计算语言学工作者需要了解的数学知识

常宝宝
北京大学计算语言学研究所，100871
chbb@pku.edu.cn

计算语言学是一门交叉学科，其中不仅涉及到语言学、计算机科学，还大量应用到数学知识。尤其是近年来，随着语料库语言学的兴起，统计等数学方法和技术在计算语言学中更是得到了越来越广泛的应用。

第一节 概率统计基础[①]

一 事件和概率

定义 1. 随机事件: 在一定条件下,可能发生也可能不发生的试验结果称为随机事件，简称**事件**，一般用大写拉丁字母 $A,B,C,\ldots$表示。

随机事件有两个特殊情况，即**必然事件**(在一定条件下，每次试验都必定发生的事件)和**不可能事件**(在一定条件下，每次试验都一定不发生的事件)，分别记为Ω和Φ。

随机事件在一次试验中是否发生，固然是无法肯定的偶然现象，但当进行多次重复试验，就可以发现其发生的可能性大小的统计规律性。具体说，如果在相同条件下进行了 n 次重复试验，事件 A 出现了 v 次，那么事件 A 在 n 次实验中出现的**频率**为是$\frac{v}{n}$。当 n 无限增大时呈现稳定性。这一统计规律性表明事件发生的可能性大小是事件本身所固有的、不以人们主观意志而改变的一种客观属性。

事件之间的关系和运算

(1) **包含** 当事件 B 发生时，如果事件 A 也一定发生，则称 A 包含 B 或 A 包含于 B 中，记作 $A\supset B$ 或 $B\subset A$。

(2) **等价** 如果 $A\supset B$ 且 $B\supset A$，即事件 A 和 B 同时发生或不发生，则称 A 和 B 等价，记作 $A=B$。

(3) **积** 表示事件 A 和 B 同时发生的事件，称为 A 与 B 的积，记作 $A\cap B$ 或 AB。

(4) **和** 表示事件 A 或事件 B 发生的事件，称为 A 与 B 的和，记作 $A\cup B$ 或 $A+B$。

(5) **差** 表示事件 A 发生而事件 B 不发生的事件，称为 A 与 B 的差，记作 $A-B$。

(6) **互斥** 如果事件 A 与 B 不可能同时发生，即 $AB=\Phi$，则称 A 与 B 是互斥的。

(7) **对立** 如果事件 A 与 B 互斥，又在每次试验中不是出现 A 就是出现 B，即 $AB=\Phi$且 $A+B=\Omega$，则称 B 为 A 的对立事件，记作 $B=\overline{A}$ 。

定义 2. 概率:事件 A 发生的可能性大小称为事件的概率，记作 $P(A)$。

[①] 若读者对概率统计方面的基本概念已经熟知，可以越过本节直接阅读下一节

当试验的次数 n 足够大，可以用事件的频率近似地表示该事件的概率，即

$$P(A)\approx\frac{v}{n}$$

概率的基本性质：

(1) $0\leqslant P(A)\leqslant 1$。

(2) $P(\Omega)=P$(必然事件)=1。

(3) $P(\Phi)=P$(不可能事件)=0。

(4) $P(A+B)=P(A)+P(B)-P(AB)$。若 A,B 互斥，则 $P(A+B)=P(A)+P(B)$，若 $A_1,A_2,\ldots,A_n$ 两两互斥，则 $P(A_1+A_2+\ldots+A_n)=P(A_1)+P(A_2)+\ldots+P(A_n)$。

(5) 若 $A\supset B$，则 $P(A)\geqslant P(B)$。

(6) 若 $A_1,A_2,\ldots,A_n$ 两两互斥，且 $A_1+A_2+\ldots+A_n=\Omega$，则

$$P(A_1+A_2+\ldots+A_n)=P(A_1)+P(A_2)+\ldots+P(A_n)=\sum_{i=1}^{n}P(A_i)=1。$$

(7) 对任意事件 A，$P(\overline{A})=1-P(A)$。

定义 3 条件概率 在事件 B 发生的条件下，事件 A 发生的概率称为事件 A 在事件 B 已发生的条件下的条件概率，记作 $P(A|B)$。当 $P(B)>0$ 时，规定

$$P(A\mid B)=\frac{P(AB)}{P(B)}$$

当 $P(B)=0$ 时，规定 $P(A|B)=0$。

由条件概率的定义，可得到**乘法公式**[②]：

$$P(AB)=P(A)P(B|A)$$

$$P(A_1A_2\ldots A_n)=P(A_1)P(A_2\mid A_1)P(A_3\mid A_2A_1)\ldots P(A_n\mid A_{n-1}A_{n-2}\ldots A_1)$$

$$=\prod_{i=1}^{n}P(A_i\mid A_{i-1}A_{i-2}\ldots A_1)$$

一般而言，条件概率 $P(A|B)$与概率 $P(A)$是不等的。但在某些情况下，它们是相等的。根据条件概率的定义和乘法公式有

$$P(AB)=P(A)P(B)$$

这时，称事件 A 与 B **相互独立**的。

贝叶斯(Bayes)**公式**:根据乘法公式，可以得到下面的重要公式，该公式称为贝叶斯公式

$$P(A\mid B)=\frac{P(B\mid A)P(A)}{P(B)}$$ [③]

[②] 乘法公式在统计自然语言处理中会多次使用到。

[③] 一般地，诸事件 $A_1,A_2,\ldots,A_n$ 两两互斥，事件 B 满足 $B\subset A_1+A_2+\ldots+A_n$，且 $P(A_i)>0(i=1,2,\ldots,n)$，$P(B)>0$，贝叶斯公式可以推广为

$$P(A_j\mid B)=\frac{P(A_j)P(B\mid A_j)}{P(A_1)(B\mid A_1)+\ldots+P(A_n)P(B\mid A_n)}=\frac{P(A_j)P(B\mid A_j)}{\sum_{i=1}^{n}P(A_i)P(B\mid A_i)}$$

实用上称，$P(A_1)$，$P(A_2)$，…，$P(A_n)$的值为验前概率，称 $P(A_1|B)$，$P(A_2|B)$，…，$P(A_n|B)$的值为验后概率，贝叶斯公式便是从验前概率推算验后概率的公式。

二 随机变量与分布函数

定义 4 随机变量 每次试验的结果可以用一个变量 $\boldsymbol{X}$ 的数值来表示，这个变量称为随机变量。

随机变量的取值随试验的结果变化，但又遵从一定的概率分布规律。它是随机现象的数量化。

定义 5 分布函数 给定随机变量 $\boldsymbol{X}$，它的取值不超过实数 x 的事件的概率 $P\{\ \boldsymbol{X}\leqslant x\}$是 x 的函数，称为 $\boldsymbol{X}$ 的概率分布函数，简称为分布函数，记作 $F(x)$，即

$$F(x)=P\{\ \boldsymbol{X}\leqslant x\} \qquad (-\infty<x<\infty)$$

分布函数的基本性质：

(1) $\lim\limits_{x\to-\infty} F(x)=0$， $\lim\limits_{x\to\infty} F(x)=1$

(2) 单调性，若 $x_1<x_2$，则 $F(x_1)\leqslant F(x_2)$

(3) $P\{a<\boldsymbol{X}\leqslant b\}=F(b)-F(a)$

定义6 离散型随机变量 如果随机变量 $\boldsymbol{X}$ 只能取有限个(或可列个[④])数值 $x_1, x_2, \ldots, x_n, \ldots$，就称 $\boldsymbol{X}$ 为离散型随机变量。在语言的统计处理中，一般仅用到离散型随机变量。

定义 7 分布密度 设 $\boldsymbol{X}$ 是一个离散型随机变量，它所有可能取的值为 $x_1, x_2, \ldots, x_n, \ldots$，$P\{\ \boldsymbol{X}=x_k\}=p_k\ (k=1,2,\ldots,n,\ldots)$，则可以用下面的表格来表达 $\boldsymbol{X}$ 取值的规律：

$\boldsymbol{X}$	x_1	x_2	…	x_n	…
概率	p_1	p_2	…	p_n	...

其中，$1\geqslant p_k\geqslant 0$ 且 $\sum\limits_k p_k=1$，则称这个表格所表示的函数为离散型随机变量的分布密度或概率函数，本书中，分布密度记作 $p(x)$。

定义 8 n 维随机变量 如果 $\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_n$ 是联系于同一组条件下的 n 个随机变量，则称($\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_n$)为 n 维随机变量。

三 随机变量的数字特征

随机变量是按照一定的规律(即分布)来取值的。在使用时，有时并不需要了解这个规律的全貌，而只需要知道它的某个侧面，这时，往往可以用一个或几个数字来描述这个侧面，这种数字部分地描述了随机变量的特性，称这种数字为随机变量的数字特征。

定义 9 数学期望 随机变量 $\boldsymbol{X}$ 的数学期望 $E(\boldsymbol{X})$是该变量取值的概率加权平均。数学期望简称期望，描述了随机变量的平均值。

若 $\boldsymbol{X}$ 为离散型随机变量，其可能的取值为 $x_k(k=1,2,\ldots)$且 $P\{\ \boldsymbol{X}=x_k\}=p_k$，则

[④] 如果数列 x_1，x_2，…，x_n，…可以和自然数建立一种映射关系,则称数列中含有可列个元素。

$$E(X)=\sum_k x_k p_k$$

定义 10 方差 随机变量$(X-E(X))^2$的数学期望称为随机变量 X 的方差，记作 $D(X)$或 Var(X)。方差描述了随机变量的取值距离其平均值(即期望值)的分散程度。即

$$D(X)=E(X-E(X))^2$$

方差定义中取平方的目的是避免正负偏差相互抵消，因为无论是正偏差大还是负偏差大，都是分散程度大。

方差描述了随机变量的取值距离其期望值的分散程度，但是它的单位是随机变量的单位的平方，有时，为了单位一致，还经常使用方差的算术平方根来描述随机变量的取值距离其期望值的分散程度。这个数字特征称为随机变量的标准差。

定义 11 标准差 随机变量 X 的标准差定义为随机变量 X 的方差的算术平方根，记作 $\sigma(X)$。即

$$\sigma(X)=\sqrt{D(X)}$$

数学期望和方差的性质：

(1) 当 c 为常数时，$E(cX)=cE(X)$

(2) $E(X\pm Y)=E(X)\pm E(Y)$

(3) 当 X，Y 相互独立时，$E(XY)=E(X)E(Y)$

(4) 当 c 为常数时，$D(cX)=c^2D(X)$

(5) 当 X，Y 相互独立时，$D(X\pm Y)=D(X)+D(Y)$

(6) 当 a 为常数时，$Ea=a$

(7) 当 a 为常数时，$Da=0$

(8) $D(X)=E(X^2)-(E(X))^2$

四 最大似然估计

定义 12 总体 研究对象的所有可能的观察结果称为总体。

定义 13 样本 从总体中抽取一部分样品，称为总体的一个样本。

要了解总体的统计规律，就应该对研究对象的所有观察结果进行研究，一一加以测定，这在实际中常常是不可能的或者不必要的。例如，某个灯泡厂需要了解灯泡的平均寿命，要实际测定它所生产的每一个灯泡的寿命是不可能的，而往往是从所有灯泡中随即选取一些灯泡进行测定。数理统计方法就是要通过研究样本来了解和判断总体的统计特性的科学方法。

设 X 为一个随机变量，并且 $P\{X=x\}=f(x,\theta_1,\theta_2,\ldots,\theta_k)$，参数$\theta_1$，$\theta_2$，…，$\theta_k$未知，$x_1$，$x_2$，…，$x_n$ 是随机变量取值的一个样本，则使函数

$$\prod_{i=1}^{n} f(x_i,\theta_1,\theta_2,...,\theta_k)$$

达到最大值的参数取值$\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k$可以被用做随机变量未知参数$\theta_1,\theta_2,\ldots,\theta_k$的估计值。这种参数估计方法一般称为**最大似然估计**方法[⑤]。

第二节 信息论基础

一 从最优编码谈起

首先看一个编码问题，假定有一个房间中有时没有人，有时甲在房间中，有时乙在房间中，有时甲乙都在房间中，房间状态服从下面的概率分布：

房间状态	房间没有人	甲在房间	乙在房间	甲乙均在房间
概率	0.5	0.125	0.125	0.25

某人受命监视房间，每五分钟记录一次房间状态，并经一个通讯设备将房间状态发送出去。我们知道，要把消息通过通讯设备发送出去，首先的问题是对消息进行编码。如果采用二进制进行编码，消息的编码将由 0 和 1 组成。一种可行的**定长编码**方案是：用 00 表示没有人在房间中，01 表示甲在房间中，10 表示乙在房间中，11 表示甲乙两人均在在房间中。按照这样的编码，发送一个消息所需要的码的长度为 2，平均发送一个消息需要$0.5\times 2+0.125\times 2+0.125\times 2+\ 0.25\times 2=2$个二进制位。

现在的问题是，房间的四种状态出现的概率并不相同，是否可以改进编码的方式，使得连续发送消息时，编码的长度最短，从而提高通讯设备的效率？可以根据下面的原则进行编码，给小概率消息赋以较长的编码，而给大概率消息赋以较短的编码。这样的话，连续发送多个消息时，可以保证发送的编码长度最短。然而这种编码方式，对于单个消息而言，编码长度是变化的，因而是一种**变长编码**。下表是符合上述原则的编码：

消息	编码
房间没有人	0
甲在房间	110
乙在房间	111
甲乙均在房间	10

利用这种方式编码，发送大概率消息(如房间没有人)仅仅需要一个二进制位，而发送小概率消息(如甲在房间)则需要三个二进制位。平均发送一个消息则需要$0.5\times 1+0.125\times 3+0.125\times 3+\ 0.25\times 2=1.75$个二进制位。可见第二种编码优于第一种编码。

一般而言，我们可以把消息看作一个取有限个值(如房间中没有人、甲在房间、乙在房间、甲乙均在房间)的随机变量 X，随机变量服从概率分布 P，如果消息 x 的分布密度为 $p(x)$，则给其分配一个长度为$\lceil -\log_2 p(x)\rceil$个二进制位的编码，这里$\lceil x\rceil$ 表示大于或等于 x 的最小整数。按照这样的编码方式，概率较大的消息的编码长度较小，概

[⑤] 最大似然估计方法的严格数学描述以及其它常用的参数估计方法请读者参阅有关数理统计的书籍，这里为了不引入更多的基础概念，对描述作了简化。

率较小的消息的编码长度教大，平均发送一个消息所需要的编码的长度为 $-\sum p(x)\log_2 p(x)$ 个二进制位。在信息论中，按照上述编码原则，如果发送一个消息所需要的编码的长度较大，则可以理解为消息所蕴涵的信息量较大，如果发送一个消息所需要的编码长度较小，则该消息所蕴涵的信息量较小，平均信息量即为发送一个消息的平均编码长度，可以用**熵**的概念来描述。

二 基本概念

定义 1 熵 设 X 是取有限个值的随机变量，它的分布密度为

$$p(x) = P\{X=x\}, \quad \text{且 } x\in \mathsf{X}$$

则，X 的熵定义为

$$H(X) = -\sum_{x\in \mathsf{X}} p(x)\log_a p(x) \qquad (2.1)$$

熵描述了随机变量的平均不确定性。一般也说，熵给出随机变量信息量的一种度量。在(2.1)式中，对数底 a 可以是任何正数，对数底 a 决定了熵的单位，若 a=2，则熵的单位称为**比特**(*bit*)。在没有特殊指明的情况下，以后本书均用比特作为熵的单位。为使上述定义在任何情况下都有意义，规定 $0\log_a 0$=0。

熵的基本性质：

(1) $H(X)\leqslant \log|\mathsf{X}|$，其中等号成立当且仅当 $p(x)=\dfrac{1}{|\mathsf{X}|}$，这里 $|\mathsf{X}|$ 表示集合 X 中的元素个数。该性质表明等概场具有最大熵。

(2) $H(X)\geqslant 0$，其中等号成立当且仅当对某个 i，$p(x_i)$=1，其余的 $p(x_k)$=0 $(k\neq i)$。这表明确定场(无随机性)的熵最小。

例子 1：假定有一种语言 P 有 6 个字母 p、t、k、a、i、u，字母的分布密度为：

P	*p*	*t*	*k*	*a*	*i*	*u*
概率	1/8	1/4	1/8	1/4	1/8	1/8

则随机变量 P 的熵为：

$$\begin{aligned} H(P) &= -\sum_{i\in\{p,t,k,a,i,u\}} p(i)\log p(i) \\ &= -[4\times\frac{1}{8}\log\frac{1}{8}+2\times\frac{1}{4}\log\frac{1}{4}] \\ &= 2\frac{1}{2}\ \text{bit} \end{aligned}$$

定义 2 联合熵 设 X、Y 是两个离散型随机变量，它们的联合分布密度为 $p(x,y)$，则 X,Y 的联合熵定义为：

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \quad (2.2)$$

定义 3 条件熵 设 X、Y 是两个离散型随机变量，它们的联合分布密度为 $p(x,y)$，则给定 X 时 Y 的条件熵定义为：

$$\begin{aligned} H(Y\mid X) &= -\sum_{x\in X} p(x)H(Y\mid X=x) \\ &= \sum_{x\in X} p(x)\left[-\sum_{y\in Y} p(y\mid x)\log p(y\mid x)\right] \\ &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y\mid x) \quad (2.3) \end{aligned}$$

联合熵和条件熵的关系可以用(2.4)式来描述，该关系一般也称为链式规则：

$H(X,Y) = H(X) + H(Y|X)$ (2.4)

信息量的大小随着消息长度的增加而增加，为了便于比较，一般使用熵率的概念，熵率一般也称为**字符熵**(per-letter entropy)或**词熵**(per-word entropy)。

定义 4: **熵率**(entropy rate)，对于长度为 n 的消息，熵率定义为：

$$H_{rate} = \frac{1}{n}H(X_{1n}) = -\frac{1}{n}\sum_{x_{1n}} p(x_{1n})\log p(x_{1n}) \quad (2.4)$$

这里 X_{1n} 表示随机变量序列 $X_1X_2\ldots X_n$，$p(x_{1n})$表示分布密度 $p(x_1\,x_2\ldots x_n)$。

可以把语言看作一系列语言单位构成的一个随机变量序列 $L = (X_1X_2\ldots X_n)$，则语言 L 的熵可以定义为这个随机变量序列的熵率：

$$H_{rate}(L) = \lim_{n\to\infty}\frac{1}{n}H(X_1, X_2,\ldots X_n) \quad (2.5)$$

根据链式规则，有：

$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$

可以推导出:

$H(X) - H(X|Y) = H(Y) - H(Y|X)$

$H(X)$和 $H(X|Y)$ 的差称为互信息，一般记作 $I(X;Y)$。$I(X;Y)$描述了包含在 X 中的有关 Y 的信息量，或包含在 Y 中的有关 X 的信息量。下面的图描述了互信息和熵之间的关系。

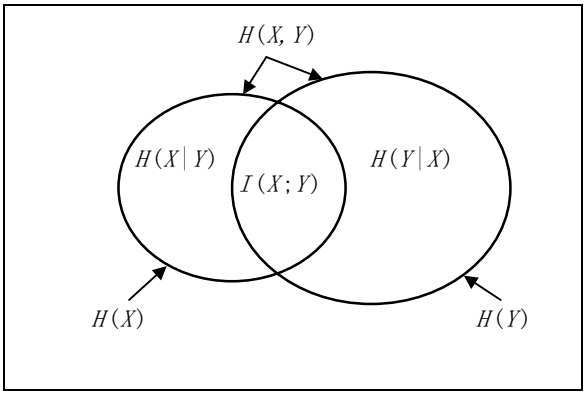

图 2.1 互信息和熵的关系

$$I(X;Y)=H(X)-H(X\mid Y)$$

$$=H(X)+H(Y)-H(X,Y)$$

$$=\sum_x p(x)\log\frac{1}{p(x)}+\sum_x p(y)\log\frac{1}{p(y)}+\sum_{x,y}p(x,y)\log p(x,y)$$

$$=\sum_{x,y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

故互信息定义如下:

定义 5: **互信息**(mutual information)，随机变量 *X*,*Y* 之间的互信息定义为:

$$I(X;Y)=\sum_{x,y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \qquad (2.6)$$

互信息的性质:

1) $I(X;Y)\geqslant 0$ 等号成立当且仅当 *X* 和 *Y* 相互独立。
2) $I(X;Y)=I(Y;X)$ 说明互信息是对称的。

(2.6)式给出了两个随机变量之间的互信息，在计算语言学中，更为常用的是两个具体事件之间的互信息，一般称之为**点式互信息**。

定义 6: **点间互信息**(pointwise mutual information)，事件 *x*, *y* 之间的互信息定义为:

$$I(x,y)=\log\frac{p(x,y)}{p(x)p(y)} \qquad (2.7)$$

一般而言，点间互信息为两个具体事件之间的相关程度提供了一种度量，即：

当 $I(x,y)>>0$ 时，*x* 和 *y* 高度相关。

当 $I(x,y)=0$ 时，*x* 和 *y* 高度相互独立。

当 $I(x,y)<<0$ 时，*x* 和 *y* 呈互补分布。

定义 7: **相对熵**(relative entropy)，设 *p*(*x*)、*q*(*x*)是随机变量 *X* 的两个不同的分布密度，则它们的相对熵定义为：

$$D(p\|q)=\sum_{x\in X}p(x)\log\frac{p(x)}{q(x)} \qquad (2.8)$$

规定 $0\log\frac{0}{q}=0$， $p\log\frac{p}{0}=\infty$。

相对熵一般也称谓 *Kullback-Leibler* 发散度或 *Kullback-Leibler* 距离。它提供了一

种度量同一个随机变量的不同分布的差异的方法。从信息论的角度看，如果一个随机变量 X 的分布密度是 $p(x)$，而人们却错误的使用了分布密度 $q(x)$，相对熵描述了因为错用分布密度而增加的信息量。

定义 8: **交叉熵**(cross entropy)，设随机变量 X 的分布密度为 $p(x)$，在很多情况下 $p(x)$是未知的，人们通常使用通过统计手段得到的 X 的近似分布 $q(x)$，则随机变量 X 的交叉熵定义为：

$$H(X,q)=\sum_{x\in X} p(x)\log q(x) \qquad (2.9)$$

三 噪音信道模型

利用信息论的基本概念，香农曾经给出了一个通信系统的抽象模型，该模型可以用图 2.2 来表示：

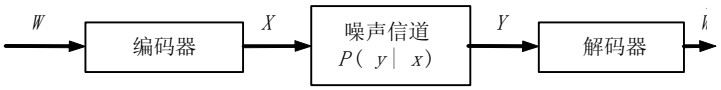

图 2.2 噪声信道模型

W 是欲经信道传输的消息，在传输之前，首先进行编码使其适于信道传输，编码后的消息为 X，由于信道噪声的存在，在信道末端，人们并不能精确接收到 X，而是接收到有噪声在内的编码 Y，信道概率 $p(y|x)$描述了编码 x 因噪声而变成 y 的概率，当接收方接到含有噪声的编码后，其任务就变为将 Y 解码，得到最为可能的消息 $\hat{W}$ 。作为通信系统而言，人们最为关心的是，如何将消息编码，以便消息在有噪声存在的情况下有效可靠地发送到接收方。噪音信道模型在通信以及编码领域得到了广泛的应用。

从 50 年代起，就有人试图利用这一模型解决语言问题，试图用一种量化的手段的对待语言。由于种种原因，最初的努力并未获得成功。70 年代，人们又一次将噪声信道模型引入语言处理领域。目前噪声信道模型已经在一系列语言问题的处理中获得了比较成功的应用。

在利用噪声信道处理语言问题时，人们并不关心编码问题，而更多关心的是，在有噪声存在的情况下，如何解码将输出还原为信道输入。因此，人们常常面对的是图 2.3 中的信道模型。

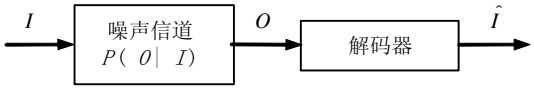

图 2.3 语言处理应用中的噪声信道模型

这个问题可以用公式(2.10)来描述。即在观察到 O 的情况下，人们可以通过计算找到一个最为可能产生 O 的输入 $\hat{I}$ 来作为原信道输入 I 的估值。

$$\hat{I} = \arg\max_{I} p(I \mid O) \qquad (2.10)$$

利用 Bayes 定理，可以得到公式(2.11)

$$\hat{I} = \arg\max_{I} \frac{p(I)p(O \mid I)}{p(O)} \qquad (2.11)$$

由于 $p(O)$是一个常数，故有：

$$\hat{I} = \arg\max_{I} p(I)p(O \mid I) \qquad (2.12)$$

在(2.12)式中出现了两个概率分布(1) $p(I)$，一般称为语言模型；(2) $p(O|I)$，一般称为信道模型。

作为一个例子，这里看一下如何把噪声信道模型应用与语言的翻译问题，假定人们希望把一篇英语文本翻译成汉语文本。对于这个任务，人们可以假定信道的输入是汉语文本，由于噪声的干扰，在信道末端人们看到的是英语文本，翻译问题就变成如何根据信道输出的英语文本恢复为信道输入端的汉语文本的问题。利用信道模型，人们为翻译问题找出了一个整齐的数学描述。除了翻译问题，信道模型在一系列其它语言处理问题上得到了较好的应用，例如光学字符识别(OCR)，词性自动标注和语音识别等等。

第三节 隐马尔可夫模型

隐马尔可夫过程(Hidden Markov process)是一种十分重要的随机过程，从目前计算语言学的研究来看，基于隐马尔可夫过程的语言模型在语言信息处理领域中有着十分广泛的用途。隐马尔可夫模型的基本理论形成于 60 年代末期和 70 年代初期。70 年代中期，CMU 的 J.K.Baker 以及 IBM 的 F.Jelinek 等人把该模型应用于语音识别问题，语音识别的研究和开发表明,基于隐马尔可夫模型的方法是一种十分成功的方法。后来，借鉴隐马尔可夫模型在语音识别领域中的成功经验,人们也在语言信息处理的其他领域采用这种模型，目前比较成功的例子包括：词类的自动标注、汉语文本的音字转换等等。

隐马尔可夫过程可以看作是马尔可夫过程的一种扩充，为了引入隐马尔可夫模型，对马尔可夫过程有所了解是必要的。隐马尔可夫模型的英语译文是 Hidden Markov Model,一般简写为 HMM。

一 马尔可夫过程

在日常生活中，我们会碰到很多系统，这些系统的运作和时间有关，随着时间的推移,系统表现为一组状态的变化，在某个具体的时刻对这个系统进行观察，会发现这个系统处在某个状态。例如，天气的变化，假如可以用阴天(雨雪天气)、晴天和多云来概括天气所有可能的变化，那么，随着时间的推移，天气总是在这三种情况之间进行变化，但在某个时刻进行观察，天气又只能是晴天、阴天和多云中的一种情况。而晴天、阴天和多云正是天气这个

系统的三种状态。天气的这种变化可以用图 3.1 来描述。

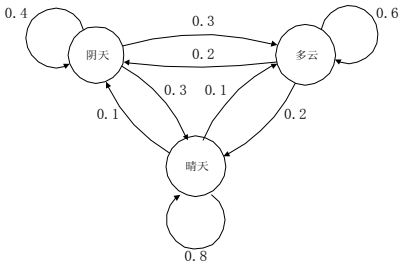

图 3.1 天气变化的马尔可夫模型

在某个具体时刻(例如今天),无法准确预测下一个时刻(例如明天)天气情况,今天如果是晴天，并不能保证明天一定也是晴天，也可能是阴天或多云。如果对大量的历史数据进行分析，找出历史上所有的晴天，在观察其后的一天的天气情况，就能得到后一天天气情况的概率，应用前面介绍条件概率的定义，可以表示为：

P(晴天|晴天)　P(阴天|晴天)　P(多云|晴天)

天气随着时间的变化从一个状态按照一定概率转换到另外一个状态，恰恰就是马尔可夫过程的一个具体的例子。完整的马尔可夫模型可用下面的数学描述进行定义

假设 S 是一个由有限个状态组成的集合$\{1, 2, 3, \ldots, n-1, n\}$, q_t 表示在时刻 t 时系统所处的状态。 严格来说 q_t 的情况将会依赖于它的所有历史情况 $q_{t-1}, q_{t-2}, \ldots, q_1$，但一般只考虑一种比较特殊的情况，即一阶马尔可夫过程，q_t 的取值只和 q_{t-1} 的取值有关，即下面的等式满足时的情况。

$$P(q_t = j \mid q_{t-1} = i, q_{t-2} = k, \ldots) = P(q_t = j \mid q_{t-1} = i)$$

令：

$$a_{ij} = P(q_t = j \mid q_{t-1} = i) \quad 1 \le i, j \le n$$

则对于所有的 i,j 有下面的关系成立：

$$a_{ij} \ge 0$$

$$\sum_{j=1}^{n} a_{ij} = 1$$

则一阶马尔可夫过程可以用一个二元组(S, A)描述。S 是状态的集合，而 A 是所有状态转移概率组成的一个 n 行 n 列的矩阵，其中每一个元素 a_{ij} 为从状态 i 转移到状态 j 的概率。

例如，对上述图 3.1 中描述的天气变化的马尔可夫过程，如果用 1 表示阴天，2 表示多云，3 表示晴天的话，则该模型可以形式化地描述为：

$$M = (S, A),$$

其中，

$$S = \{1, 2, 3\}$$

$$A=\begin{bmatrix}0.4 & 0.3 & 0.3\\ 0.2 & 0.6 & 0.2\\ 0.1 & 0.1 & 0.8\end{bmatrix}$$

二 隐马尔可夫过程

在马尔可夫过程中，每一个状态都唯一对应一个事件，例如对于上述关于天气的马尔可夫模型，如果连续观察 8 天的天气变化情况，假定为(晴天，晴天，晴天，阴天，阴天，晴天，多云，晴天)，我们也一定知道一个唯一的状态转换序列(3, 3, 3, 1, 1, 3, 2, 3)。如果把晴天称为状态 3 的输出，阴天称为状态 1 的输出，多云称为状态 2 的输出。根据观察到的输出序列就可以决定模型中的状态转换序列。如果每一个状态并不对应一个唯一的输出，而是按照一定的概率分布输出多个可能的观察值，就形成了一个隐马尔可夫过程。

关于隐马尔可夫模型，L.Rabiner 曾经给出过一个经典的例子：在一个房间中，假定有 N 个坛子，每个坛子中都装有各种颜色的小球，并且假定总共有 M 中不同颜色的小球。一个精灵在房间中首先随机地选择一个坛子，再从这个坛子中随机选择一个小球，并把小球的颜色报告给房间外面的人员记录下来作为观察值。该精灵然后再把该球放回目前选择的坛子，以目前的坛子为条件再随机选择一个坛子，从中随机选择一个小球，并报告小球的颜色，然后放回小球，如此继续…，随着时间的推移，房间外的人会得到由这个过程产生的一个小球颜色的序列。

在这个过程中，如果把每一个坛子对应与一个状态，可以用状态转移概率矩阵来描述坛子的选择过程。并且每个状态可能按照特定的概率分布输出不同颜色的小球。与天气变化的马尔可夫过程不同，观察人员仅仅根据看到的小球颜色的序列，并不能确定坛子的选择过程，也就是说，根据观察序列，人们不能确定状态转移序列，状态转移过程被隐藏起来了。所以这种随机过程一般称为隐马尔可夫过程。

一个完整的隐马尔可夫过程包含下面五个要素：

1. 一组状态的集合 $S=\{1,2,3,\ldots,N\}$。尽管观察人员不能直接观察到状态之间的转换，但在隐马尔可夫过程的实际应用中，状态往往有确定的含义，在坛子和小球的试验中，状态和坛子之间有一种对应关系，状态 n 对应坛子 n。
2. 一组输出(或观察符号)的集合 $V=\{v_1,v_2,v_3,\ldots,v_M\}$。观察符号对应着模型的物理输出，在坛子和小球的试验中，观察符号对应的是小球的颜色，如果共有 3 中不同颜色的小球，则观察符号集中共含有 3 个元素，即 V = { 红， 白，蓝 }。
3. 状态转移概率矩阵 $A=[a_{ij}]$。是一个 N 行 N 列的矩阵，含义和一阶马尔可夫过程中的状态转移概率矩阵相同。其中

$$a_{ij}=P(q_{t+1}=j\mid q_t=i),1\le i,j\le N$$

4. 观察符号的概率分布 $B=\{b_j(k)\}$。$b_j(k)$ 表示在状态 j 时输出观察符号 v_k 的概率。则有：

$$b_j(k)=P(v_k\mid j),1\le k\le M,1\le j\le N$$

5. 初始状态概率分布 $\pi=\{\pi_i\}$。表示时刻 1 选择某个状态的概率。则有：

$$\pi_i=P(q_1=i)$$

因此，隐马尔可夫模型可以表示为一个五元组。如果用λ 来表示隐马尔可夫模型，则

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

一般也简写为

$$\lambda = (A, B, \pi)$$

可以把隐马尔可夫模型看为一个观察值的生成装置，按照一定的步骤，隐马尔可夫模型可以生成下面的观察序列：

$$O = (o_1 o_2 o_3 \ldots o_T)$$

其中 o_i 表示时刻 i 的观察值。

1．按照初始状态概率分布 π 选择一个初始状态(即时刻 1 时模型所出的状态) $q_1 = i$。
2．令 $t = 1$。
3．按照状态 i 观察符号的概率分布 $b_i(k)$选择一个观察值 $o_t = v_k$ 。
4．按照状态转移概率分布 a_{ij} 选择一个后继状态 $q_{t+1} = j$ 。
5．若 $t < T$，令 $t = t + 1$，并且转移到算法第 3 步继续执行，否则结束。

在上述模型的定义中，实际上假定了状态转移仅仅依赖与前一个状态的选择，这样的隐马尔可夫过程称为一阶隐马尔可夫过程，同马尔可夫过程类似，状态选择可能不仅仅依赖于前一个状态的选择，而依赖于前 k 个状态的选择，这样的隐马尔可夫过程称为 k 阶隐马尔可夫过程。

可见，隐马尔可夫模型是一个双重随机过程，其中一重随机过程不能直接观察到，通过状态转移概率矩阵描述。另一重随机过程输出可以观察的观察符号，这由输出概率来定义。

当把隐马尔可夫模型用于实际问题时，有三个问题需要解决：

1) 给定隐马尔可夫模型$\lambda = (A, B, \pi)$ 和一个观察序列 $O = (o_1 o_2 o_3 \ldots o_T)$，如何有效地计算出观察序列的概率，即 $P(O|\lambda)$? 这是一个计算问题。
2) 给定隐马尔可夫模型$\lambda = (A, B, \pi)$ 和一个观察序列 $O = (o_1 o_2 o_3 \ldots o_T)$，如何寻找一个状态转换序列 $q = (q_1\ q_2\ q_3\ \ldots\ q_T)$，该状态转换序列最有可能产生上述观察序列(或在某种意义下，最好地解释了上述观察序列)？这是一个估计问题。
3) 在模型参数未知或不准确的情况下，如何根据观察序列 $O = (o_1 o_2 o_3 \ldots o_T)$求得模型参数或调整模型参数，即如何确定一组模型参数，使得 $P\ (O|\lambda)$最大？这是一个训练问题。

三 向前算法和向后算法

对隐马尔可夫模型而言，状态转换序列是隐藏的，一个观察序列可能由任何一种状态转换序列产生。因此要计算一个观察序列的概率值，就必须考虑所有可能的状态转换序列，图 3.2 表示了产生观察序列 $O = (o_1 o_2 o_3 \ldots o_T)$的所有可能的状态转换序列。

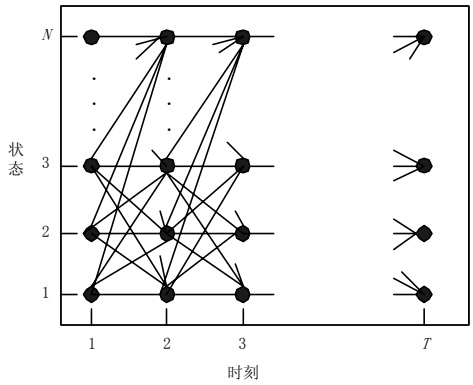

图 3.2 观察序列 $O = (o_1 o_2 o_3 \dots o_T)$的所有可能的状态转换序列

对于某一个状态转换序列 q = (q_1 q_2 q_3 … q_T)产生观察序列 $O = (o_1 o_2 o_3 \dots o_T)$的概率可以通过下面的公式计算：

$$P (O|q, \lambda) = b_{q1}(o_1)\, b_{q2}(o_2)\, b_{q3}(o_3) \dots b_{qT}(o_T)$$

而状态转换序列 q = (q_1 q_2 q_3 … q_T)的概率可以通过下面的公式计算：

$$P (q \mid \lambda) = \pi_{q1}\, a_{q1q2}\, a_{q2q3} \dots a_{qT\text{-}1qT}$$

则 O 和 q 的联合概率为：

$$P (O, q \mid \lambda) = P (O|q, \lambda)\, P (q \mid \lambda)$$

因为要考虑所有的状态转换序列，所以

$$P(O \mid \lambda) = \sum_{q} P(O, q \mid \lambda) = \sum_{q1, q2, \dots qT} \pi_{q1} b_{q1}(o_1) a_{q1q2} b_{q2}(o_2) \dots a_{qT-1qT} b_{qT}(o_T) \quad (1)$$

该公式直观含义为，模型在时刻 1 以概率π_{q1}处于 q_1 状态，并按照概率 $b_{q1}(o_1)$生成输出 o_1，模型按照概率 a_{q1q2}在时刻 2 转换为状态 q_2，并且 q_2 状态按照概率 $b_{q2}(o_2)$产生输出 o_2，如此继续，直到模型按照概率 $a_{qT\text{-}1qT}$在时刻 T 转换为状态 q_T，并且按照概率 $b_{qT}(o_T)$产生输出 o_T 为止。

从理论上可以通过上述公式计算一个观察序列的值，然而实际上严格按照这个公式进行计算是不现实的。该计算需要进行$(2T-1)N^T$次乘法运算，N^T-1 次加法运算，如果有 5 个状态，即 N=5，若 T=100，则会需要大约 10^{72} 次运算。因此必须寻找更为有效的计算方法。

考虑在给定模型λ，时刻 t，处在状态 i，并且部分观察序列为 $o_1 o_2 o_3 \dots o_t$ 的概率，假定该概率为$\alpha_t(i)$，则：

$$\alpha_t(i) = P(o_1 o_2 o_3 \dots o_t, q_t = i \mid \lambda)$$

$\alpha_t(i)$一般称为向前变量。可以用下面的方法以归纳方式计算向前变量：

1. 初始化

$$\alpha_1(i) = \pi_i\, b_i(o_1), \quad 1 \le i \le N$$

2．若已知 $\alpha_t(i)$，$1 \le i \le N$，则可以通过下面的公式归纳计算得到$\alpha_{t+1}(i)$：

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(o_{t+1}), \quad 1 \le t \le T-1, 1 \le j \le N$$

归纳计算可以通过图 3.3 得到解释：

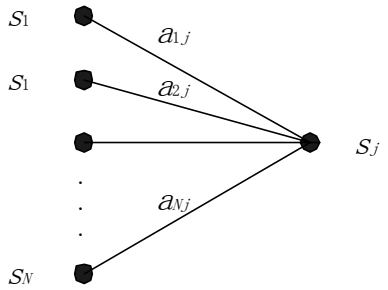

图 3.3 到达 j 状态的所有转换

在时刻 t，模型可能处在任何一个状态，它们都可以在时刻 $t+1$ 转换到 j 状态，并且产生输出 o_{t+1}。因为$\alpha_t(i)$是在时刻 t 观察到 $o_1 o_2 o_3 \ldots o_t$，并且处在状态 i 的概率。那么乘积$\alpha_t(i)a_{ij}$即为观察到 $o_1 o_2 o_3 \ldots o_t$，并在 t 时刻处在状态 i，经由状态 i 而在 $t+1$ 时刻到达状态 j 的概率，因为从任何一个状态都可以到达 j 状态，则求和就得到了观察到 $o_1 o_2 o_3 \ldots o_t$，并且在 $t+1$ 时刻处在状态 j 的概率，乘以状态 j 输出 o_{t+1} 的概率 $b_j(o_{t+1})$，就得到了在时刻 t+1，观察到 $o_1 o_2 o_3 \ldots o_t o_{t+1}$，并且处在状态 j 的概率，也就是$\alpha_{t+1}(j)$。

这样的归纳计算一直可以这样计算向前计算下去，直到求出所有的$\alpha_T(i)$，也就是观察到整个观察序列，并且在 T 时刻处在状态 i 的概率。因为 T 时刻可能处在任何一个状态，所以，该计算过程终止于下面的计算。

3．结束

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

由于这个过程是从时刻 1 开始，逐步向前归纳计算，所以这个过程一般称为向前过程。按照这个过程计算，计算效率大为提高，总共需要 $N(N+1)(T-1)+N$ 次乘法和 $N(N-1)(T-1)$次加法。同样例如 $N=5, T=100$ 时，大约需要 5000 次计算。这比直接按照(1)大大节约了时间。

按照同样的道理，也可以采用一种向后的过程计算$P(O|\lambda)$，为此首先定义向后变量 $\beta_t(i)$。

$\beta_t(i) = P(o_{t+1} o_{t+2} \ldots o_T \mid q_t = i, \lambda)$

其含义为，给定模型λ，时刻 t 处在状态 i，观察到 $o_{t+1} o_{t+2} \ldots o_T$ 的概率。同向前过程类似，可以按照下面的方式归纳计算出所有的$\beta_t(i)$，并进而计算出 $P(O|\lambda)$。

1．初始化

$\beta_T(i) = 1, \quad 1 \le i \le N$

2．归纳计算

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$

3．终止

$$P(O \mid \lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)$$

向后过程的计算复杂度和向前过程是相同的，均为 $O(N^2T)$。

四 韦特比算法

隐马尔可夫模型的第二个问题是估计出一个能最好解释观察序列的状态转换序列，显然可以通过枚举所有的状态转换序列，并对每一个状态转换序列 q 计算 $P(O, \mathrm{q} \mid \lambda)$，使 $P(O, \mathrm{q} \mid \lambda)$ 取最大值的 q^*就是能最好解释观察序列的状态转换序列，即：

$$q^* = \arg\max_q P(O, q \mid \lambda)$$

但同样，这并不是一个有效的计算方法。为了更为有效地计算最佳状态转换序列，同向前过程类似，首先定义如下的变量：

$$\delta_t(i) = \max_{q1,q2,\ldots qt-1} P(q_1 q_2 \ldots q_{t-1}, q_t = i, o_1 o_2 \ldots o_t \mid \lambda)$$

$\delta_t(i)$的含义是模型在时刻 t 处于状态 i，观察到 $o_1 o_2 o_3 \ldots o_t$ 的最佳状态转换序列的概率。和向前变量类似，$\delta_t(i)$可以通过下面的公式归纳计算。

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] b_j(o_{t+1})$$

这样，可以一直归纳计算出$\delta_T(i)$，即模型在时刻 T 处于状态 i，观察到 $o_1 o_2 o_3 \ldots o_T$ 的最佳状态序列的概率，那么，模型在时刻 T，观察到 $o_1 o_2 o_3 \ldots o_T$ 的最佳状态序列的概率，可以通过下面的公式得到：

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

可以上述归纳过程中设立 T 个数组 $\psi_1(N), \psi_2(N), \ldots \psi_T(N)$，从而记住最佳路径的构成，过程如下：

1. 初始化

   $\delta_1(i) = \pi_i b_i(o_1)$，$1 \le i \le N$

   $\psi_1(i) = 0$

2. 归纳计算

   $$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(o_t), 2 \le i \le N, 1 \le j \le N$$

   $$\psi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}]$$

3. 归纳终止

   $$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

   $$q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)]$$

4. 回溯最佳状态转换序列

   $$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T\text{-}1, T\text{-}2, \ldots, 1$$

上述算法一般称为韦特比算法，$\delta_t(i)$一般也称为韦特比变量，可以发现韦特比算法和向前算法是相似的，是一种有效地寻求最佳状态转换序列的方法。

以上的算法需要做乘法，韦特比算法一般在实现时，一般要把模型参数取对数，这样乘

法运算变为加法，算法如下：

0．预处理

$\tilde{\pi}_i = \log(\pi_i)$，$1 \le i \le N$

$\tilde{b}_i(o_t) = \log[b_i(o_t)]$，$1 \le i \le N$，$1 \le t \le T$

$\tilde{a}_{ij} = \log(a_{ij})$，$1 \le i,j \le N$

1．初始化

$\tilde{\delta}_1(i) = \log(\delta_1(i)) = \tilde{\pi}_i + \tilde{b}_i(o_1)$，$1 \le i \le N$

$\psi_1(i) = 0$

2．归纳计算

$\tilde{\delta}_t(i) = \log(\delta_t(i)) = \max_{1 \le i \le N}[\tilde{\delta}_{t-1}(i) + a_{ij}] + \tilde{b}_j(o_t)$

$\psi_t(j) = \max_{1 \le i \le N}[\tilde{\delta}_{t-1}(i) + a_{ij}]$，$2 \le t \le T, 1 \le j \le N$

3．终止

$\tilde{P}^* = \max_{1 \le i \le N}[\tilde{\delta}_T(i)]$

$q_T^* = \arg\max_{1 \le i \le N}[\tilde{\delta}_T(i)]$

4．回朔最佳状态转换序列

$q_t^* = \psi_{t+1}(q_{t+1}^*)$，　$t = T\text{-}1, T\text{-}2, \ldots, 1$

改进后的算法时间复杂度为 $O(N^2T)$，但不再有乘法运算。但增加了预处理的开销，预处理只需要进行一次，对于多数系统而言，预处理开销是微不足道的。

五 Baum-Welch 算法

3.2 节提出的和隐马尔可夫模型相关的三个问题中，前两个问题均假定模型已知，而第三个问题是已知观察序列，求最佳模型的问题，这是这三个问题中最为困难的一个。对于一个隐马尔可夫模型而言，关键有三组参数，状态转移概率矩阵 A、状态输出概率 B 以及初始状态概率分布π。因此该问题是如何选择$\lambda = (A, B, \pi)$，使得在该模型下，已知的观察序列的概率 $P(O|\lambda)$为最大？把观察序列视为训练样本，A、B、π 视为未知参数，问题就是一个参数估计的问题。

Baum-Welch 算法是一种重复渐进的估计过程，通过逐步叠代寻找最为可能的模型参数。为了描述这个过程，首先定义给定模型λ和观察序列 O，在时刻 t 处在状态 i，时刻 t+1 处在状态 j 的概率 $\xi_t(i,j)$，即：

$\xi_t(i,j) = P(q_t = i, q_{t+1} = j \mid O, \lambda)$

根据前面定义的向前变量和向后变量，$\xi_t(i,j)$可以进一步写成：

$$\xi_t(i,j) = \frac{P(q_t = i, q_{t+1} = j, O \mid \lambda)}{P(O \mid \lambda)} = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)}$$

公式的分子部分计算可以表示为图 3.4。

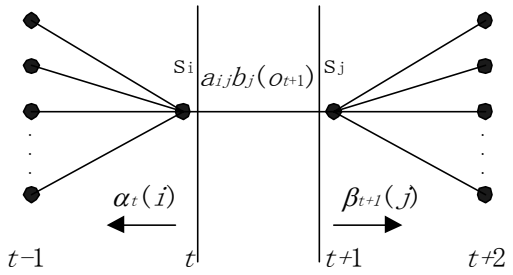

图 3.4 计算 $P(q_t = i, q_{t+1} = j, O \mid \lambda)$

如果在给定模型以及观察序列的情况下，t 时刻处在状态 i 的概率为$\gamma_t(i)$，则有：

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j)$$

则$\gamma_1(i)$，$\gamma_2(i)$，…，$\gamma_{T-1}(i)$之和(即 $\sum_{t=1}^{T-1} \gamma_t(i)$)可以解释为观察序列 O 中从状态 i 出发的转换的期望次数。而 $\sum_{t=1}^{T-1} \xi_t(i, j)$ 可以解释为观察序列 O 中从状态 i 到状态 j 的转换的期望次数。

利用上述结论，关于π, A, B，一种合理的估计方法如下：

$\overline{\pi}_i = \gamma_1(i)$ ，即在 $t = 1$ 时处在状态 i 的期望次数(或期望频率)。

$\overline{a}_{ij} = \dfrac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$ ，也就是从状态 i 到状态 j 的转换的期望次数除以从状态 i 出发的转换的期望次数。

$\overline{b}_j(k) = \dfrac{\sum_{t=1}^{T} \gamma_t(j) \times \delta(o_t, v_k)}{\sum_{t=1}^{T} \gamma_t(j)}$，其中，当 $o_t = v_k$ 时，$\delta(o_t,v_k)=1$，当 $o_t \neq v_k$ 时，$\delta(o_t,v_k)=0$，这样在该等式中，等号右面分子部分的含义即为在状态 j 观察到 v_k 的期望次数，而分母部分则表示处在状态 j 的期望次数。

利用上述三个估算公式，可以通过叠代求精的方式估计参数π, A, B。首先为π, A, B 选择一组初始值，作为最初的模型参数。但初始值的选择要满足隐马尔可夫模型定义对参数关系的限制，即有：

$$\sum_{i=1}^{N}\pi_i = 1$$

$$\sum_{j=1}^{N}a_{ij} = 1,\ 1 \le i \le N$$

$$\sum_{k=1}^{M}b_j(k) = 1,\ 1 \le j \le N$$

然后利用上述三个估算公式，计算得到一组新的模型参数，这组新的模型参数又可以作为一个新的估算过程的开始点，再次进行估算，如此反复进行，直到参数值收敛。Baum 等人证明要么估算值 $\bar{\lambda}$ 和估算前的参数值λ相等，要么估算值 $\bar{\lambda}$ 比估算前的参数值λ更好的解释了观察序列 O。参数最终的收敛点并不一定是一个全局最优值，但一定是一个局部最优值。Baum-Welch 算法是一类称为 EM (Estimation-Maximisation:估计-最大化)算法的一个例子，这类算法均可保证收敛于一个局部最优值。

参考文献

① 同济大学数学教研室主编，《工程数学—概率论》，高等教育出版社，1982

② L.Rabiner, Fundamentals of Speech Recognition, Prentice Hall, 1993

③ 孟庆生，《信息论》，西安交通大学出版社，1986

④ C. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999

⑤ E. Charniak, Statistical Language Learning, MIT Press, 1993

⑥ B.Krenn, C. Samuelsson, The Linguist's Guide to Statistics Don't Panic, 1997, http://www.cogsci.ed.ac.uk/~chrisbr/linguists-guide-to-statistics.ps

⑦ M. Oakes, Statistics for Corpus Linguistics, Edinburgh University Press, 1998