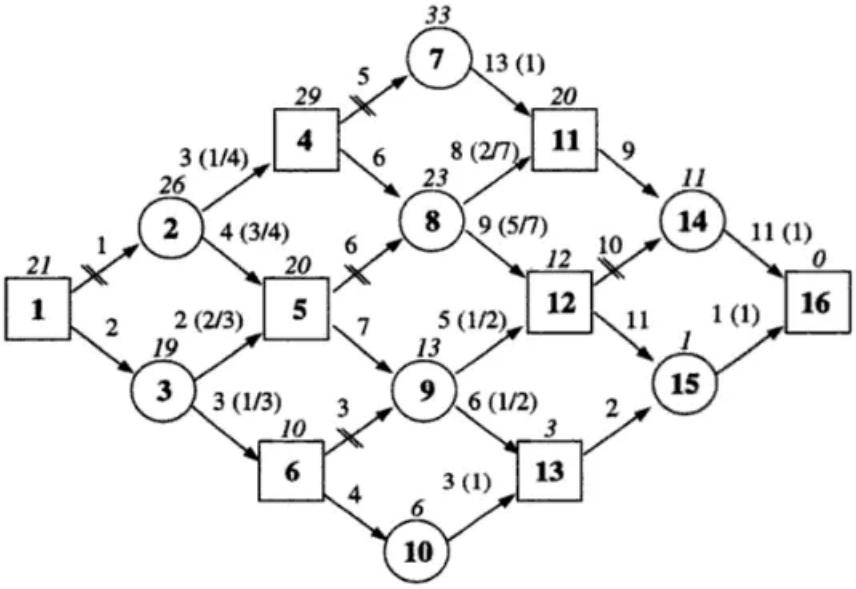
同样采用后向的分析方法，对于节点10，随机最短路径问题的f(10)与一般最短路径问题的f(10)的计算结果相同。但是对于节点9，我们不能直接选择下一个状态，而是以1/2的概率选择节点12，以1/2的概率选择节点13，因此

$$f(9)=\frac{1}{2}\left[d_{9,12}+f(12)\right]+\frac{1}{2}\left[d_{9,13}+f(13)\right]=\frac{1}{2}[5+12]+\frac{1}{2}[6+3]=13$$

后向递归：

$$\text{决策节点：}\quad f(i)=\min_j\{d_{ij}+f(j)\}$$

$$\text{机会节点：}\quad f(i)=\sum_j\left[d_{ij}+f(j)\right]p(j|i)$$



如图所示，最小期望路径长度是21，不同于上一个最短路径问题的答案。

02 马尔可夫决策的应用-顾客促销优惠问题

（一）问题定义和分析：

假设市场上有一家互联网零售公司e-Rite-Way，他们拥有所有客户的电子邮箱地址。

他们将客户分为两个状态：

State 1 (inactive state): 该客户在过去一个月都没有任何购买记录

State2（active state）：该客户在过去一个月曾经有过购买记录

公司经理正在考虑定期向客户发送电子贺卡，并附上力度较轻或者较大的促销优惠以激励客户的购买行为。对于每位客户，公司可以采取以下三种行为：

Decision	Description	$ Direct Cost
a		$c(a)$
0	Nothing active carried out	0
1	Gift sent and minor price promotion offered	0.50
2	Gift sent and major price promotion offered	0.50

公司经理希望客户得到促销优惠后能够变为活跃用户。经过调查分析，他们发现不同状态的客户对于不同的促销力度有如下反应：

Customer State	Decision	Probability of No Purchase	Probability of Purchase
s	a	p_{s1}^a	p_{s2}^a
1	0	0.9899	0.0101
1	1	0.9293	0.0707
1	2	0.8586	0.1414
2	0	0.8081	0.1919
2	1	0.7273	0.2727
2	2	0.5051	0.4949

对于这种经济问题，回报应该考虑货币的时间价值，因此令α=0.99为一个月的贴现因子。

令 r_{sj}^a 为随之而来的回报， $r_{s2}^0 = 8, r_{s2}^1 = 7, r_{s2}^2 = 3$

每类客户做出每种决定的即时回报的现值为

$$r(s,a) := -c(a) + \alpha \sum_j p_{sj}^a r_{sj}^a$$

Customer State	Decision	Direct Cost	Expected Present Value of Purchase	Immediate Return
s	a			$r(s,a)$
1	0	0	$0.99[0.9899 \times 0 + 0.0101 \times 8]$	0.08
1	1	-0.5	$0.99[0.9293 \times 0 + 0.0707 \times 7]$	-0.01
1	2	-0.5	$0.99[0.8586 \times 0 + 0.1414 \times 3]$	-0.08
2	0	0	$0.99[0.8081 \times 0 + 0.1919 \times 8]$	1.52
2	1	-0.5	$0.99[0.7273 \times 0 + 0.2727 \times 7]$	1.39
2	2	-0.5	$0.99[0.5051 \times 0 + 0.4949 \times 3]$	0.97

假设客户在期末没有剩余价值，在一期问题中使预期收益最大化的政策是使即时收益最大化的政策。**不应向任何一种客户类型提供促销。**但这就是问题的答案吗？那为什么现实生活中，会存在如此多有针对性的促销优惠行为呢？

其实，由于时间跨度短，这种不向任何一种客户类型提供促销的策略忽略了**将不活跃的客户转化为活跃的客户的价值**，他们更有可能在未来再次购买。

可以试着考虑**两个月的时间范围。**

例如，在考察时间范围的开始，考虑向不活动的客户发送一个主要的促销：在状态1时选择动作2。第一个月的即时回报为-0.08。在下个月初，您将像在一个周期问题中一样采取最佳行动。所以后续收益的最优期望现值为

$$0.99[0.8586\times0.08+0.1414\times1.52]=0.2808$$

所以总的期望现值是-0.08+0.2808=0.2008

s	Return form Decision 0	Return form Decision 1	Return form Decision 2	Optimal Decision	Optimal Value
1	0.1736	0.1700	0.2008	2	0.2008
2	1.8728	1.8580	1.7548	0	1.8728

因此，经理应该用两个周期（月）的时间范围来使用的策略是：在开始的时候，给不活跃的客户发送一个重大的促销，而对活跃的客户则什么也没有。一个月后，不发送任何客户。分别分析三期和四期问题。设表示从状态s开始的n个周期内返回的最大期望现值。我们刚才计算了每个状态s的，我们刚刚计算 使用的以下递归，称为最优性方程：

$$f_n(s)=\max_a\{r(s,a)+\alpha\sum_j p_{sj}^a f_{n-1}(j)\}$$

最优性方程表明，长度为n的时间步上的最优回报可以计算如下。选择初始状态s，并考虑任意动作a。然后获取预期的即时回报，并添加与系统在此期间结束时可能移动到的每个状态相关的最优回报的期望现值。

分别将考虑时间范围延伸到3和4个时期，计算结果如下。大家可以思考一下，在这种情况下，采取怎样的策略才是最优的呢？

Table: Analysis of Three-Period Problem

	Conditional Exp. Return	Conditional Exp. Return	Conditional Exp. Return	Optimal Decision	Optimal Value
s	$a=0$	$a=1$	$a=2$		
1	0.2955	0.3058	0.3529	2	0.3529
2	2.036	2.040	1.988	1	2.040

Table: Analysis of Four-Period Problem

	Conditional Exp. Return	Conditional Exp. Return	Conditional Exp. Return	Optimal Decision	Optimal Value
s	$a=0$	$a=1$	$a=2$		
1	0.4462	0.4575	0.5056	2	0.5056
2	2.1899	2.1949	2.1462	1	2.1949

（三）马尔科夫决策过程（Markov decision processes, MDP）求解问题

带机会节点和决策节点的最短路径问题和给予哪种优惠促销问题都是有限时间步马尔科夫决策过程的简单实例。有限步马尔科夫决策过程可以形式化如下：

N：周期数N，称为planning horizon或time horizon。

S：状态空间，在一个周期开始，过程（Process）都处于状态空间S的某个状态s，如果S是离散的，那么这个过程（Process）称为链（Chain）。

s：状态，s∈S。

A (s)：动作集，表示在状态s时可以采取的动作。

a：过程（Process）处在某个时期的某个状态时可能采取的动作。

p_{sj}^{a} ：在某一时期，当前状态是s，采取动作a，转移到次态j的概率。之所以说这种过程演化是马尔科夫的，是由于过程转移到下一个状态的概率只和当前状态s和动作a有关。

r(s,a)：在状态s采取动作a的即时回报（immediate return）

v_T ：终值函数，当planning horizon的终期处于状态s的价值。

δ：决策规则，决策规则是一个指明状态空间的状态可以采取哪些动作的函数，对于每一个s∈S，有δ(s)∈A(s)。

Δ：可行决策 (admissible decision rules) 的集合。形式上，我们有δ∈Δ。

π：策略（strategy），$\pi = (\delta_1, \delta_2, \ldots, \delta_n)$

Π：策略空间 (strategy space)，Π=Δ×Δ×…×Δ

$v_t(\pi, s)$ ：从时期t以状态s开始，在t,t+1,… ,N 的时间序列中采取策略 π 产生的期望现值。期望现值是价值贴现到当期的值。当我们不需要考虑时间价值时，设贴现率为0即可。

$$v_t(\pi, s) = E[\sum_{i=t}^{N} \alpha^{i-t} r\big(s_i, \delta_i(s_i)\big) + \alpha^{N-t+1} v_T(s_{N+1})$$

$v_t(\pi)$ ：期望现值向量，向量的各个分量是当t时期处于状态空间的各个状态，采取策略π 得到的期望现值。

v(π)：当t=1时，可以省略下标t。

π^* ：是时期t下状态s的最优策略，如果

$$v_t(\pi^*, s) \geq v_t(\pi, s) \quad for\ every\ \pi \in \Pi$$

是时期t的最优策略，如果

$$v_t(\pi^*) \geq v_t(\pi) \quad for\ every\ s \in S$$

f_t ：时期t的最优值函数（optimal value function）

$f_t(s)$ ：在时期t以状态s开始的过程的最优值

$$f_t(s) := \sup_{\pi \in \Pi} v_t(\pi, s)$$

有了上述符号定义，可以得到马尔科夫决策过程中最重要的结论——最优性方程

最优性方程（optimality equations）：

$$f_t(s) = \sup_{a \in A(s)} \{r(s,a) + \alpha E[f_{t+1}(s_{t+1})]\}$$

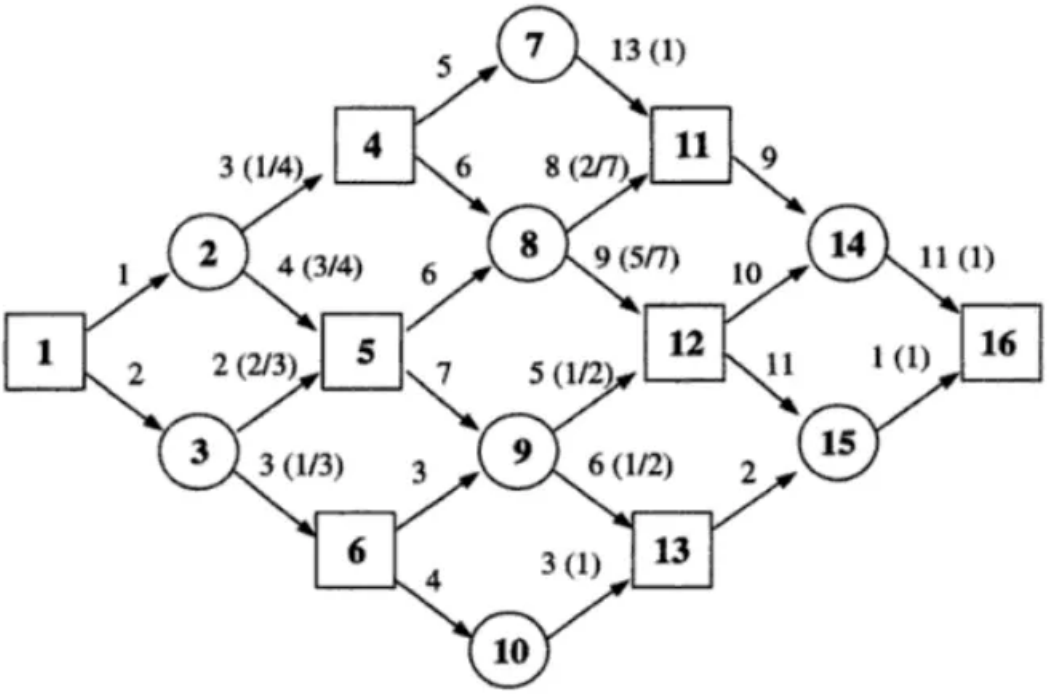
$$= \sup_{a \in A(s)} \left\{ r(s,a) + \alpha \sum_{j \in S} p_{sj}^{a} f_{t+1}(j) \right\}$$

$$其中 f_{N+1}(s) = v_T(s)，\ 对于所有 s$$

换句话说，最优性方程表明，在任意初始状态的最优值可以表示为该状态下所有可行决策的即时回报的最优值，而这个最优值考虑了当前状态和采取动作所关联的所有状态的最优值。因此，任何时期任何状态的最优值都可以通过后向递归

得到。

带机会节点和决策节点的最短路径问题可以写成如下马尔科夫决策过程：



Horizon: T=6
State：{1,2,3,… ,16}
Action: A(1)={2,3},A(2)={4,5},A(3)={5,6},A(4)={7,8},A(5)={8,9},A(6)={9,10},A(7)={11},A(8)={11,12},A(9)={12,13},A(10)={13},A(11)={14},A(12)={14,15},A(13)={15},A(14)={16},A(15)={16}
Return: r(1,2)=1,r(1,3)=2,r(2,4)={3},r(2,5)=4,… ,r(14,16)=11,r(15,16)=1
Probability:

$$p_{1,2}^{2}=1,\ p_{1,3}^{3}=1,\ p_{2,4}^{4}=\frac{1}{4},p_{2,5}^{5}=\frac{3}{4},\ldots,p_{14,16}^{16}=1,p_{15,16}^{16}=1$$

Terminal value is zero。

对于马尔科夫决策过程的介绍就到这里，本人对于这些知识的理解还有许多不透彻的地方，可能存在许多错漏和不恰当的解释，欢迎大家批评指正。