机器学习：随机过程(Stochastic Process)与强化学习(Reinforcement Learning)

---

1. 复习

PCA 的核心问题：

$$w_1 = \arg\max_{||w_1||=1} w_1^T X X^T w_1$$

根据瑞利商理论，该最优化问题可转为求 XX^T 的最大特征值，可转换为对 X 作奇异值分解，即对 XX^T 做特征值分解。

2. 导言

监督学习可以理解为一个开环控制系统：

$$X \to f^* \to y$$

强化学习可以理解为一个带反馈的闭环控制系统(最优控制，贝尔曼原理)。

应用：$Alpha\ Go$ 围棋，机器人，自动驾驶。

3. 最优控制

介绍最优控制的基本思想：

$$J(u_0, \cdots, u_\infty) = \min \sum_{i=0}^{\infty} x_i^T P x_i + u_i^T Q u_i$$

即根据不同控制器 $u_k = Kx_k$，得到不同的状态反馈矩阵 K 。希望找到一个 K ，使得 J 最小。

4. 马尔科夫过程

这部分是随机过程的内容，老师上课没有讲，选看。马尔科夫过程是理解强化学习中马尔科夫决策过程的基础。篇幅原因，一些性质的推导就不列写了，大家可以自行翻阅相关资料。

4.1. 随机过程概述

随机过程，通常指随着时间或空间变化的一组随机变量，股票的价格随着时间波动(K 线图)，人说话的声音信号再时间线内变化的序列，天气随着不同天数的改变，都可以看做是随机过程。索引下标 x 决定了不同类型的随机过程。以时间为例，如果时间是离散的，则为离散随机过程；如果是连续时间，则为连续随机过程。

随机变量 y 不决定系统的类型，只决定系统的输出状态，值可以连续也可以离散，气温为连续值，取值为 $[-60, +50]$ 内的实数，天气为离散值，取自集合 $\{晴, 雨, 阴\}$ 。

离散随机过程表示如下：

$$X_0, \cdots, X_t, \cdots$$

各个时刻的随机变量之间存在着概率关系，我们需要计算这组随机变量的联合概率或者是条件概率。

Lesson 12 时间序列将讲述更多的基于随机过程的算法，这一章我们只关注马尔科夫过程。

4.2. 马尔科夫过程($Markov\ Process$)与马尔科夫性

对于普通的随机过程，当前时刻的状态 X_t 和更早时刻的状态均有关系，即存在如下条件概率：

$$p(X_t | X_0, \cdots, X_{t-1})$$

马尔科夫过程是一种特殊的随机过程，其核心假设是系统在当前时刻的状态值只和上一个时刻的状态值有关，和更早的时刻无关，称为无记忆性。利用这个马尔科夫假设，如果满足：

$$p(X_t | X_0, \cdots, X_{t-1}) = p(X_t | X_{t-1})$$

这也就是一阶马尔科夫假设，满足此假设的系统具有马尔科夫性。反复利用这个式子，可以得到随机变量序列联合概率的一个简洁计算方法：

$$p(X_0, \cdots, X_T) = p(X_T | X_{T-1}) p(X_{T-1} | X_{T-2}) \cdots p(X_1 | X_0) p(X_0)$$

该式表明如果系统有马尔科夫性，则序列的联合概率由各个时刻的条件概率值 $p(X_t | X_{t-1})$ 和初始概率 $p(X_0)$ 决定。

4.3. 马尔科夫链($Markov\ Chain$)

根据系统的状态 X ，时间 t 是否连续，可以将马尔科夫过程分为四种类型。

	可数状态空间	连续状态空间
离散时间	有限或可数状态空间的马尔科夫链	可测状态空间的马尔科夫链
连续时间	连续时间的马尔科夫过程	具有马尔科夫性的连续型随机过程

时间或状态离散的马尔科夫过程被称为马尔科夫链。马尔科夫链是马尔科夫过程的数学表示。这里我们重点介绍离散时间下的马尔科夫链，$p(X_t|X_{t-1})$ 表示系统上一个时刻的状态为 X_{t-1} ，下一个时刻转移到 X_t 的概率。对于有限或者无限可数的状态空间的马尔科夫链，可以用状态转移矩阵来表示此条件概率值。假设系统有 m 个状态：

$$P_{m\times m}=\begin{pmatrix}p_{11} & \cdots & p_{1m}\\ \vdots & & \vdots\\ p_{m1} & \cdots & p_{mm}\end{pmatrix}$$

其中，$p_{ij}=p(X_t=j|X_{t-1}=i)$ ，即从状态 i 转移到状态 j 的概率，且成立 $p_{ij}\geqslant 0$ 。

当前时刻无论除以哪一种状态 i ，下一时刻必然会转为 m 个状态中的一个，因此成立 $\sum_{j=1}^{m}p_{ij}=1$ 。即任意一行的矩阵元素之和为 1 。

重点考虑离散时间，且状态数有限的情况。对于状态连续的马尔科夫链，每个时刻各个状态的值由概率密度函数描述，状态转移概率为条件密度函数。如果任何时刻的状态转移概率相同，则称为时齐马尔科夫链。此时只有一个状态转移矩阵，任何时刻都适用。

以天气为例，$\{晴1,雨2,阴3\}$ ，符合马尔科夫假设，假设其状态转移矩阵为：

$$P=\begin{pmatrix}0.7 & 0.2 & 0.1\\ 0.4 & 0.5 & 0.1\\ 0.3 & 0.4 & 0.5\end{pmatrix}$$

如果昨天是阴天，则今天为雨天的概率是 $p_{32}=0.1$ 。可绘制状态转移图，也成为状态机。

接下来讨论状态 π 和 P 的关系。系统初始处于哪种状态也是随机的，用行向量 π 表示，需要满足：

$$\pi_i\geqslant 0\quad \sum_{i=1}^{m}\pi_i=1$$

以保证 π 是一个合法的概率分布，这是一个多项分布。

以天气为例，假设初始处于晴天概率是 0.5 ，阴天是 0.4 ，雨天是 0.3 ，则 π 为：

$$\pi=(0.5\ 0.4\ 0.3)$$

根据马尔科夫性，出现状态序列为 $X_0,\cdots,X_T$ 的概率为：

$$p(X_0,\cdots,X_T)=p(X_T|X_{T-1})p(X_{T-1}|X_{T-2})\cdots p(X_1|X_0)p(X_0)=\pi_{X_0}\prod_{t=1}^{T}p_{X_{t-1}X_t}$$

在这里，$p(X_0)=\pi_{X_0}$ 。

初始 3 天都为晴天的概率为：

$$p(X_0=1,X_1=1,X_2=1)=\pi\times p_{11}\times p_{11}=0.245$$

某一天为晴天的概率为：

$$
\begin{aligned}
p(X_t=1) &= p(X_{t-1}=1)p(X_t=1|X_{t-1}=1) + p(X_{t-1}=2)p(X_t=1|X_{t-1}=2) + p(X_{t-1}=3)p(X_t=1|X_{t-1}=3) \\
&= 0.7 \times p(X_{t-1}=1) + 0.4 \times p(X_{t-1}=2) + 0.3 \times p(X_{t-1}=3)
\end{aligned}
$$

由此得到，如果 t 时刻各个状态出现的概率为 π_t ，由于状态矩阵的第 i 列为从上一个时刻各个状态转入当前时刻状态 i 的概率，因此 t 时刻的状态为 i 的概率为：

$$
\pi_{t,i} = \sum_{j=1}^{m} \pi_{t-1,j} p_{ji}
$$

对于所有状态，写成矩阵形式为：

$$
\pi_t = \pi_{t-1} P
$$

反复利用这个式子可以得到：

$$
\pi_t = \pi_{t-1} P = \pi_{t-1} PP = \cdots = \pi_0 P^t
$$

这也是随机变量序列联合概率的矩阵表示，即给出初始状态 π_0 和状态转移矩阵 P ，可以计算出任意时刻的状态概率分布。

这个结论可以推广到多步状态转移概率。多步状态转移概率指，从一个状态开始，经过多次状态转移后，到达另一个状态的概率。定义 n 步转移概率为从 i 转移到 j 的概率：

$$
p_{ij}^{(n)} = p(X_n = j | X_0 = i)
$$

对于时齐马尔科夫链，即满足任何时刻的状态转移概率相同，则有：$p_{ij}^{(n)} = p(X_{k+n} = j | X_k = i)$ 。

以 n 步转移概率为元素的矩阵称为 n 步转移概率矩阵：

$$
P_{m\times m}^{(n)} = \begin{pmatrix}
p_{11}^{(n)} & \cdots & p_{1m}^{(n)} \\
\vdots & & \vdots \\
p_{m1}^{(n)} & \cdots & p_{mm}^{(n)}
\end{pmatrix}
$$

$n=1$ 时，该矩阵退化为普通的状态转移矩阵 $P^{(1)} = P$ 。

n 步状态转移概率满足 $Chapman-Kolmogorov$ 方程($C-K$ 方程)：

$$
p_{ij}^{(n)} = \sum_{k=1}^{m} p_{ik}^{(l)} p_{kj}^{(n-l)}
$$

该方程表示，从状态 i 经过 n 次转移到状态 j 的概率，等于从状态 i 经过 l 次转移到状态 k 的概率，乘以从状态 k 经过 $n-l$ 次转移到状态 j 的概率，对所有状态 k 进行求和的结果。可通过全概率公式和条件概率定义证明。

根据 $C-K$ 方程，可以推导出：

$$
P^{(n+l)} = P^n P^l
$$

拆分上式 n 次，可以得到状态转移矩阵和 n 步转移矩阵的关系：

$$
P^{(n)} = P^n
$$

4.4. 状态的性质与分类

4.4.1. 互通与可约

如果可以从状态 i 转移到状态 j，即存在 $n \geqslant 0$ 使得 $p_{ij}^{(n)} > 0$，则称从 i 到 j 是可达的，记为 $i \to j$。如果 $i \to j$ 且 $j \to i$，则称这两个状态是互通的，记为 $i \leftrightarrow j$。互通意味着两个状态可以互相转移。

互通具有自反性($i \leftrightarrow i$)，对称性($i \leftrightarrow j$ 可推出 $j \leftrightarrow i$)，传递性($i \leftrightarrow j$ 且 $j \leftrightarrow k$ 则 $i \leftrightarrow k$)。因此互通是一种等价关系，所有互通的状态属于一个等价类，可以按照互通性将所有状态划分成若干个不相交的子集。

如果一个马尔科夫链任意两个状态都是互通的，则称它是不可约的($irreducible$)。如果至少存在两个状态不互通，则称它是可约的。不可约的马尔科夫链任意两个顶点之间都有路径存在，即用图表示的话，它是一个强连通图。判断可不可约方式很简单，如果任意两个状态之间都是可达的(存在路径)，则马尔科夫链不可约。

4.4.2. 周期性

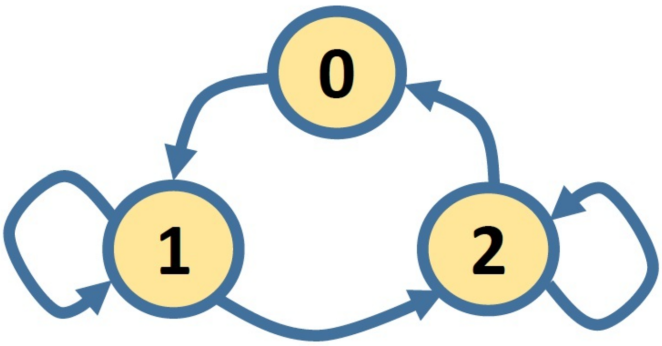
状态 i 的周期 $d^{(i)}$ 定义为假设从该状态出发，经过 n 步之后回到该状态，所有 n 的最大公约数：

$$d^{(i)} = gcd\{n > 0 : p_{ii}^{(n)} > 0\}$$

如果对所有的 $n > 0$ 都有 $p_{ii}^{(n)} = 0$，则称周期无穷大($+\infty$)，表示一定回不到原状态；如果某状态周期 $d^{(i)} > 1$，则称该状态是周期的； 如果状态的周期为 1，则它是非周期的。

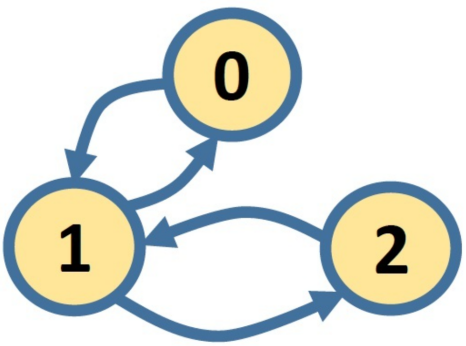
重要结论：如果两个状态互通，则它们的周期相同。

因此可以得到推论：1. 如果不可约的马尔科夫链有周期性状态 i，则其所有的状态均为周期性状态。2. 如果不可约的马尔科夫状态有一个非周期的状态 i，则其所有的状态都是非周期状态。下面给一个例子：



$0 \to 0$ 的步数可以是 $3, 4$，最大公约数为 1，因此 0 是非周期的。 而这又是个不可约的马尔科夫链，因此所有状态都是非周期的。

而下面这张就是一个周期为 2 的马尔科夫链。状态互通意味着它周期相同，因此对于不可约的周期性马尔科夫链，其周期等于任何一个状态的周期。



4.4.3. 常返性

定义 $f_{ij}^{(n)}$ 表示从状态 i 出发在时刻 n 首次进入状态 j 的概率。即：

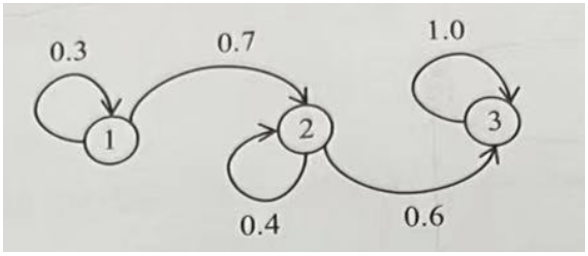
$$f_{ij}^{(0)}=0 \quad f_{ij}^{(n)}=p(X_n=j,X_k\neq j,k=1,\cdots,n-1|X_0=i)$$

定义 f_{ij} 表示从状态 i 出发迟早将转移到状态 j 的概率。即它为所有首次进入 j 的可能性之和：

$$f_{ij}=\sum_{n=1}^{+\infty}f_{ij}^{(n)}$$

f_{ii} 表示状态 i 出发迟早会返回到自己的概率。如果 $i\neq j$，则当且仅当从 i 到 j 可达时，$f_{ij}>0$ 。

如果 $f_{ii}=1$，则称状态 i 是常返的($recurrent$)，否则是非常返的($transient$)。常返意味着从一个状态出发会以概率 1 再次进入该状态，即迟早会返回该状态。



以这个马尔科夫链为例，对于状态 1，$f_{11}^{(1)}=0.3,f_{11}^{(2)}=0,f_{11}^{(3)}=0,\cdots$。状态 1 出发在时刻 $2,3,\cdots$ 首次进入状态 1 的概率为 0，因为已经离开这个状态，就无法再返回，从而有 $f_{11}=\sum\limits_{n=1}^{+\infty}f_{11}^{(n)}=0.3$ 。这说明状态 1 是非常返的。从该状态出发，有 0.7 的概率再也回不来。

对于状态 2，$f_{22}^{(1)}=0.4,f_{22}^{(2)}=0,f_{22}^{(3)}=0,\cdots$，从而有 $f_{22}=\sum\limits_{n=1}^{+\infty}f_{22}^{(n)}=0.4$ 。这说明状态 2 是非常返的。从该状态出发，有 0.6 的概率再也回不来。

对于状态 3，$f_{33}^{(1)}=1,f_{22}^{(2)}=0,f_{22}^{(3)}=0,\cdots$，从而有 $f_{33}=\sum\limits_{n=1}^{+\infty}f_{33}^{(n)}=1$ 。这说明状态 3 是常返的。

状态 i 常返的充要条件是，$\sum\limits_{n=1}^{+\infty}p_{ii}^{(n)}=+\infty$，证明略。

不可约的马尔科夫链的所有状态的常反性相同，要么全常返要么全非常返。

如果 $i\leftrightarrow j$，且 j 是常返的，则 $f_{ij}=1$ (i 迟早会转移到 j)。

如果一个状态 j 是非常返的，则对所有的 i 均有：$\sum\limits_{n=1}^{+\infty}p_{ij}^{(n)}<+\infty$，即从任何状态 i 出发到达 j 的次数的数学期望是有限的。当 $n\to+\infty$ 时，$p_{ij}^{(n)}\to 0$ 。

可以对常返做进一步分类。如果定义 μ_{ii} 为返回 i 所需要的平均转移次数(平均返回时间)，即：

$$\mu_{ii}=\begin{cases}+\infty, & i\text{非常返}\\ \sum\limits_{n=1}^{+\infty}nf_{ii}^{(n)}, & i\text{常返}\end{cases}$$

假设 i 常返，如果 $\mu_{ii}<+\infty$，则 i 是正常返的。正常返意味着从一个状态出发不但会以 1 的概率再次返回(迟早回来)，而且返回该状态的平均时间是有限的，μ_{ii} 表达式的级数收敛。同样，如果 $\lim\limits_{n\to+\infty}p_{ii}^{(n)}>0$，即 n 次转移从 $i\to i$ 的概率大于 0，则 i 是正常返。

假设 i 常返，如果 $\mu_{ii} = +\infty$，则 i 是零常返的。零常返只保证返回的概率是 1，但是平局返回时间是 $+\infty$。同样，如果 $\lim\limits_{n\to+\infty} p_{ii}^{(n)} = 0$，即 n 次转移从 $i \to i$ 的概率不存在，则 i 是正常返。

如果一个马尔科夫链的状态数是有限的，则只存在正常返和非常返的状态，不存在零常返的状态。因此，有限状态且不可约的马尔科夫链的所有状态都是正常返的。

4.5. 平稳分布和极限分布($Stationary\ \&\ Limiting\ Distribution$)

马尔科夫链有一个有趣的性质，对于任意的初始状态分布，随着状态转移的进行，最后系统的状态概率分布趋向于一个稳定的收敛值。

定义平稳分布：假设状态空间的大小为 m，状态的概率分布为向量 π，对于状态转移矩阵为 P 的马尔科夫链，如果存在一个概率分布 π 满足：

$$\pi P = \pi$$

则称此分布 π 为平稳分布。即当前时刻服从此分布，转移到下一时刻还是服从此分布。或者说，上一时刻处于某种状态的概率，和下一时刻从各个状态进入该状态的概率相同。这也意味着如果平稳分布作为初始状态，则任何次转移后，状态的概率分布不变。

根据定义可推论，平稳分布是 P^T 归一化的特征向量，且特征值为 1。因为 $(\pi P)^T = P^T\pi^T = \pi^T$，即 $(P^T - I)\pi = 0$，这就是矩阵特征值和特征向量(Lesson 5 监督学习之分类(Perceptron, Fisher, Logistic, Softmax, Bayes))的表达式，特征值为 1，平稳分布 π 就是这个齐次方程归一化的非 0 解。

因此，给定一个马尔科夫链的状态转移矩阵 P，求解 $(P^T - I)x = 0$ 即可得到其平稳分布。其次方程组的解不唯一，平稳分布需要满足前文 $\pi_i \geqslant 0, \sum\limits_{i=1}^{m}\pi_i = 1$ 的约束条件，才可以确定唯一解。以前面的天气转移矩阵举例：

$$P = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.4 & 0.5 & 0.1 \\ 0.3 & 0.4 & 0.5 \end{pmatrix}$$

列写方程：$(P^T - I)\begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix} = 0$，即：

$$\begin{cases} \pi_1 = 0.7\pi_1 + 0.4\pi_2 + 0.3\pi_3 \\ \pi_2 = 0.2\pi_1 + 0.5\pi_2 + 0.4\pi_3 \\ \pi_3 = 0.1\pi_1 + 0.1\pi_2 + 0.3\pi_3 \\ \pi_1 + \pi_2 + \pi_3 = 1 \end{cases}$$

通过Lesson 1 解方程的高斯消元法，解方程可以得到唯一解：$\pi_1 = 0.554, \pi_2 = 0.321, \pi_3 = 0.125$。

可以通过特征值分解(Lesson 8 无监督学习(聚类, 信号分解, 流形降维))计算所有可能的平稳分布。求解 P^T 的特征值，得到 $\lambda_1 = 1, \lambda_2 = 0.3, \lambda_3 = 0.2$。对于 $\lambda = 1$，$P^T - I$ 初等行变换，得到 $\pi_1 = \frac{31}{7}t, \pi_2 = \frac{18}{7}t, \pi_3 = t$，添加归一化约束条件 $\pi_1 + \pi_2 + \pi_3 = 1$，得到：$\pi_1 = \frac{31}{56}, \pi_2 = \frac{18}{56}, \pi_3 = \frac{7}{56}t$，这与前面的结论一致。$\lambda_2, \lambda_3$ 不符合平稳定义的要求(特征值不为 1)，不予考虑。该天气转移矩阵的平稳分布存在且唯一。

不是所有马尔科夫链都存在平稳分布且唯一。比如下面的例子：

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

该马尔科夫链是可约的，任意量状态均不互通。$\lambda = 1$ 是 P^T 的三重特征值。列写方程：$(P^T - I)\begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix} = 0$。任意非 0 向量都是这个方程的非 0 解，这意味着任何合法的概率分布 π 都是平稳分布。

平稳分布也可以表示为当时间趋于 $+\infty$ 时，每个状态出现次数的比例。令 $N_i(n)$ 表示到 n 时刻进入状态 i 的总次数，则有：

$$\pi_i = \lim_{n \to +\infty} \frac{N_i(n)}{n}$$

定义极限分布：极限分布是当时间趋于 $+\infty$ 时，状态 j 出现的概率，该概率与起始状态 i 无关，以任意的 i 初始进行演化，最后转移到状态 j 的概率都是相同的。

$$\pi_j = \lim_{n \to +\infty} p(X_n = j | X_0 = i) = \lim_{n \to +\infty} p_{ij}^{(n)} > 0,\ j = 1, \cdots, m$$

存在性和唯一性：

- 如果一个马尔科夫链是非周期、不可约的，当且仅当所有的状态都是正常返($\lim_{n \to +\infty} p_{ii}^{(n)} > 0$)时，平稳分布存在且唯一，且此时平稳分布等于极限分布。
- 推论：根据前文结论，如果一个马尔科夫链是非周期、不可约的，且其状态数是有限的，可以推出它的所有状态一定是正常返的，因此平稳分布存在且唯一。
- 如果一个马尔科夫链是非周期、不可约的，且其所有状态全是非常返的，或全是零常返的，则对 $\forall i, j$ 有 $\lim_{n \to +\infty} p_{ii}^{(n)} = 0$，此时平稳分布不存在。
- 如果一个马尔科夫链是可约的，通常情况下存在多个平稳分布。

4.5.1. 二者关系

极限分布是时间趋于 $+\infty$ 时，状态 j 出现的概率。

$$\pi_j = \lim_{n \to +\infty} p(X_n = j | X_0 = i) = \lim_{n \to +\infty} p_{ij}^{(n)} > 0,\ j = 1, \cdots, m$$

且该概率与起始状态 i 无关，因此可以简写为：

$$\pi_j = \lim_{n \to +\infty} p(X_n = j),\ \forall j \in S$$

因此， 根据上式和前文的式子 $\pi_t = \pi_{t-1} P = \pi_{t-1} P P = \cdots = \pi_0 P^t$ ，对 $\pi_t = \pi_0 P^t$ 两侧取极限可得：

$$\pi = \lim_{n \to +\infty} \pi_0 P^{n+1} = (\lim_{n \to +\infty} \pi_0 P^n) P = \pi P$$

这意味着极限分布就是平稳分布。

根据结论平稳分布等于极限分布的唯一性条件和 $C-K$ 方程($p_{ij}^{(n)} = \sum_{k=1}^{m} p_{ik}^{(l)} p_{kj}^{(n-l)}$)，可以推导出，如果平稳分布存在，则状态转移矩阵幂的极限也存在，且等于平稳分布。

$$\lim_{n \to +\infty} P^{(n)} = \lim_{n \to +\infty} P^n = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_m \\ \pi_1 & \pi_2 & \cdots & \pi_m \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_m \end{pmatrix}$$

极限分布和平稳分布刻画了马尔科夫链的重要属性，实际应用中有着重要的价值。

4.6. 细致平衡条件($Detailed\ Balance$)

通过细致平稳条件，可以构造出状态转移矩阵 P ，使其平稳分布 π 满足给定的状态概率分布 π 的马尔科夫链。

定义细致平衡条件：如果 P 对于所有的 i, j 满足：

$$\pi_i p_{ij} = \pi_j p_{ij}$$

即对于 $\forall i,j$ ，处于状态 i 的概率乘以从 $i \to j$ 的转移概率等于处于状态 j 的概率乘以从 $j \to i$ 的转移概率，则 π 为马尔科夫链的平稳分布。证明略

需要注意的是，P 和 π 满足细致平衡条件是 π 为 P 的平稳分布的充分非必要条件。即只能通过细致平衡条件推出平稳分布，但是平稳分布不一定满足细致平衡条件。

直观上看，平稳分布意味着对于同一个状态而言，从其转出的概率和转入其的概率相等。细致平衡条件的要求则更严格，它意味着对于任意两个状态 i,j ，$i \leftrightarrow j$ 的概率相同。

对于状态连续的马尔科夫链，细致平衡条件同样成立，即：

$$p(x)p(x'|x) = p(x')p(x|x')$$

细致平衡条件常用语马尔科夫链的蒙特卡洛采样算法的实现。

5. 隐马尔科夫模型(HMM)

对于概率图模型，Lesson 3.5 参数估计(MLE, MAP, Bayes, KNN, Parzen, GMM, EM算法)已经介绍了变分推断算法，这里我们介绍基于马尔科夫模型的扩充模型，隐马尔科夫模型。自动控制原理下中，状态空间的分析与拟合一章和隐马尔科夫模型有很多共通之处，可以进行对比学习。

很多应用中系统的状态值是很难观察到的，是隐变量，只能得到一组观测值。隐马尔科夫模型为通过观测值来推断出状态值提供了解决方案，它实质上是增加了观测模型后的马尔科夫链。

隐马尔科夫模型是著名的有向图模型，结构最简单的动态贝叶斯网络，常用于时序数据的建模，语音识别，自然语言处理等，它描述了观测变量和状态变量之间的概率关系。和马尔科夫过程对状态建模不同，隐马尔科夫模型除了对状态建模，还对观测值建模。

比如说，视频中人的动作，站立、坐下、行走，是状态值，识别前无法得到确切值。人的关键点坐标是观测值，能直接得到。HMM 通过观测值推断出状态值，从而识别动作。

首先定义观测序列，它是直接能够观察或者计算得到的值：

$$x = \{x_1, \cdots, x_T\}$$

这是一个随机变量序列，任何时刻的观测值来自有限的观测集：$V = \{v_1, \cdots, v_m\}$ ，即 x 的取值范围。

定义状态序列，它是不客观的，隐藏的，也是模型中的隐变量：

$$z = \{z_1, \cdots, z_T\}$$

它也是一个随机变量序列，任何时刻的状态值来自有限的状态集：$S = \{s_1, \cdots, s_n\}$ ，即 z 的取值范围。状态序列是一个普通的马尔科夫链，其状态转移矩阵为 A ，状态转移矩阵在马尔科夫链中详细介绍了性质和分类，这里不做赘述。状态随着时间线演化，每个时刻的状态值决定了观测值。

状态转移矩阵 A 的元素 $a_{ij} = p(z_{t+1} = s_j | z_t = s_i)$ 。

定义观测矩阵 B ，其元素为 $b_{ij} = p(x_t = v_j | z_t = s_i)$ 。该值表示根据 t 时刻状态值为 s_i 时获得观测值为 v_j 的概率。它和状态转移矩阵 A 一致，也需要满足同样的约束条件：

$$b_{ij} \geqslant 0 \quad \sum_{i=1}^{n} b_{ij} = 1$$

B 的第 i 行是状态为 s_i 时观测值为各个值的概率分布。

假设初始概率分布为 π ，其元素为 $\pi_i = P(y_1 = s_i)$ ，即系统初始时个状态的概率分布。

HMM 可以抽象为五元组：

$$(S, V, \pi, A, B)$$

一般假设马尔科夫链是时齐的，即 A, B 在任何时刻都不变，简化问题计算。

系统在 1 时刻处于状态 z_1 ，在该状态下得到观测值 x_1 。接下来从状态 $z_1 \to z_2$ ，在 z_2 下又得到了观测值 x_2 ，以此类推，就得到了整个观测序列。用算法描述这个前向过程为：

- 根据 π 选择初始状态 z_1 。
- 根据 z_t 和 B 得到了 x_t 。
- 根据 z_t 和 A 得到了 z_{t+1} 。
- $t = t + 1$ ，转到第二步，直到时刻 T 。

可以得到，当出现状态序列为 z 且观测序列为 x 的概率为：

$$\begin{aligned} p(z, x) &= p(z)p(x|z) \\ &= p(z_T|z_{T-1})p(z_{T-1}|z_{T-2}) \cdots p(z_1|z_0)p(x_T|z_t)p(x_{T-1}|z_{T-1}) \cdots p(x_1|z_1) \\ &= \Big(\prod_{t=1}^{T} a_{z_{t-1}z_t}\Big) \prod_{t=1}^{T} b_{z_t z_t} \end{aligned}$$

上式的形式满足条件独立性假设，式中，$p(z_1|z_0) = p(z_1)$ 为状态的初始概率。

基于这个式子可以看出，不管是根据给定的模型计算产生某观测序列的概率，根据给定的模型和观测序列找最匹配的状态序列，根据给定的观测序列调整模型参数让该观测序列出现的概率最大，这三个问题都很好解决。

根据以往的观测序列推测最可能出现的观测值，根据语音信号找到最匹配的文字，根据训练样本得到最优的模型参数，恰好就对应着这三个问题。

以第二个问题为例，对于天气的马尔科夫链，假设我们无法直接得知天气情况 $\{\text{晴}, \text{雨}, \text{阴}\}$，但是能得知一个人在各种天气下的活动情况，假设有三种观测 $\{\text{睡觉}, \text{跑步}, \text{逛街}\}$ 。即天气是状态值，活动是观测值。

状态转移矩阵和观测矩阵通过样本学习而得到。假设通过学习后，得到了这一问题的观测矩阵为：

$$B = \begin{pmatrix} 0.1 & 0.7 & 0.2 \\ 0.5 & 0.2 & 0.3 \\ 0.7 & 0.1 & 0.2 \end{pmatrix}$$

例如 $b_{13} = 0.2$ 表示在晴天的天气状态下，观测到这个人逛街的概率是 0.2 。

状态转移矩阵为：

$$A = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.4 & 0.5 & 0.1 \\ 0.3 & 0.4 & 0.5 \end{pmatrix}$$

在给定了观测序列 x 的条件下，可以计算出状态序列 z 出现的概率，即条件概率 $p(z|x)$ 。

6. 强化学习的算法思想

西瓜书中的种西瓜的例子我觉得说的很生动形象。首先，种瓜动作中，是无法立即知道最终奖励值的，只能得到一个当前的反馈值(瓜看起来更大了)，只有当西瓜收获时，才能知道这个瓜到底种的好不好(最终的奖励)，并且才能根据瓜的情况来修改下一次种西瓜的策略。因此，多次种瓜不断摸索(迭代)，根据当前瓜的生长情况(状态)，采取不同的种瓜动作，根据采取种瓜动作后瓜的生长情况改变(状态转移，与随之而来的惩罚和奖励)，不断调整种瓜动作，直到找到最优种瓜策略(状态 $\to$ 动作)，让算法最终收敛，这也就是强化学习的核心思想。

用数学符号表示为：对于一个强化学习模型，在当前状态 s 下执行某一动作 a ，然后进行下一个状态，并受到一个反馈(奖励值) R ，反复执行，直到收敛。算法的目标函数即确定一个策略函数 π ，实现从状态到动作的映射 $a = \pi(s)$ 。

详细来说，即在每个时刻 t ，算法在状态 s_t 下执行了动作 a_t 后，系统进入下一个状态 s_{t+1} ，并给出一个奖励值 R_{t+1} 。整个过程中，算法随机执行动作，收到惩罚和奖励反馈，从而学习到想要的行为策略 π 。系统对正确的动作做出奖励，错误的做出惩罚，训练完成后用得到的策略函数 π 进行状态 $\rightarrow$ 动作的预测。

和监督学习的区别：强化学习中的 $State$ 对应着监督学习中的样本 X ，而 $Action$ 对应着监督学习中的标签 y，"策略"则对应着监督学习中的分类 / 回归函数 h ，奖励和惩罚对应着监督学习中的损失函数 $loss$ 。但不同的是强化学习中没有有标签样本对($X-y$)，即没人能告诉机器在什么状态下应该做什么样的动作，只有等到最终的结果揭晓，才能通过反思之前的动作是否正确来进行学习(奖励和惩罚)。因此，强化学习可以看成具有"延迟标记信息"的监督学习问题。

7. 马尔科夫决策过程(MDP)

三者关系：马尔科夫过程是描述状态随时间变化的随机过程，隐马尔科夫模型是增加了观测序列的马尔科夫过程，马尔科夫决策过程是增加了动作与奖励机制的马尔科夫过程。

强化学习算法需要对问题的不确定性建模，即将解决的问题抽象为一个马尔科夫决策过程。

7.1. MDP 的符号定义

MDP 可由五符号进行定义：

$$\mathcal{M} \triangleq (\mathcal{S}, \mathcal{A}, \{P_{sa}\}, \gamma, \mathcal{R})$$

其中 $\mathcal{S}$ 是状态空间，即状态的集合(状态变量可取的空间)，每个状态记为 s 。举例说明，对于围棋，就是当前的棋局 $3^{19\times19}$ (19×19 的棋盘，每个块有 3 种状态)，其状态是离散的。对于自动驾驶，就是现在的位置、速度等连续值。

$\mathcal{A}$ 是动作空间，即行为的集合，每个动作记为 a 。在每种状态 s 下，可以执行的动作记为 $A(s)$ 。举例说明，围棋就是空的地方落子；自动驾驶就是汽车的速度 (v_x, v_y) 。

$\{P_{sa}\}$ 是状态转移概率，也可以定义为 $P^a_{s\rightarrow s'}$ ，表示在状态 s 下，执行动作 a ，状态在下个时刻转移成 s' 的 概率。

$$p(s'|s,a) = p(s_{t+1}=s'|s_t=s,a)$$

它需要满足 $\sum p(s'|s,a)=1$ 。根据马尔科夫性，下一时刻的状态只和当前时刻的状态和当前时刻采取的动作有关。以围棋为例，下一时刻的动作，只和当前时刻的棋局、当前时刻的落子动作，以及接下来对手在当前时刻落子的动作有关。而对手的落子是不确定的，满足随机性。

γ 是遗忘因子，用于定义累积回报和价值函数， $\gamma\in(0,1)$ 。

$\mathcal{R}:\mathcal{S}\times\mathcal{A}\rightarrow\mathbb{R}$ 为奖励函数，即从 s 到 s' 执行了 a 的回报。t 时刻的回报记为 $\mathcal{R}_t$ ，由当前时刻的状态，动作和下一时刻的状态决定

$$\begin{gathered} X_{k+1}=Ax_k+Bu_k,\ u_k=Kx_k \\ x_k\in\mathbb{R}^n,\ u_k\in\mathbb{R}^m \\ \text{对应}\ \mathcal{S},\mathcal{A} \end{gathered}$$

7.2. MDP 的动力学定义

MDP 的动力学可由下式定义：

$$s_0\rightarrow^{a_0}s_1\rightarrow^{a_1}s_2\cdots$$

根据上面的 MDP 符号定义，整个动力学过程可以定义为：对于系统，将动作 a 作用于状态 s 上，通过 P 来计算 $s\rightarrow s'$ ，进行状态转移，转移的过程中，环境会根据 R 函数反馈给系统一个奖励值，系统根据奖励值调整策略，直到找到最优。

假设策略 π 定义为 $\mathcal{S}\rightarrow\mathcal{A}$ ，即 $a=\pi(s)$ 。表示根据该策略，在不同的状态下就能找到不同的动作。

$$V^\pi(s)=\mathbb{E}^\pi(\mathcal{R}(s_0,\pi(s_0))+\gamma\mathcal{R}(s_1,\pi(s_1))+\cdots\quad|s_0=s,\pi)$$

类比于最优控制中的 u，强化学习的目标即为根据当前状态 $\mathcal{S}$，找到一组最优的动作 $(a_0, a_1, \cdots)$，以最大化 $E(\mathcal{R}(s_0, a_0) + \gamma\mathcal{R}(s_1, a_1) + \cdots)$，即找到能使这个长期累积奖赏最大化的策略。

8. 有模型学习 - 策略评估与 $Bellman$ 方程求解

如果马尔科夫决策过程的五元组已知，即机器已经对环境建模，这称之为有模型学习。基于策略评估和动态规划的求解方式就是典型的有模型学习算法。

这里我们只介绍有模型学习，其他的优化算法包括无模型学习，值函数近似，模仿学习等，大家可以翻阅相关资料。

8.1. 策略评估

根据强化学习的算法思想，从 t 时刻起的收益函数定义为：

$$G_t = \mathcal{R}_{t+1} + \gamma\mathcal{R}_{t+2} + \gamma^2\mathcal{R}_{t+3} + \cdots = \sum_{k=0}^{+\infty}\gamma^k\mathcal{R}_{t+k+1}$$

长期累积奖赏需要采取折扣累积的方法，因此引入 γ 折扣因子，平衡短期收益和长远收益，同时保证上面的级数收敛。如果级数收敛，则有：

$$G_t = R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots) = R_{t+1} + \gamma G_{t+1}$$

强化学习的目标就是最大化这个累计奖励，本质上是一个优化问题。

算法需要确保所有的状态按照某一策略执行，得到的累积回报均为最大化。因此，定义状态价值函数($States\ Value$)，表示状态 s 下预计累计回报的期望值。或者说，表示在状态 $s_t = s$ 情况下，反复按照策略 π 执行，所得到的累计奖励的数学期望：

$$V_\pi(s) = \mathbb{E}_\pi[G_t|s_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{+\infty}\gamma^k R_{t+k+1}|s_t = s\right]$$

使用数学期望是因为系统具有随机性，需要对所有情况的累计奖励计算均值。

类似的可以定义状态-动作价值函数($Action\ Value$)，表示状态 s 下采取动作 a 的预计累计回报的期望值。更细致地描述，表示当前状态 $s_t = s$ 的情况下执行动作 $a_t = a$，然后反复按照策略 π 执行，所得到的累计奖励的数学期望：

$$Q_\pi(s, a) = \mathbb{E}_\pi[G_t|s_t = s, a_t = s] = \mathbb{E}^\pi\left[\sum_{k=0}^{+\infty}\gamma^k R_{t+k+1}|s_t = s, a_t = a\right]$$

s 是当前状态，为一个限制条件，跳变过程中 $s_0 = s$ 。

8.2. 符号定义

定义 G_t 为从 t 时刻起的奖励函数。

定义状态价值函数 $V_\pi(s)$ 表示状态 s 下预计累计回报的期望值。

定义状态-动作价值函数 $Q_\pi(s, a)$ 表示状态 s 下采取动作 a 的预计累计回报的期望值。

除此之外，定义 $P^a_{s\to s'}$ 表示采取行为 a 后状态 s 转为 s' 的概率：

$$s_0 \to^{a_0} = \begin{cases} s_1, & P^{a_0}_{s_0\to s_1} \\ s'_1, & P^{a_0}_{s_0\to s'_1} \\ s''_1, & P^{a_0}_{s_0\to s''_1} \end{cases}$$

定义 $R^a_{s\to s'}$ 表示采取行为 a 后状态 s 转为 s' 所获得的奖励。

定义 $\pi(s, a)$ 表示状态 s 下根据策略 π 采取行为 a 的概率。

8.3. 迭代公式推导

根据上面的动作概率的定义和马尔科夫性，即系统下一时刻的状态仅由当前时刻状态决定，和更早的状态无关的性质，可以推导出状态价值函数 V_π 的迭代公式：

$$\begin{aligned} V_\pi(s) &= \mathbb{E}[G_t|s_t=s] \\ &= \mathbb{E}[R_{t+1}+(\gamma R_{t+2+\cdots})|s_t=s] \\ &= \mathbb{E}[R_{t+1}+\gamma G_{t+1}|s_t=s] \\ &= \sum_{a\in\mathcal{A}}\pi(s,a)\sum_{s'\in\mathcal{S}}P^a_{s\to s'}(R^a_{s\to s'}+\gamma\mathbb{E}[[G_t|s_{t+1}=s']) \\ &= \sum_{a\in\mathcal{A}}\pi(s,a)\sum_{s'\in\mathcal{S}}P^a_{s\to s'}(R^a_{s\to s'}+\gamma V_\pi(s')) \\ &= \mathbb{E}[R_{t+1}+\gamma V_\pi(s_{t+1})|s_t=s] \end{aligned}$$

通俗理解：$V_\pi = 1+\frac{1}{2}+\frac{1}{4}+\cdots = 1+\frac{1}{2}(1+\frac{1}{2}+\frac{1}{4}+\cdots) = \cdots$ 。

类似的，状态-动作价值函数 Q_π 的公式也可以写出：

$$Q_\pi(s,a) = \sum_{s'\in\mathcal{S}}P^a_{s\to s'}(R^a_{s\to s'}+\gamma V_\pi(s'))$$

即：

$$V_\pi(s) = \sum_{a\in\mathcal{A}}\pi(s,a)Q_\pi(s,a)$$

上面的等式又被称为贝尔曼方程(贝尔曼等式)，用于表示动态规划问题中相邻状态关系的方程。用上面的递归等式来计算值函数，实际上就是一种动态规划的算法。

8.4. 策略改进 - 最优 $Bellman$ 方程

上面的式子定义了策略的评估方式。对策略进行评估后，如果不是最优策略，则希望对其改进，理想的最优策略应该能最大化累积奖赏：

$$\pi^* = \arg\max_\pi \sum_{s\in\mathcal{S}} V^\pi(s)$$

一个强化学习任务可能有多个最优策略，最优策略对应的值函数 $V^*(s)$ 为最优值函数，即

$$\forall s\in S, V^*(s) = V_{\pi^*}(s)$$

由于最优值函数的累积奖赏值应最大，因此前面的贝尔曼方程 $V_\pi(s)$ 可以进行一些改动，将对动作的概率求和改为取最优：

$$V^*(s) = \max_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}P^a_{s\to s'}(R^a_{s\to s'}+\gamma V^*(s'))$$

这是一种贪心的做法。因为这是 MDP 过程，当前状态只与上一状态相关，所以能使当前状态下值函数最大的动作，一定是最大化当前状态下累积奖赏的动作，也即最优状态值函数：

$$V^*(s) = \max_{a\in\mathcal{A}} Q_{\pi^*}(s,a)$$

将式子代回前面的状态-动作值函数，得到最优状态-动作值函数：

$$Q^*(s,a) = \sum_{s'\in\mathcal{S}}P^a_{s\to s'}(R^a_{s\to s'}+\gamma\max_{a'\in\mathcal{A}}Q^*(s',a'))$$

上述关于最优值函数的等式，称为最优 $Bellman$ 方程，其唯一解是最优值函数。

最优 $Bellman$ 方程揭示了非最优策略的改进方式，即将策略选择的动作改变为当前的最优动作。假设动作改变后对应的策略为 π'，改边动作的条件为 $Q_\pi(s,\pi'(s)) \geqslant V_\pi(s)$，则如下不等式成立：

$$\begin{aligned} V_\pi(s) &\leqslant Q_\pi(s,\pi'(s)) = \sum_{s'\in\mathcal{S}} P_{s\to s'}^{\pi'(s)}(R_{s\to s'}^{\pi'(s)} + \gamma V_\pi(s')) \\ &\leqslant \sum_{s'\in\mathcal{S}} P_{s\to s'}^{\pi'(s)}(R_{s\to s'}^{\pi'(s)} + \gamma Q_\pi(s',\pi'(s'))) \\ &= \cdots = V_{\pi'}(s) \end{aligned}$$

省略号省略的部分就是将 Q 值再代入为 V，和第二步一样循环迭代，这一步说明每一次满足 $Q_\pi(s_i,\pi'(s_i)) \geqslant V_\pi(s_i)$，都会增大不等式，而不等式右端又恒等于 $V_{\pi'}(s)$，则在 $\pi \to \pi'$ 的不断的迭代中，一定有 $V_\pi(s) \leqslant V'_\pi(s)$。

由上可知，值函数对于策略的每一点改进都是单调递增的，或者说至少是不递减的。因此对于当前策略 π，必然可以将其改进为：

$$\pi'(s) = arg\max_{a\in\mathcal{A}} Q_\pi(s,a)$$

直到 $\pi'(s) = \pi(s)$，即希望通过 $Bellman$ 方程找 π^*，使得不等式的左右两边相等，通过不断迭代直至收敛。

8.5. 算法描述：策略迭代和值迭代

首先说明策略迭代，8.3. 中我们知道了如何评估一个策略的值函数，8.4 我们知道了如何在策略评估后进行改进以获得最优策略。二者结合就是策略迭代算法：即从初始策略出发，先进行策略评估，然后改进策略，评估改进的策略，再进一步改进策略，直到策略收敛，不再改变。

1. $\forall s \in \mathcal{S}$，初始化 $V(s) = 0$。
2. $Loop$：
3. 评估：迭代 $\pi \to \pi'$，更新 $V(s)$，直至收敛(即 $\max\limits_{s\in\mathcal{S}} |V(s) - V'(s)| < \theta$)：

$$V'(s) = \sum_{a\in\mathcal{A}} \pi(s,a) \sum_{s'\in\mathcal{S}} P_{s\to s'}^{a}(R_{s\to s'}^{a} + \gamma V_\pi(s'))$$

4. 改进：$\forall s \in \mathcal{S}, \pi'(s) = arg\max\limits_{a\in\mathcal{A}} Q_\pi(s,a)$。
5. if 策略收敛，即 $\forall x : \pi'(s) = \pi(s) : Goto\ 6.$
   $Else : Continue\ 2.$
6. $end\ Loop$
7. $Output$：最优策略 π

策略改进和值函数的改进是一致的，因此可以将策略改进视为值函数的改进，得到值迭代算法：

1. $\forall s \in \mathcal{S}$，初始化 $V(s) = 0$。
2. $Loop$：
3. 迭代 $\pi \to \pi'$ 直至收敛到 $V^*(s)$，即 $\max\limits_{s\in\mathcal{S}} |V(s) - V'(s)| < \theta$ 时停止 $Loop$

$$\forall s \in \mathcal{S}, \quad V(s) = \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} P_{s\to s'}^{a}(R_{s\to s'}^{a} + \gamma V(s'))$$

4. $Output : \pi(s) = arg\max\limits_{a\in\mathcal{A}} Q_\pi(s,a)$

在模型已知时，强化学习的目标任务为基于动态规划的最优化问题。这里并未涉及模型的泛化能力，而是为每一个状态找到最优的动作。

除了有模型学习的动态规划算法，其他的优化算法包括无模型学习中的蒙特卡洛强化学习，时序差分学习($Sarsa$ 算法，Q 学习)，值函数近似，模仿学习中的直接模仿学习和逆强化学习等。