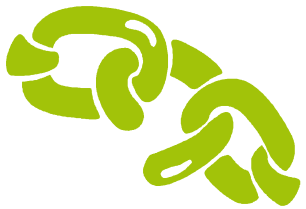
第11章
信息理论方法及其应用

本章主要内容



本章主要内容

11.3 最小交叉熵原理

- 11.3.1 最小交叉熵原理
- 11.3.2 交叉熵的性质
- 11.3.3 最小交叉熵推断的性质
- 11.3.4 交叉熵法

11.4 信息理论方法的应用

- 11.4.1 信息论在分子生物学中应用
- 11.4.2 最大熵谱估计和最小交叉熵谱估计
- 11.4.3 最大熵建模及其在自然语言处理中应用
- 11.4.4 最大熵原理在经济学中的应用
- 11.4.5 信息理论方法应用展望

11.1 信源熵的估计

本节主要内容：

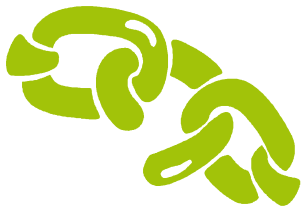
1. 离散信源序列熵的估计

2. 连续信源熵的估计

11.1 熵估计

- 参数估计

  给定一个信源概率密度的参数模型，首先从可能的密度函数空间中搜索最可能的密度函数，其次计算最可能密度函数所对应的熵。



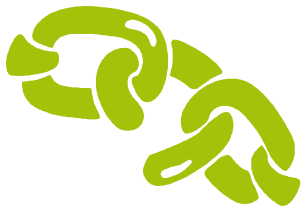
11.1 熵估计（续）

- 非参数估计

  （1）标准的最大似然或“插入”法，就是先估计信源符号的概率或概率密度，然后再将估计的概率代入熵的计算公式中来计算熵；

  （2）对于离散信源，可以利用无损信源压缩编码算法进行熵估计，通常使用通用压缩编码；

  （3）利用其他熵估计算法。



11.1.1 离散信源序列熵的估计

- 插入熵估计
- 通用信源压缩编码熵估计
- 模板匹配熵估计

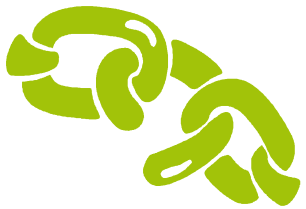
插入熵估计（一）——离散无记忆信源熵估计

设一个离散无记忆信源具有未知的概率分布 $P=\{p(i),i\in A\}$，符号集 $A=\{1,2,\cdots,q\}$，$H=H(P)$ 为信源的熵，现用给定信源序列作为训练序列 $x_1,x_2,\cdots,x_n$，对信源熵进行插入法估计。

首先利用训练序列估计信源符号的概率，表示为 $\hat{p}_n(i),i=1,2,\cdots q$ 其中 n 为训练样本数，通常估计值依赖于样本数。符号的概率采用下式估计：

$$\hat{p}_n(i)=\frac{\sum_{k=1}^{n}I_i(x_k)}{n}$$

其中 $I_i(x)=\begin{cases}1 & x=i\\ 0 & x\neq i\end{cases}$，$\hat{p}_n(i)$ 称为经验分布。



插入熵估计（一）——离散无记忆信源熵估计

可以证明 $\hat{p}_n(i)$ $p(i)$是 $E[\hat{p}_n(i)] = p(i)$的最大似然估计，且是无偏的，即

将概率的估计值代入信源熵的的公式，得信源熵的估计值：

$$\hat{H}_{MLE}(p_n) = -\sum_{i=1}^{q} \hat{p}_n(i) \log \hat{p}_n(i)$$

虽然概率的估计是无偏的，但熵的估计却不是无偏的。根据熵的上凸性有

$$E[\hat{H}_{MLE}(p_n)] = E[-\sum_{i=1}^{q} \hat{p}_n(i) \log \hat{p}_n(i)] \leq -\sum_{i=1}^{q} E[\hat{p}_n(i)] \log[E(\hat{p}_n(i))] = H(P)$$

可见，用插入估计得到的熵值的平均要比实际熵值低，是实际熵的欠估计。所以在进行熵估计时，要做适当的修正，以保证熵的估计具有较小的偏差。

插入熵估计（一）——离散无记忆信源熵估计

- Miller等（1954年）提出，修正后的熵按下式计算

$$\hat{H}_{MM} = \hat{H}_{MLE}(p_n) + (\hat{m} - 1)/(2n)$$

其中，$\hat{m}$ 为估计的信源符号集的大小，n为训练样本数，熵的单位是奈特。

- 还有一种偏差修正法，称为jackknife法，是统计学中对估计器的偏差和方差进行估计的有效方法，这种方法特别用于标准方法难于应用的场合。熵估计的公式为

$$\hat{H}_{JK} = n\hat{H}_{MLE}(p_n) - \frac{n-1}{n}\sum_{i=1}^{n}\hat{H}(-i)$$

其中，$\hat{H}(-i)$ 表示用样本$(x_1, \cdots x_{i-1}, x_{i+1}, \cdots, x_n)$ 进行插入估计所得的熵。

插入熵估计（一）——离散无记忆信源熵估计

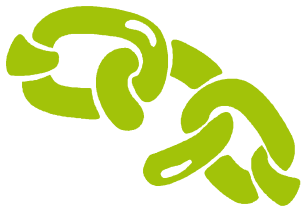
例11.1

一个二元离散无记忆信源，符号0和1的概率分别为1/4和3/4，长度为32的训练序列为1 1 1 0 0 0 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 0 1 1 1 0 1 1 1 1;

（1）求信源熵的最大似然插入估计；

（2）利用Miller修正法估计信源熵；

（3）利用jackknife修正法估计信源熵。



插入熵估计（一）——离散无记忆信源熵估计

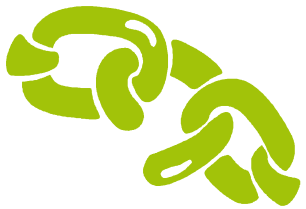
解：

信源熵 $H(X) = -(1/4)\times\log_2(1/4) - (3/4)\times\log_2(3/4) = 0.8113$ 比特

（1）信源符号概率的ML估计为： $\hat{p}_0 = 7/32$ ， $\hat{p}_1 = 25/32$;

信源熵的最大似然插入估计为

$\hat{H}_{MLE} = -(7/32)\times\log_2(7/32) - (25/32)\times\log_2(25/32) = 0.7519$ 比特；



插入熵估计（一）——离散无记忆信源熵估计

（2）利用Miller修正估计信源熵为

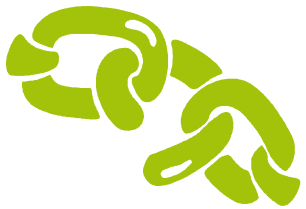
$$\hat{H}_{MM} = \hat{H}_{MLE} + (\hat{m}-1)/2N = 0.7519 + 1/(2\times 32)\times \log_2 e = 0.7744$$ 比特；

（3）$\hat{H}_{-i}(x_i = 0) = -(6/31)\times\log_2(6/31) - (25/31)\times\log_2(25/31) = 0.7088$ 比特；

$\hat{H}_{-i}(x_i = 1) = -(7/31)\times\log_2(7/31) - (24/31)\times\log_2(24/31) = 0.7706$ 比特；

利用jackknife修正估计信源熵为

$\hat{H}_{JK} = 32\times\hat{H}_{MLE} - (31/32)[7\times\hat{H}_{-i}(x_i = 0) + 25\times\hat{H}_{-i}(x_i = 1)]$ 比特。



插入熵估计（二）——一阶马氏链熵估计

设信源是一个J状态的一阶马氏链，其中状态集合$S=\{1,2,\cdots,J\}$信源序列$(x_0,x_1,\cdots,x_n)$为训练序列，现对信源的熵进行插入估计。

定义示性函数

$$I_{ij}(x,y)=\begin{cases}1 & x=i,y=j\\ 0 & \text{其它}\end{cases}$$

在训练序列中状态（i，j）同时发生次数的估计为

$$\hat{m}_n(i,j)=\sum_{k=1}^{n}I_{ij}(x_{k-1}=i,x_k=j)$$

状态i发生次数的估计为

$$\hat{m}_n(i)=\sum_{j=1}^{J}\sum_{k=1}^{n}I_{ij}(x_{k-1}=i,x_k=j)$$

插入熵估计（二）——一阶马氏链熵估计

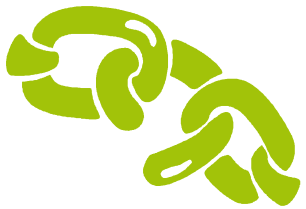
状态转移概率的估计：

$$\hat{p}_n(j \mid i) = \hat{m}_n(i,j) / \hat{m}_n(i)$$

平稳概率的估计：

$$\hat{\pi}_n(i) = \hat{m}_n(i) / n$$

注意，如果训练样本数量不够，会出现某些状态的概率统计值为零的情况。当$\hat{m}_n(i)=0$，必有$\hat{p}_n(j|i)=\hat{\pi}_n(i)=0$；当$\hat{m}_n(i,j)=0$，必有$\hat{p}_n(j|i)=0$。为避免出现有些状态观察不到的情况，应该进行多次独立的观察。

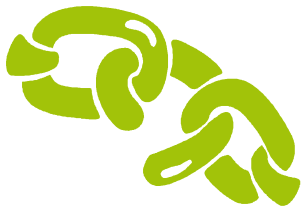


插入熵估计（二）——一阶马氏链熵估计

信源熵的估计：

$$\hat{H}(p_n) = -\sum_{i=1}^{J} \hat{\pi}_n(i) \sum_{j=1}^{J} \log \hat{p}_n(j \mid i)$$

与独立信源熵的估计类似，熵的估计也不是无偏的，用插入估计得到的熵值的平均要比实际熵值低。这也可根据熵的上凸性推出。



插入熵估计（三）——N次扩展源的熵估计

$$\hat{p}_n(x_1,x_2,\cdots,x_N)=\frac{n(x_1,x_2,\cdots,x_N)}{n} \qquad (11.17)$$

其中，$n(x_1,x_2,\cdots,x_N)$ 为序列中状态 $(x_1,x_2,\cdots,x_N)$ 的个数

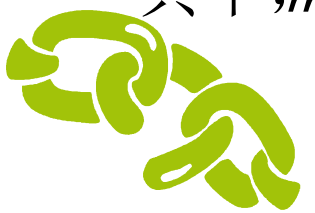
N次扩展源的熵估计为

$$\hat{H}_{MLE}(\mathbf{X}_1^N)=-\sum_{\hat{p}(x_1,x_2,\cdots x_N)}\hat{p}_n(x_1,x_2,\cdots,x_N)\log\hat{p}_n(x_1,x_2,\cdots,x_N) \qquad (11.18)$$

修正后的熵为：

$$\hat{H}(\boldsymbol{X}_1^N)=\hat{H}_{MLE}(\boldsymbol{X}_1^N)+(\hat{m}-1)/2n+(1/12n^2)(1-\sum_{P(x_1,\cdots;x_N)}1/p(x_1,\cdots;x_N)+o(n^{-3}) \qquad (11.19)$$

其中，$\hat{m}$ 为 $\hat{p}_n(x_1,x_2,\cdots,x_N)>0$ 的个数。

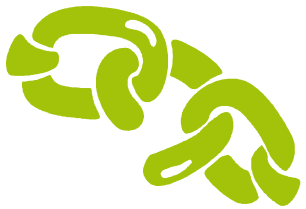


插入熵估计（四）——高阶有记忆马氏源熵估计

- 根据 $H(\mathbf{X}_1^N)=H(\mathbf{X}_1^{N-1})+H(X_N|\mathbf{X}_1^{N-1})$ ，
  所以一个N-1阶马氏源熵的估计为：

$$\hat{H}(X_N|\mathbf{X}_1^{N-1})=\hat{H}(\mathbf{X}_1^N)-\hat{H}(\mathbf{X}_1^{N-1}) \qquad (11.20)$$

- 研究表明，插入法熵估计随L的增大，熵估计精度增加。对于低阶马氏源，该方法能得到较精确的熵估计，但对于记忆长度较大的信源序列，数据长度往往不够，不能得到精确的估计结果。



通用信源压缩编码熵估计

根据香农第一定理，无损信源编码码率的下界是信源的熵，即通过理想的信源编码后，编码器码率可以压缩到信源的熵。因此可以计算无损编码后的压缩率（输出文件比特长度与输入文件比特长度的比）R，而$R \geq H(X)$，如果采用性能理想的无损信源编码算法，编码压缩率R可以作为信源熵的估计。

$$R \approx H(X) \qquad (11.21)$$

当然，压缩性能越好，估计值越接近信源的熵。

通用信源压缩编码熵估计

- 有很多通用信源压缩编码算法可用做熵估计，该方法相对于插入法的优点是，不依赖信源的模型，也不假设任何特殊的信源结构，主要考虑的是算法的收敛速度。

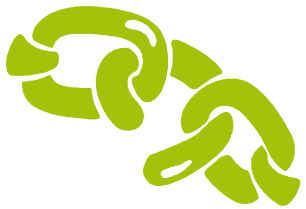
- 有几种常用的信源压缩编码算法可用做熵估计，例如：LZ77系列（自适应模板匹配编码）、GZIP算法（LZ77 + Huffman编码）、BZIP（基于Burrows-Wheeler变换 + Huffman编码）CWT（上下文树加权 + 算术编码）等。

模板匹配熵估计

对于记忆逐渐消失的信源，设k和n为任意选择的正整数，给定观察，信源熵可由下式估计：

$$\hat{H}(n,k)=\frac{k\log n}{\sum_{i=1}^{k}L_i(n)} \qquad (11.22)$$

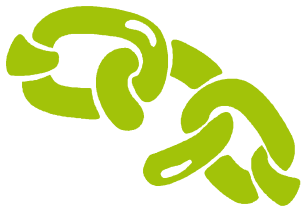
从估计过程可以看到，随着i的连续增加，匹配长度的计算在长度为n的滑动窗内进行。所以这种熵估计方法也称为滑动窗熵估计。



模板匹配熵估计

可采用如下方法进行估计值的矫正：

对给定信源序列进行信源符号概率的最大似然估计，再以估计的概率产生与信源序列长度相同的信源序列复制品，对这些复制序列做窗长不同的滑动窗熵估计。因为复制序列的熵是可以精确计算的，所以用复制序列的熵减去估计熵就得到估计偏差。在所有复制品序列中对这种偏差进行平均就得到平均估计偏差。修正后的熵估计值是滑动窗熵估计加偏差值。



模板匹配熵估计

例11.2：

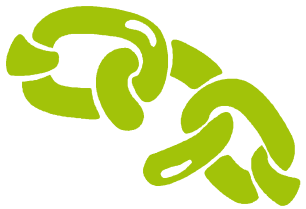
对于例11.1的二元离散无记忆信源，用相同的训练序列，用滑动窗熵估计法估计信源的熵，滑动窗口长度n=8，不要求对估计结果进行矫正。

解

求得 $L_i(n)$，$i=1,\cdots,24$，分别为3，3，4，4，4，2，1，0，2，2，7，6，5，4，5，4，4，4，4，4，3，3，2，1；

熵的估计为

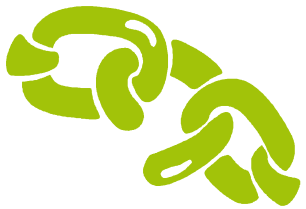
$$\hat{H}(8,24)=\frac{24\times\log_2 8}{\sum_{i=1}^{k}L_i(n)}=72/81=0.8889 \text{ 比特}$$



11.1.2 连续信源熵的估计

连续信源熵的估计主要有以下两种方法：

- 用一个参数集合近似实际的概率密度，而这个参数集合所对应的熵是已知的；
- 基于对概率密度的直接估计，例如利用直方图等，然后计算信源的熵，实际上就是插入估计。



11.2 最大熵原理

本节主要内容：

1. 最大熵原理的描述
2. 熵集中定理
3. 几种重要的最大熵分布

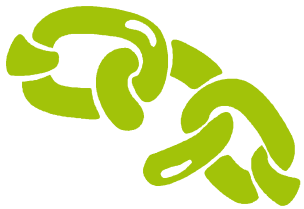
最大熵原理的描述

最大熵原理：

- 在寻找满足某些约束的概率分布时，选择满足这些约束具有最大熵的概率分布。

利用最大熵原理的依据：

- 主观依据
- 客观依据



最大熵原理的描述——离散信源

设离散信源熵：

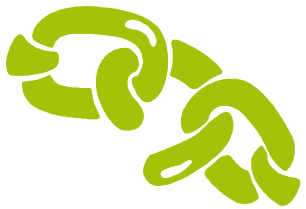
$$H = -\sum_{i=1}^{n} p_i \log p_i \qquad (11.23)$$

满足约束

$$\sum_{i=1}^{n} p_i = 1 \qquad (11.24)$$

$$\sum_{i}^{n} p_i g_r(x_i) = a_i \ , \quad r = 1,2,\cdots,m \qquad (11.25)$$

其中，（11.24）是概率的归一化约束；$g_r(x_i)$ 是已知函数，（11.25）通常是概率矩的描述（包括均值或方差）或其他特征的平均值，a_i 是已知常数，在解决实际问题时这些常数可能由训练数据得到。



最大熵原理的描述——离散信源

定理11.1 （离散最大熵分布定理）
满足（11.24）和（11.25）的约束使（11.23）达到最大值的概率分布为：

$$p_i = Z^{-1}\exp[-\sum_{r=1}^{m}\lambda_r g_r(x_i)] \quad ,i=1,2,\cdots,n \qquad (11.26)$$

其中，
$$Z=\sum_{i=1}^{n}\exp[-\sum_{r=1}^{m}\lambda_r g_r(x_i)] \qquad (11.27)$$

最大熵为
$$H_{\max}=\ln Z+\sum_{r=1}^{m}\lambda_r a_r \qquad (11.28)$$
（奈特）

参数λ_r，$r=1,2,\cdots,m$，由下式确定：

$$a_r = Z^{-1}\sum_{i=1}^{n} g_r(x_i)\prod_{k=1}^{m}\alpha_k^{g_k(x_i)} \quad ,r=1,2,\cdots,m \qquad (11.29)$$

其中，$\alpha_r=\exp(-\lambda_r)$ 。

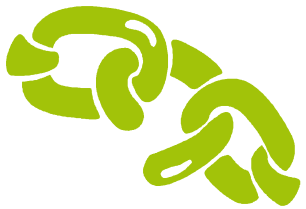
最大熵原理的描述——离散信源

例11.3 做1000次抛掷骰子的试验，求抛掷点数的平均值。

解

由于抛掷次数很多，所以各点出现的频率近似等于出现的概率。假定在每次抛掷后，骰子6个面中的每一个面朝上的概率都相同，即为1/6。这里我们利用了“不充分理由原理”，因为除知道骰子有6个面外，我们没有其他任何别的信息。

抛掷点数的平均值：$m=(1+2+3+4+5+6)/6=3.5$



最大熵原理的描述——离散信源

例11.4 做1000次抛掷骰子的试验后得知抛掷点数的平均值为4.5，求骰子各面朝上的概率分布。

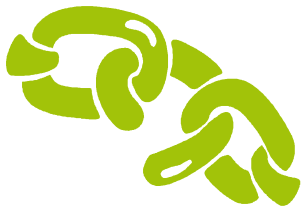
解

很明显，骰子的各面朝上的概率是不均匀的。除概率的归一性外，我们知道的信息仅有平均值，这对于确定6个面的概率是不完整的信息，必须利用最大熵原理。

根据题意，平均值的约束写为

$$p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + 6p_6 = 4.5$$

结合（11.25），有m=1g,$(x_i) = i$ a_1,= 4.5 ，且只有待定常数 Z、α_1 ，



最大熵原理的描述——离散信源

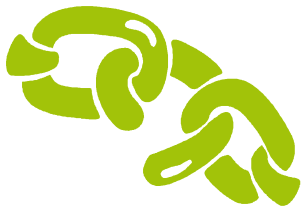
由（11.29），得

$$4.5=\frac{\alpha_1+2\alpha_1^2+3\alpha_1^3+4\alpha_1^4+5\alpha_1^5+6\alpha_1^6}{\alpha_1+\alpha_1^2+\alpha_1^3+\alpha_1^4+\alpha_1^5+\alpha_1^6}$$

$\alpha_1=1.44925$；代入（11.26），得

所求概率分布为：$p_i=\frac{\alpha_1^i}{\alpha_1+\alpha_1^2+\alpha_1^3+\alpha_1^4+\alpha_1^5+\alpha_1^6}=\frac{1.44925^i}{26.6637}$

$$(p_1,p_2,p_3,p_4,p_5,p_6)=(0.0543,0.0788,0.1142,0.1654,0.2398,0.3475)$$



最大熵原理的描述——离散信源

例11.5 求例11.4概率分布所对应的熵。

解

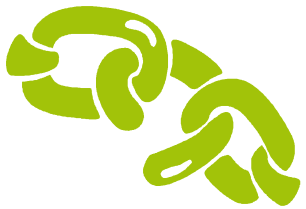
$$H_{\max} = H(0.0543, 0.0788, 0.1142, 0.1654, 0.2398, 0.3475)$$

$$= 1.6135\text{奈特} = 2.3279\text{比特}$$

由（11.27）得 $Z = 26.6637$；

也可利用（11.28），得

$$H_{\max} = \log Z + \lambda_1 a_1 = 3.2833 - (\ln 1.44925) \times 4.5 = 1.6135 \text{ 奈特}$$



最大熵原理的描述——连续情况

信源的熵

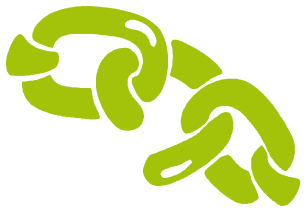
$$h = -\int_a^b p(x)\ln p(x)dx \qquad (11.32)$$

满足

$$\int_a^b p(x)dx = 1 \qquad (11.33)$$

$$\int_a^b p(x)g_r(x)dx = a_r \ , r = 1,2,\cdots,m \qquad (11.34)$$

与推导离散情况类似，可以得到以下结果。



最大熵原理的描述——连续情况

定理11.2 （连续最大熵分布定理）

满足（11.33）和（11.34）的约束使（11.32）达到最大值的概率密度为：

$$p(x) = Z^{-1} \exp[-\sum_{r=1}^{m} \lambda_r g_r(x)] \qquad (11.35)$$

其中，

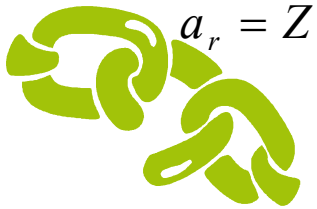
$$Z = \int_a^b exp[-\sum_{r=1}^{m} \lambda_r g_r(x)]dx \qquad (11.36)$$

最大熵等于

$$h_{\max} = \ln Z + \sum_{r=1}^{m} \lambda_r a_r \qquad (11.37)$$

参数λ_r，$r = 1,2,\cdots,m$，由下式确定：

$$a_r = Z^{-1} \int_a^b g_r(x) \exp[-\sum_{k=1}^{m} \lambda_k g_k(x)]dx \text{，} r = 1,2,\cdots,m \qquad (11.38)$$



最大熵原理的描述——连续情况

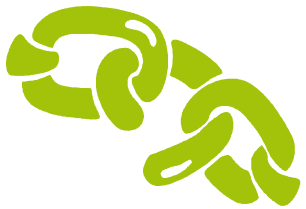
例11.6　连续信源X的取值区间为（a,b），求达到最大熵的X的分布度$p(x)$和相应的最大熵$h$max 。

解

因为只有归一化约束，由（11.36），得Z=b-a，

由（11.35），所求分布密度 $p(x)=1/(b-a)$，

最大熵为 $h_{\max}=\log(b-a)$。



熵集中定理

定理11.3（熵集中定理）满足约束（11.25）的一组概率 $p_1,\cdots,p_n$ 所产生的熵在如下范围：

$$H_{\max}-\Delta H \le H(p_1,\cdots,p_n)) \le H_{\max} \qquad (11.42)$$

其中

$$2N\Delta H = \chi_k^2(1-F) \qquad (11.43)$$

$H_{\max}$ 为在约束（11.24）、（11.25）下，（11.23）的最大值。（11.43）的含义是，当N足够大时，$2N\Delta H$ 渐近为维数为k（=n-m-1，n为信源符号数，m为约束方程个数），置信度为1-F的 χ^2 分布的值。通常，在很高的置信度的条件下，ΔH的也值很小。

熵集中定理

例11.7 求例11.4置信度95%和99.99%时信源熵的范围。

解

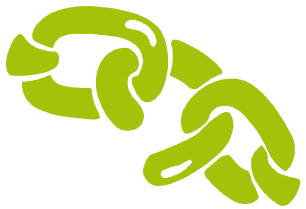
根据题意，$2N\Delta H$ 为自由度6-1-1=4的 χ^2 分布，χ^2 查表，

(1) 在置信度95%条件下，得 $2N\Delta H = 9.488$，$\Delta H = 0.00474$ ， 信源熵的范围：

$$1.609 \leq H \leq 1.614 \text{（奈特）；}$$

(2) 在置信度99.99%条件下，得 $\Delta H = \chi_4^2(0.9999)(2N)^{-1} = 0.012$ ，信源熵的范围：

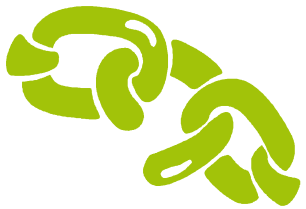
$$1.602 \leq H \leq 1.614 \text{（奈特）}$$



熵集中定理

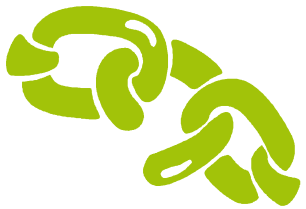
- 结论：

  在提供的信息不完全的情况下，最大熵分布不仅以最多的实现方式实现，而且随着试验次数的增多绝大多数可能的分布的熵都接近最大熵。当次数N时，除具有最大熵的分布外，其他满足约束的分布都是非典型的，出现的概率几乎为零。可以认为，具有最大熵的分布是所有满足给定约束的概率分布的代表；最大熵法是一种保险的策略，它能防止我们预测出数据没有提供的虚假结果。



11.2.3几种重要的最大熵分布

- 满足均值约束的连续最大熵分布是指数分布
- 满足均值约束的离散最大熵分布是几何分布
- 满足均值和均方值约束的最大熵分布是高斯分布
- 满足几何平均值约束的最大熵分布是幂律分布



满足均值约束的连续最大熵分布是指数分布

例11.8　连续信源X的取值区间为[0，∞），均值E(X)=μ，求达到最大熵的X的分布密度和相应的最大熵。

解

根据（11.35）有$p(x)=Z^{-1}e^{-\lambda_1 x}$，

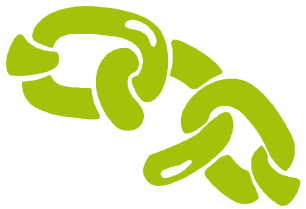
其中，　$Z=\int_0^{\infty}e^{-\lambda_1 x}dx=1/\lambda_1$　；

根据约束条件有 $\mu=\int_0^{\infty}\lambda_1 x e^{-\lambda_1 x}dx=1/\lambda_1$

所求分布密度为 $p(x)=\frac{1}{\mu}e^{-x/\mu}$　$,x\geq 0$　（11.44）

根据（11.37），得最大熵为

$$h_{\max}=\log(e\mu)$$　（11.45）



满足均值约束的连续最大熵分布是指数分布

例11.9　离散信源X的取值为{Ei，i=1，2，……}，满足约束，$\sum_i p_i E_i = E$ ，求达到最大熵的X的概率分布pi和相应的最大熵$H（X）$。

解

根据（11.26），得$p_i = Z^{-1} e^{-\lambda_1 E_i}$

其中，$Z = \sum e^{-\lambda_1 E_i}$。令 $\lambda_1 = 1/(k_B T)$，其中 kB 为波耳兹曼常数，T为绝对温度，那么 $p_i = Z^{-1} e^{-E_i / k_B T}$　　（11.46）

这就是物理学中的波耳兹曼分布，也称波耳兹曼—吉布斯（Boltzmann-Gibbs）分布，也是指数分布，是平衡统计力学的基本定律。它指出，一个能量为Ei的某特殊状态的概率满足波耳兹曼分布。一个重要的特例就是，Ei=i，这时离散分布称为几何分布。

满足几何平均值约束的最大熵分布是幂律分布

例11.11

设连续信源X的取值为正实数，概率密度为p(x)，且满足约束，

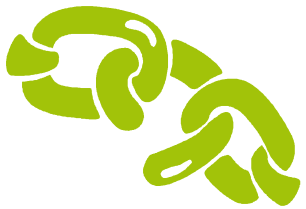
$$\int_1^{\infty} p(x)\ln x dx = \mu \quad , \quad \mu > 0 \qquad (11.47)$$

求具有最大熵的分布密度。

解

根据（11.35）有

$$p(x) = \frac{e^{-\lambda_1 \ln x}}{\int_1^{\infty} e^{-\lambda_1 \ln x} dx} = (\lambda_1 - 1)x^{-\lambda_1}$$



满足几何平均值约束的最大熵分布是幂律分布

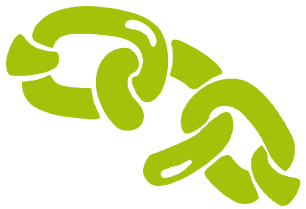
根据约束条件（11.47）有 $\mu=(\lambda_1-1)\int_1^{\infty}x^{-\lambda_1}\ln x dx=1/(\lambda_1-1)$

得，$\lambda_1=1+1/\mu$ ，所以 $p(x)=\frac{1}{\mu}x^{-(\mu+1)/\mu}$ （11.48）

（11.48）所示的分布满足幂律分布。所谓幂律是指一个随机变量的概率密度是一个幂函数，即

$$p(x)\propto x^{-\alpha} \qquad (11.49)$$

其中，α 为正数。



11.3 最小交叉熵原理

本节主要内容：

1. 最小交叉熵原理
2. 交叉熵的性质
3. 最小交叉熵推断的性质
4. 交叉熵法

11.3.1 最小交叉熵原理

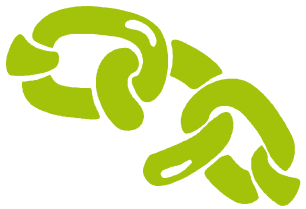
- 信息散度在信号处理领域常称为交叉熵。
- 对离散和连续信源，定义在同一概率空间的两概率测度P和Q的交叉熵分别定义为：

- 离散情况：$$D(P\|Q)=-\sum_{i=1}^{n}p_i\log\frac{p_i}{q_i}\quad(11.50)$$
- 连续情况：$$D(P\|Q)=-\int_a^b p(x)\log\frac{p(x)}{q(x)}dx\quad(11.51)$$

11.3.1 最小交叉熵原理

- 交叉熵表示概率分布P和Q之间的“距离”，也表示信源从概率P变化到概率Q所需要的信息量，所以Q称为先验概率，而P称为后验概率。

- 在很多情况下，可能存在关于概率分布的先验知识，此时可以用最小交叉熵原理推断后验概率分布。最小交叉熵原理就是，当推断一个具有先验分布Q的随机变量的概率分布P时，选择在满足X的已知约束下使交叉熵最小的概率分布。下面是最小交叉熵分布定理。



11.3.1 最小交叉熵原理

- **定理11.4**（离散最小交叉熵分布定理） 满足（11.25）的约束使（11.50）达到最小值的后验概率分布为：

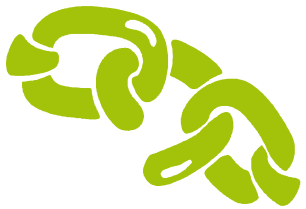
$$p_i = Z^{-1} q_i \exp[-\sum_{r=1}^{m} \lambda_r g_r(x_i)]，\ i = 1,2,\cdots,n \qquad (11.52)$$

其中，$Z = \sum_{i=1}^{n} q_i \exp[-\sum_{r=1}^{m} \lambda_r g_r(x_i)]$ （11.53）

参数 λ_r，$r = 1,2,\cdots,m$，由下式确定：

$$a_r = Z^{-1} \sum_{i=1}^{n} q_i g_r(x_i) \prod_{k=1}^{m} \alpha_k^{g_k(x_i)}，r = 1,2,\cdots,m \qquad (11.54)$$

其中，$\alpha_r = \exp(-\lambda_r)$ 。



11.3.1 最小交叉熵原理

- **定理11.5** （连续最小交叉熵分布定理） 满足（11.34）的约束使（11.51）达到最小值的后验概率密度为：

$$p(x)=Z^{-1}q(x)\exp[-\sum_{r=1}^{m}\lambda_r g_r(x)] \qquad (11.55)$$

其中，$Z=\int q(x)exp[-\sum_{r=1}^{m}\lambda_r g_r(x)]dx$ （11.56）

λ_r $r=1,2,\cdots,m$ 参数，由下式确定：

$$a_r=Z^{-1}\int_a^b g_r(x)q(x)\exp[-\sum_{k=1}^{m}\lambda_k g_k(x)]dx \quad r=1,2,\cdots,m \qquad (11.57)$$

以上两定理的证明与最大熵分布定理的证明类似，此处略。

11.3.1 最小交叉熵原理

- 设满足约束条件（11.25）或（11.34）的概率分布的集合为 P，那么最小交叉熵原理可以描述为：

$$p = \arg\min_{p' \in P} D(p' \| q) = \arg H(p,q) \qquad (11.58)$$

其中，$H(p,q) = \min_{p' \in P} D(p' \| q)$ （11.59）

满足（11.58）的概率密度称为最小交叉熵的解。

- 通过比较可知，最大熵原理是最小交叉熵原理的特殊情况，此时的先验概率是均匀分布。
- 也可以说，最小交叉熵原理是最大熵原理的推广。

11.3.1 最小交叉熵原理

- 例11.12　设先验概率为N维独立高斯分布随机矢量：

$$q(\boldsymbol{x})=\prod_{k}(2\pi\sigma_k^2)^{1/2}\exp[-(x_k-m_k)^2/2\sigma_k^2]$$

N维随机矢量满足约束：

$$\int_{\boldsymbol{x}} x_i\, p(\boldsymbol{x})\, d\boldsymbol{x} = \mu_i \quad , \quad i=1,\cdots,N$$

$$\int_{\boldsymbol{x}} (x_i-\mu_i)^2\, p(\boldsymbol{x})\, d\boldsymbol{x} = \nu_i \quad , \quad i=1,\cdots,N$$

求使交叉熵最小的后验分布密度。

11.3.1 最小交叉熵原理

- 解

根据（11.55），得

$$p(x)=\exp\{-(\lambda_0+\sum_{r=1}^{N}[\lambda_{1r}x_r+\lambda_{2r}(x_r-\mu_r)^2]\}\prod_k(2\pi\sigma_k^2)^{1/2}\exp[-(x_k-m_k)^2/2\sigma_k^2]$$

$$=\prod_k(2\pi\sigma_k^2)^{1/2}\exp\{-\lambda_0-\sum_{i=1}^{N}[\lambda_{1i}x_i+\lambda_{2i}(x_i-\mu_i)^2+(x_i-m_i)^2/2\sigma_i^2]\}$$

可以看到，仍然是独立的高斯分布，再根据给定的约束条件得 $p(x)=\prod_i(2\pi\nu_i)^{1/2}\exp[-(x_i-\mu_i)^2/2\nu_i]$

可见，所求的密度仍然是高斯分布，不过均值和方差被新的约束值代替。

11.3.2 交叉熵的性质

交叉熵的性质总结如下：

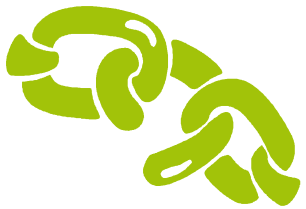
1）非负性

$D(p\|q)\geq 0$，仅当 $P=Q$ 时等式成立。（11.60）

2）下凸性

$D(p\|q)$ 是p的下凸函数，确切地说，已知概率分布（离散或连续）p_1，p_2，q 和正数 α，且 $0<\alpha<1$，$p=\alpha p_1+(1-\alpha)p_2$，有

$$D(p\|q)\geq \alpha D(p_1\|q)+(1-\alpha)D(p_2\|q) \qquad (11.61)$$



11.3.2 交叉熵的性质

3）可加性

这类似于熵的可加性，就是独立联合分布的交叉熵等于各独立合分布的交叉熵的和。确切地说，已知概率分布 $p(\boldsymbol{x})=p_1(x_1)\cdots p_N(x_N)$，$q(\boldsymbol{x})=q_1(x_1)\cdots q_N(x_N)$ 有

$$D(p\|q)=\sum_{i=1}^{N}D(p_i\|q_i) \tag{11.62}$$

4）坐标变换下的不变性

在离散情况转化成置换下的不变性。确切地说，设概率密度 $p(\boldsymbol{x})$，$q(\boldsymbol{x})$ 有变换 $\boldsymbol{y}=f(\boldsymbol{x})$，有

$$D(p(\boldsymbol{x})\|q(\boldsymbol{x}))=D(p(\boldsymbol{y})\|q(\boldsymbol{y})) \tag{11.63}$$

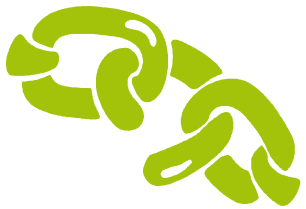
11.3.2 交叉熵的性质

5）勾股性质

定理11.6 设$q(x)$为先验概率密度，$p(x)$、$r(x)$为满足（11.34）约束的后验概率密度，且 $D(p\|q)=H(p,q)$，那么

$$D(r\|q)=D(r\|p)+D(p\|q) \qquad (11.64)$$

（11.64）式类似于勾股定理。如果把满足约束的概率分布看成一个子空间C，则r，p都在C内，而q在C外。p与q的距离长度最短，相当于连接p、q的矢量和C垂直，从C中任何一点r到q的距离的平方等于r到p距离的平方加上p到q的距离的平方。



11.3.2 交叉熵的性质

- 证：

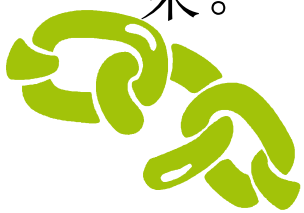
$$D(r\|q)=\int r(x)\log\frac{r(x)}{q(x)}dx=\int r(x)\log\frac{r(x)}{p(x)}dx+\int r(x)\log\frac{p(x)}{q(x)}dx$$

$$\overset{a}{=}\int r(x)\log\frac{r(x)}{p(x)}dx+\int r(x)[-(\lambda_0+\sum_{r=1}^{m}\lambda_r g_r(x)]dx$$

$$\overset{b}{=}\int r(x)\log\frac{r(x)}{p(x)}dx+\int p(x)\log\frac{p(x)}{q(x)}dx$$

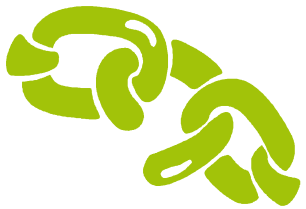
$$=D(r\|p)+D(p\|q)$$

- 其中，a：$p(x)$满足（11.55），b：$r(x)$和 $p(x)$ 都满足约束。



11.3.3 最小交叉熵推断的性质

- J.E.Shore等提出4个公理化条件：（1）惟一性；（2）在坐标系变换下不变性；（3）系统独立性；（4）子集独立性。提出这些公理的指导性原则就是，如果问题可以用多于一种的方法解决，那么结果要一致。他们证明了，在给定先验概率和约束条件下，由最小交叉熵原理所推出的后验概率，即最小交叉熵的解是惟一满足上述公理条件下的结果。所以，最小交叉熵原理是惟一的一致推断系统，所有其它的推断系统都将导致矛盾。



11.3.3 最小交叉熵推断的性质

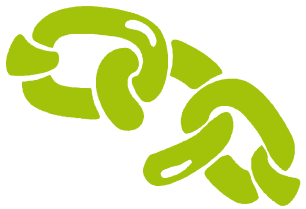
- 下面对最小交叉熵推断满足以上4个公理化条件做具体说明：

  （1）惟一性

  最小交叉熵的解是惟一的。这可以从是p的下凸函数推出。

  （2）坐标变换下的不变性

  在两个不同的坐标系中的最小交叉熵解具有坐标变换关系。也就是说，先求最小交叉熵的解，再变换得到的后验概率密度和先变换再求最小交叉熵解结果是相同的。



11.3.3 最小交叉熵推断的性质

（3）系统独立性

系统独立性含义是，对于多维系统，对各维分别进行最小交叉熵推断实现和用联合概率用最小交叉熵直接推断实现得到相同的后验概率。下面以2维分布为例来说明。设先验概率密度 $q(x,y)=q_1(x)q_2(y)$，后验概率密度 $p(x,y)$ 及其边际概率密度 $p_1(x)$ 和 $p_2(y)$，现通过 $H(p,q)=H(p,q_1q_2)$ 估计 $p(x,y)$。

因为 $D(p\|q_1q_2)-D(p_1p_2\|q_1q_2)=\iint p(x,y)\log\dfrac{p(x,y)}{p_1(x)p_2(y)}dxdy$

$$=D(p\|p_1p_2)\geq 0$$

所以，求 $\arg H(p,q)$ 相当于求 $\arg H(p_1p_2,q_1q_2)$，而

$$D(p_1p_2\|q_1q_2)=D(p_1\|q_1)+D(p_2\|q_2)$$

因此相当于分别求 $\arg H(p_2,q_2)$ 和 $\arg H(p_1,q_1)$。

11.3.3 最小交叉熵推断的性质

（4）子集的独立性

子集独立性含义是，用最小交叉熵推断后验概率时，可以用整个概率密度计算来实现也可以将概率密度划分成若干状态下条件概率密度分别实现。

设先验与后验概率密度分别划分成若干不相交区间 $\{s_i\}$，且 $q(x)=\sum_i m_i q_i(x\mid u_i)$，$p(x)=\sum_i n_i p_i(x\mid v_i)$

$$D(p\|q)=\sum_i\int_{s_i} n_i p(x\mid u_i)\log\frac{n_i p_i(x\mid u_i)}{m_i q_i(x\mid v_i)}dx$$

$$=\sum_i n_i D(p_i\|q_i)+\sum_i n_i\log\frac{n_i}{m_i} \qquad (11.65)$$

所以，求 $\arg H(p,q)$ 相当于对所有 i 求 $\arg H(p_i\|q_i)$

11.3.4 交叉熵法

在信息处理中，往往要求一个概率密度接近另一个目标概率密度，而目标概率密度的参数未知的。这样，将式（11.51）$p(x)$作为目标概率密度，$q(x)$为含有参数的概率密度，写成 $q(x,u)$，可以通过改变u使交叉熵最小。由于

$$D(P\|Q)=\int p(x)\log\frac{p(x)}{q(x)}dx=\int p(x)\log p(x)dx-\int p(x)\log q(x)dx$$

因此，使 $D(P\|Q)$ 最小，相当与使上式第二项最大，即

$$u^*=\arg\max_u\int p(x)\log q(x,u)dx \qquad (11.66)$$

上式就是当前被称为交叉熵法的理论依据。

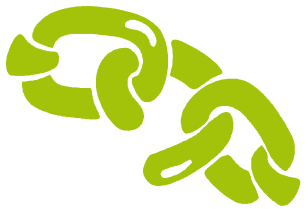
11.3.4 交叉熵法

交叉熵法是一个迭代算法，包含两步：

1）应用一个动态参数集产生随机数据样本；

2）应用数据样本本身对控制随机数据产生的参数进行更新，以进一步改进数据样本。

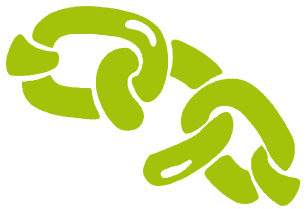
交叉熵法首先由Rubinstein在1997年提出，用做估计稀有事件概率的自适应算法，后来作为解决很多优化问题，特别是NP难题的通用而有效的工具。交叉熵法已经成功应用到很多复杂的优化问题，例如邮递员旅行问题、二次分配问题、最大割集问题等。



11.4 信息理论方法的应用

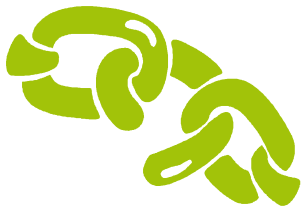
本节主要内容：

1. DNA序列的熵估计和压缩
2. 最大熵谱估计和最小交叉熵谱估计
3. 最大熵建模及其在自然语言处理中的应用
4. 最大熵原理在经济学中的应用
5. 信息理论方法应用展望



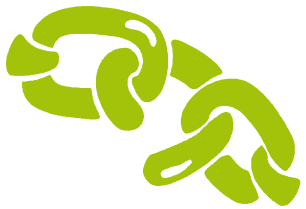
11.4.1 DNA序列的熵估计和压缩

- 近些年来，生物学领域的研究取得很大进展，很多物种完整DNA（脱氧核糖核酸）序列已经被发现，关于基因的一系列重要问题已成为研究的热点，这里就包括DNA序列的处理以及有效存储和传送问题，DNA序列熵的估计和压缩是用来解决这类问题的方法之一。



11.4.1 DNA序列的熵估计和压缩

我们知道，细胞核中的DNA是生物的遗传物质。它的单体是核苷酸，由一个碱基，脱氧核糖分子（S）和磷酸分子（P）构成。碱基有四种：腺嘌呤（A）、鸟嘌呤（G）、胞嘧啶（C）和胸腺嘧啶（T）。因此共四种核苷酸，简记为A，G，C，T，所以DNA序列可以表示字母表为$\{A，T，G，C\}$的符号串。在DNA的编码区，能够对蛋白质进行编码的序列称为外显子（exon），而不能对蛋白质进行编码的序列称为内含子（intron）。对各种遗传序列的测试表明，内含子和外显子的熵存在很大的差别。

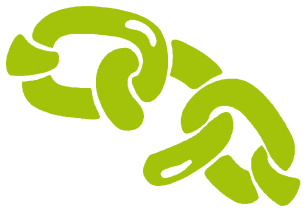


11.4.1 DNA序列的熵估计和压缩

- 前面介绍的离散序列熵估计算法都可用来估计内含子和外显子序列的熵，但由于DNA序列中外显子序列较短，要求熵估计器的收敛必须足够快，以便得到精确的估计值。
- 因为DNA序列是有记忆的，最常用的方法就是k次扩展估计方法。先把序列分割成长度为k的子序列，对每个k，估计长度为k的子序列的概率，利用插入法得到每符号熵的估计值如下式表示：

$$\hat{H}(k)=\frac{1}{k}\sum_{\boldsymbol{x}\in(A,T,G,C\}^k}-\hat{p}_k(\boldsymbol{x})\log\hat{p}_k(\boldsymbol{x}) \qquad (11.67)$$

其中，$\hat{p}_k(\boldsymbol{x})$ 为所估计的长度为k的子序列的概率。

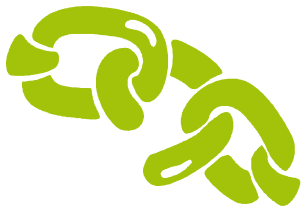


11.4.1 DNA序列的熵估计和压缩

- 也可用滑动窗模板匹配法进行熵估计。Farach等用此算法来测试外显子和内含子熵差别，发现外显子的平均熵73%的时间较大，而内含子的变化80%的时间较大。
- Loewenstern和Yianilos提出一个称作CDNA的估计DNA序列熵的算法。他们观察到，DNA序列包含很多重复，偶尔也可以预测。该算法使用两个参数表示这种不精确匹配，一个参数w表示子串的长度，另一个参数h表示汉明距离。这两个参数用来构建一个预测专家平台pw，h，它们各有不同的w和h的值，然后应用期望值最大法对各个专家平台参数的加权值进行训练，使得将它们组合成一个单独预测器时预测能力最强。

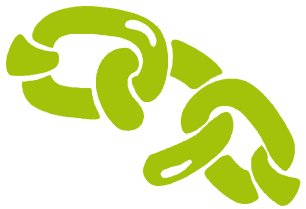
11.4.1 DNA序列的熵估计和压缩

- Kevin Lonctot等提出一种称做语法变换分析和压缩（GTAC）熵估计器，应用一种新的数据结构以线性时间解最长非重叠模式问题，其特点是运行时间短，估计值精确。该方法以基于语法的编码分析为基础，并利用了DNA序列的反向互补特性。这种估计器是通用的，适用于任何序列。



11.4.1 DNA序列的熵估计和压缩

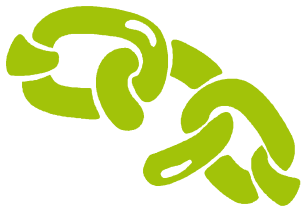
- 如前所述，无损压缩编码也可用做熵的估计，所以提出了若干无损压缩编码用于DNA序列熵的估计和压缩。实践证明，适用于文本压缩的通用无损信源编码往往对DNA序列的压缩效果不好，主要原因是：（1）这些算法大部分仅具有渐近最佳特性，而DNA序列往往没有足够的长度；（2）这些算法未利用DNA序列本身的特性。
- 我们知道，DNA序列有如下特性：（1）序列中存在重复，（2）存在近似重复（有个别差错的重复），（3）存在反向互补重复（reverse complement），（4）局部频率非均匀。



11.4.1 DNA序列的熵估计和压缩

- 在DNA序列中，A与T互补，G与C互补。如果两序列***x***=*x*1…*x*n 和***y***=*y*1…*y*n 中 xi与yn+1-i(1in)互补，则称***x*** 和***y***反向互补，也称回文（palindrome）。例如AAACGT与ACGTTT就是反向互补。

- 迄今为止已经提出多种DNA序列的压缩算法，主要有BioCompress(BioCompress-2)，Cfact， GenCompress， CTW+LZ，DNCompress，DNASequitur，DNAPack和GenomeCompress等。

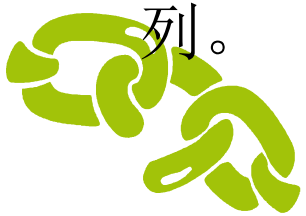


11.4.1 DNA序列的熵估计和压缩

- Biocompress和它的第2版Biocompress-2(Grumbach和Tahi,1994)是第一个DNA专用压缩算法，其要点是：（1）精确检测直接和反向互补重复，（2）在每一步，选择从当前位置开始与前面开始的最长匹配，用LZ编码，其中一个子符号串编成一对整数，一个表示匹配长度，另一个表示匹配位置，（3）不重复的数据段用2比特编码，（4）当未发现重复时，Biocompress-2用2阶算术编码器编码。
- Cfact（Rivals 等，1996）算法的要点是：（1）寻找最长的精确匹配，（2）两次通过（保证增益），（3）用后缀树寻找最长重复，（4）对于不重复部分用2比特编码。

11.4.1 DNA序列的熵估计和压缩

- GenCompress（Chen等，1999）算法要点：（1）考虑近似重复，（2）在每一步，DNA序列中还未编码部分（后缀）的最佳前缀（增益函数），（3）如果在使用最佳前缀中无任何增益，就在缓冲器中加一个字母，（4）汉明距离和编辑距离用做近似重复。
- CTW+LZ（Matsumoto等，2000）算法要点：（1）GenCompress和CTW（上下文树加权）的组合，（2）长的精确或近似的重复用LZ77型算法，而短重复用CWT编码，（3）用局部启发式解贪婪选择问题，（4）执行时间长，不能用于长序列。



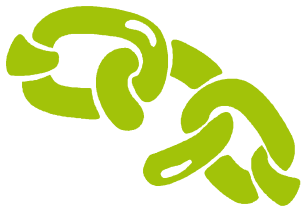
11.4.1 DNA序列的熵估计和压缩

- DNCompress（Chen等，2000）要点：（1）使用LZ压缩，（2）用PatternHuter工具进行预处理，寻找所有的近似重复包括反向互补重复，（3）对近似重复和非重复区域编码，（4）执行时间短。
- DNASequitur（Cherniavsky 和Ladner，2004）是一个基于语法的压缩算法，其要点是：（1）提供一个上下文无关的语法来表示输入数据，（2）Sequitur（Nevill-Manning和Witten，1997）的要点是，Digram Uniqueness（在语法中，相邻符号对的出现不多于1次）和Rule Utility（每条原则至少用2次，起始原则除外）；（3）DNASequitur：为适配DNA序列，Sequitur的改进型。

11.4.1 DNA序列的熵估计和压缩

- DNAPack（Behshad Behzadi等，2005 ）是一个基于动态规划的算法，要点：（1）汉明距离用于重复和反向互补，（2）用CTW或2阶算术编码对非重复区编码，（3）用动态规划选择重复。（4）使用加速技术，也适用于长序列。

- GenomeCompress（Umesh Ghoshdastider等）算法据报道是压缩最好的算法。



11.4.1 DNA序列的熵估计和压缩

- 所有这些方法有很多共性，其中包括：

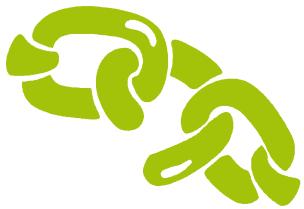
（1）利用DNA序列本身的特性，将序列分成重复（包括近似重复的反向互补）段和非重复段；

（2）采用类似于LZ的算法压缩重复段；

（3）采用性能好的通用压缩算法，例如算术编码、CWT等算法压缩非重复段；

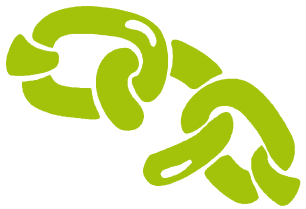
（4）采用改进的搜索算法寻找重复段。

- 资料表明，这些算法可以把DNA序列压缩到1.7比特左右。



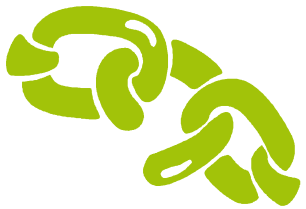
11.4.2 最大熵谱估计和最小交叉熵谱估计

我们知道，信号的时间相关函数可直接利用信号的时间波形样值计算，而相关函数和功率谱是互为富氏变换的关系。所以信号功率谱的估计通常要通过计算相关函数来实现。常规的谱估计方法要对信号的样值序列加窗，利用有限时间内的样值计算相关函数，然后进行功率谱的估计。如果所使用的时间段太长，就不能保证信号的平稳性，而使用的时间段太短，就会降低功率谱的分辨率。



11.4.2 最大熵谱估计和最小交叉熵谱估计

而且由于加窗的影响，还使窗内的信号失真，同时还迫使窗外的信号为零，而窗外实际的信号未必是零。还有一种谱估计方法方法，就是将自相关函数向未知区域延伸，但如何合理延伸是一个关键问题。Burg提出最大熵谱估计的方法，在自相关函数的延伸时，使由功率谱所确定的熵率最大，即在所计算的自相关函数的约束下，把使信源熵率最大的功率谱作为估计的结果。



11.4.2 最大熵谱估计和最小交叉熵谱估计

1．最大熵谱估计

- 一个限带高斯连续时间信源的熵率与它的功率谱的关系由下式确定：

$$h(X)=\frac{1}{2}\log(2\pi e)+\frac{1}{4W}\int_{-W}^{W}\log[S(f)]df \quad (11.68)$$

其中，W为信号的带宽，S（f）为信号的功率谱密度。

- 通过计算得到的信号的自相关函数序列就是功率谱的约束，即

$$R(k)=\int_{-W}^{W}S(f)\exp(j2\pi fk\Delta t)df \qquad -p\le k\le p$$

，　　（11.69）

其中，Δt为时间抽样间隔，2p+1为自相关函数样值的个数。

11.4.2 最大熵谱估计和最小交叉熵谱估计

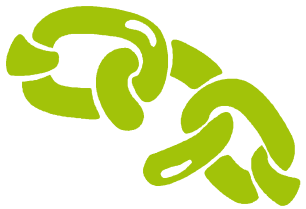
- 最大熵谱估计就是求在给定约束下使熵率达到最大的信号功率谱。下面求在（11.69）的约束下，（11.68）的极值。
- 令 $J=\frac{1}{2}\log(2\pi e)+\frac{1}{4W}\int_{-W}^{W}\log[S(f)]df-\sum_{k=-p}^{p}\lambda_k\int_{-W}^{W}S(f)\exp(j2\pi fk\Delta t)df$

  求J对S(f)“导数”，并令其为零，得

$$S(f)=\frac{1}{\sum_{k=-p}^{p}\lambda_k\exp(j2\pi fk\Delta t)} \qquad (11.70)$$

可以证明 $\sum_{k=-p}^{p}\lambda_k e^{j2\pi fk\Delta t}=\left|1+\sum_{k=1}^{p}a_k e^{-j2\pi fk\Delta t}\right|^2/\sigma^2$ （11.71）

其中，$a_k(k=1,\cdots,p),\sigma$ 可通过与自相关函数有关的方程组求解。



11.4.2 最大熵谱估计和最小交叉熵谱估计

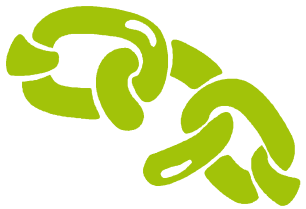
将（11.71）代入（11.70），得到最大熵功率谱估计为

$$S(f)=\frac{\sigma^2}{\left|1+\sum_{k=1}^{p}a_k e^{-j2\pi fk\Delta t}\right|^2} \qquad (11.72)$$

满足（11.72)式的信号x(t)称为自回归过程，其时间序列满足：

$$x_n=-\sum_{i=1}^{p}a_i x_{n-i}+z_i \qquad (11.73)$$

其中，xi 为x(t）的抽样值，zi为均值为0，方差为σ2的高斯白噪声序列。



11.4.2 最大熵谱估计和最小交叉熵谱估计

根据（11.73），得 $R(0)=\sum_{i=1}^{p}a_iR(-i)+\sigma^2$ （11.74）

$$R(j)=\sum_{i=1}^{p}a_iR(j-i)$$

（11.75）

(11.74)和（11.75)称做Yule-Walker方程。利用x(t) 的抽样值x(i)估计相关函数R(i)，再用Yule-Walker方程来计算 a_k 和 σ ，最后根据（11.72)式得到信号频谱的估计。根据相关函数R(i)估计的方式不同可分为自相关法和协方差法以求解方程组。

- 最大熵谱估计比常规的谱估计的分辨率有很大提高，成为当前重要的实用谱估计算法。

11.4.2 最大熵谱估计和最小交叉熵谱估计

2. 最小交叉熵谱估计

- 最小交叉熵谱估计可以看成自相关函数的另一种延伸方式，这里考虑到一个先验估计，在谱估计时，使被估计的过程和先验估计之间的交叉熵最小。如果先验估计是平坦谱，那么最小交叉熵谱估计就归结为最大熵谱估计。

- 设概率密度p属于某概率集合P，该集合是已知的，但p本身未知，q为先验密度，同时还有p满足的约束条件。最小交叉熵谱估计的原理就是：在所有满足约束的密度中，选择与先验密度q交叉熵最小的概率密度p，如（11.58)式。

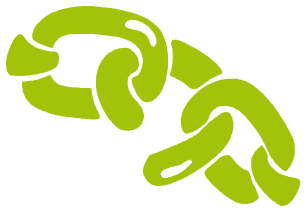
- 由于利用了先验信息，最小交叉熵谱估计比最大熵谱估计的性能有改善。

11.4.3 最大熵建模及其在自然语言处理中应用

1. 最大熵建模基本原理

- 建模就是构造一个精确表示随机过程行为的随机模型，估计在给定上下文x条件下输出y 的概率p（y|x），其中x为模型的输入，y为输出。

- 为设计一个适合某种过程的模型，需要对该过程的行为进行一段时间的观察，收集样本值作为训练数据。设训练样本集有N对样本值，表示为$\{(x1, y1), (x2, y2), \cdots, (xN, yN)\}$。



11.4.3 最大熵建模及其在自然语言处理中应用

- 定义两种分布，一是经验分布，就是通过训练数据得到的分布；二是模型分布，就是信源实际的分布。训练集合中数据对的分布称为经验分布，定义为：

$$\tilde{p}(x,y)=\frac{1}{N}\times(x,y)\text{在训练集合中出现的次数} \qquad (11.76)$$

通常，一个特殊的数据对要么不出现，要么出现多次。

- 最大熵建模就是以训练数据为依据，用最大熵原理构造一个产生训练样本经验分布 $\tilde{p}(x,y)$ 的统计模型，这里估计的是条件概率 $p(y|x)$ 。

11.4.3 最大熵建模及其在自然语言处理中应用

- 建模的一个重要步骤就是从训练数据中提取特征。特征或特征函数指的是x与y之间存在的某种特定关系，可以用一个输出为0或1的二值函数（或示性函数）表示。 特征实际上是一种映射：$f_i:\varepsilon\to(0,1)$，其中，$\varepsilon\in A\times B$ 。A为y的符号集，表示一个可能的类集合；B为x的符号集，为上下文集合。

- 对于一个特征(x0,y0)，定义特征函数：

$$f_{x_0,y_0}(x,y)=\begin{cases}1 & \text{若y}=y_0\text{且x}=x_0\\ 0 & \text{其他}\end{cases} \qquad (11.77)$$

11.4.3 最大熵建模及其在自然语言处理中应用

- 实际上，特征函数的定义与所解决的问题有关。以文本分类问题为例。假设有4类文本：政治、经济、体育和文艺。每个词在不同类的文本中出现的概率是不同的，特别是具有代表性的词类。例如，“货币”一词经常出现在经济类的文本中，而“比赛”一词经常出现在体育类的文本中。对于一个特征（“球”，“体育”），其中“球”属于上下文集合，“体育”属于类集合，其特征函数定义为：

$$f_{球,体育}(x,y)=\begin{cases}1 & if\ y=\text{“体育”且“球”在x中出现}\\ 0 & \text{其他}\end{cases}$$

- 用经验分布对特征求平均是有用的统计量，对每一个特征，表示为

$$E_{\tilde{p}}(f)=\sum_{x,y}\tilde{p}(x,y)f(x,y) \qquad (11.78)$$

11.4.3 最大熵建模及其在自然语言处理中应用

- 用模型 $p(y|x)$ 表示对f的期望值为

$$E_p(f)=\sum_{x,y}\tilde{p}(x)p(y|x)f(x,y) \qquad (11.79)$$

- 其中，$\tilde{p}(x)$为训练样本中x的经验分布。我们令经验分布特征平均值与模型分布特征平均值相同，即要求对每一个特征有，$E_p(f)=E_{\tilde{p}}(f)$，或

$$\sum_{x,y}\tilde{p}(x)p(y|x)f(x,y)=\sum_{x,y}\tilde{p}(x,y)f(x,y) \qquad (11.80)$$

（11.80）称为约束方程或简称约束。当样本数足够多时，可信度高的特征的经验概率与期望概率是一致的。

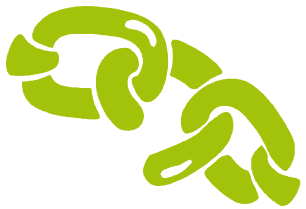
11.4.3 最大熵建模及其在自然语言处理中应用

- 设有n个特征函数{fi，i=1，…，n}，这是建模中重要的统计量，我们希望所寻找的模型符合这些统计量。定义P表示所有满足（11.80）约束的条件概率分布的集合，即

$$P = \{p = p(y \mid x) \mid E_p(f_i) = E_{\tilde{p}}(f_i), i \in \{1.2, \cdots, n\}\} \quad （11.81）$$

条件熵表示为 $H(p) = -\sum_{x,y} \tilde{p}(x)\, p(y \mid x) \log\ p(y \mid x)$（11.82）

最大熵建模就是：从满足约束条件的集合P中，选择具有最大熵的分布 $p^* \in P$ ，即 $p^* = \text{argmax}_{p \in P} H(p)$ 。



11.4.3 最大熵建模及其在自然语言处理中应用

这是一个求有约束优极值化问题，应用拉格朗日乘子法，引入拉格朗日乘子λ_i，并仿照（11.26）的推导，得

$$p_\lambda(y|x) = Z_\lambda(x)^{-1} \exp \sum_i \lambda_i f_i(x,y) \qquad (11.84)$$

$$= Z_\lambda(x)^{-1} \prod_j \alpha_j^{f_j(x,y)}$$

其中，

$$Z_\lambda(x) = \sum_y \prod_j \alpha_j^{f_j(x,y)} \qquad (11.85)$$

$$\alpha_j = \exp(\lambda_i) \qquad (11.86)$$

$$Q = \{q \mid q = Z_\lambda(x)^{-1} \prod_j \alpha_j^{f_j(x,y)}\} \qquad (11.87)$$

可以看到，最大熵建模的解p*满足$p^* \in P \cap Q$ （11.88）

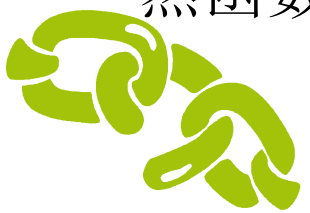
11.4.3 最大熵建模及其在自然语言处理中应用

- 由于训练序列各样本对都是独立的，所以一个N长的训练序列的对数似然函数为：

$$L_{\tilde{p}}(p) = \log \prod_{i=1}^{n} p(y_i \mid x_i) = \log \prod_{x,y} p(y \mid x)^{N\,\tilde{p}(x,y)}$$

$$= N \sum_{x,y} \tilde{p}(x,y) \log p(y \mid x) \qquad (11.89)$$

将（11.84）的结果代入（11.89），并求满足最大值的 λ_k ，通过推导得知，当满足（11.80）的约束时，（11.84）达到最大值，即满足（11.84）式的条件概率使训练序列的对数似然函数达到最大值。



11.4.3 最大熵建模及其在自然语言处理中应用

- 因此可得到以下结论：

最大熵建模的解p*满足：

（1） $p^* \in P \cap Q$ ；

（2） $p^* = \arg\max_{p \in P} H(p)$ ；

（3） $p^* = \arg\max_{p \in Q} L_{\tilde{p}}(p)$ ；

（4） p^* 是惟一的。

11.4.3 最大熵建模及其在自然语言处理中应用

- 最大熵建模在简单情况可以求出解析解，例如有一、二个约束情况。但一般情况最大熵问题没有显式解，求参数 λ^* 必须借助数值解法。有些实际问题，有时可能有上千个约束条件，计算量和花费的时间巨大，必须使用有效的算法。
- 一般化的迭代尺度算法（GIS，Generalized Iterative Scaling Algorithm)是一个专门用于最大熵问题的算法（Danroch 和Ratcliff，1972），该算法要求特征为非负值，没有解析解，收敛速度较慢。以后，D.Pietra等改进了原有的求解算法，降低了求解的约束条件，提出了IIS(Improved Iterative Scaling Algorithm)算法，增加了算法的适用性，IIS算法是目前最大熵参数求解中的常用算法。

11.4.3 最大熵建模及其在自然语言处理中应用

2. 最大熵统计模型的优缺点

- 最大熵建模方法有很多优点：

（1）与极大似然估计结果同，所建立的模型是唯一的；

（2）最大熵统计模型可以灵活地设置约束条件。通过约束条件的多少可以调节模型对未知数据的适应度和对已知数据的拟合程度；

（3）通常性能优于其他方法。

- 最大熵统计模型的缺点：

（1）运算量大；

（2）存在过拟合问题，通常在求极值时需加入先验随机函数进行平滑。

11.4.3 最大熵建模及其在自然语言处理中应用

3．最大熵建模在自然语言处理中的应用

- 最大熵建模已成功应用到自然语言处理的许多方面，其中包括：

  单词聚类(S.Pietra)

  机器翻译（A.L.Berger）

  句子边界检测，词类标注(Ratnaparkli,1998)

  自适应统计语言建模(Rosenfeld,1996)

  组块分析(Osborne,2003;Koeling,2003)

  垃圾邮件过滤(Zhang,2003)

  名实体识别(A.Borthwick)

11.4.4 最大熵原理在经济学中的应用

- 前面指出物理学中的波耳兹曼分布是一个指数分布。推导该定律的基本依据是能量守恒定律。因此，我们可以推断，在一个大系统中任何守恒的量都应该具有指数概率分布。在物理学中指数Boltzmann-Gibbs分布和封闭经济系统中的货币的平衡分布具有类似性，与能量类似，在一个封闭经济体中，货币在经济代理商之间的相互作用中在局部是守恒的，所以货币也遵循Boltzmann-Gibbs分布，其等效温度等于平均每个代理商的货币量。财富不但包含货币还包含物质财富，所以不守恒，一个早期的研究者Vilfledo Pareto在19世纪末发现，在一个人口均匀分布的地理范围内，人们之间的财富的分布按一个幂律分布，因此这种分布经常称做Pareto分布。下面利用最大熵原理推导和分析封闭的经济体中货币和财富的分布。

11.4.4 最大熵原理在经济学中的应用

1.封闭经济体中货币量的分布

- 守恒的交换市场可以看成大量的经济代理商组成，他们之间通过买卖相互作用。在这种市场中，生产活动并不存在，而且在每次交易中只是货币的交换，例如，一个保险公司中一个代理人的行为，所有其他代理人都可以看成行为的环境。他们按市场手段进行交换，此环境吸收代理人损失的货币同时又提供给代理人货币作为其收入。这里货币的总量是守恒的。
- 假设在一个经济体中，总货币量为M元，总代理商数为N。假定在经济发展的每个阶段，一个代理商随机选择另一个代理商，并给所选择的代理商汇入1元，规定无资金的代理商不能借贷。求货币的稳态分布（拥有货币量i元的代理商的概率），设交易初始每代理商具有相同的货币量。

11.4.4 最大熵原理在经济学中的应用

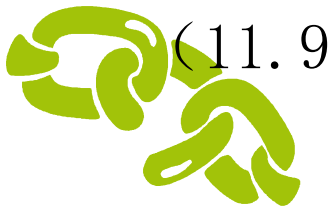
- 解可以利用最大熵原理解决这个问题。

  设 p_i 为有i元代理商的概率，为有i元的代理商个数，根据题意有 $\sum_i i \cdot n_i = M$ ，$\sum_i n_i = N$ ，当 N 很大时，所求的概率近似为占总体的比例数，即 $p_i = n_i / N$ ，所以

$$\begin{aligned} \sum_i i \cdot p_i &= M / N \\ \sum_i p_i &= 1 \end{aligned} \qquad (11.90)$$

  实际上，确定货币的稳态分布，就是求在满足（11.90）约束下对应最大熵的分布。与（11.25）的条件对比，得

$$p_i = \frac{e^{-\lambda_1 \cdot i}}{\sum_i e^{-\lambda_1 \cdot i}} \qquad i = 1,2,\cdots,M \qquad (11.91)$$



11.4.4 最大熵原理在经济学中的应用

- 由 $T = M/N = \sum_i i \cdot e^{-\lambda_1 \cdot i} / \sum_i e^{-\lambda_1 \cdot i} = \frac{e^{-\lambda_1}}{1-e^{-\lambda_1}} \approx 1/\lambda_1$，得

$$\sum_i e^{-\lambda_1 \cdot i} = \frac{1}{1-e^{-\lambda_1}} \approx T \qquad (11.92)$$

$$\lambda_1 \approx 1/T \qquad (11.93)$$

将（11.92）和（11.93）代入（11.91），得

$$p_i = \frac{1}{T} e^{-i/T} \qquad (11.94)$$

- 其中，T为平均每代理商的货币数。（11.94）表明，在总货币守恒的市场中，货币的分布服从Boltzmann-Gibbs分布。这个结论也可用来分析一个社会流通系统，根据对收入分布的数据分析，在很多国家在大多数人口收入分布可用Boltzmann-Gibbs分布描述。
- 有些数据表明，该模型与现实数据统计的拟合度较好。

11.4.4 最大熵原理在经济学中的应用

2．封闭经济体中财富的分布

- 设由很多代理商构成一个经济体，每代理商i在时刻t的财富为$wi(t)$，经济体中的总财富为$w(t)=\sum_i w_i(t)$；代理商财富可能值集合为{wi}，令Ri=wi(ti)/wi(t0)为代理商i的相对财富，对所有i；每代理商i的经济增长率为ln（Ri）；求在平均经济增长率为C*ln (R)的约束条件下，具有最大熵的财富分布。
- 解

  为便于计算，用连续变量x近似Ri，概率密度函数p(x)近似原来的离散分布，分布的熵为

$$h(X)=-\int_1^{\infty} p(x)\ln p(x))dx \qquad (11.95)$$

  满足

$$\int_1^{\infty} p(x)dx = 1 \qquad (11.96)$$

11.4.4 最大熵原理在经济学中的应用

平均增长率约束为 $\int_1^{\infty} p(x)\ln x dx = c\ln(R)$ （11.97）

其中，$\ln x$ 为经济增长率，c为常数。现求在平均经济增长率为规定值条件下，财富分布熵的最大值。

根据例11.9的结果，所求的分布应该是幂律分布。根据（11.35）、（11.36）式，

得财富的幂律分布为

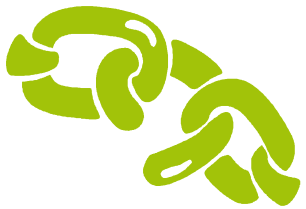
$$p(x) = Z^{-1}e^{-\lambda \ln x} = (\lambda - 1)x^{-\lambda} \quad （11.98）$$

再根据（11.34），得

$$\lambda - 1 = 1/[c\ln(R)] \quad （11.99）$$

所以

$$p(x) = c\ln(R)x^{-1-1/[c\ln(R)]} \quad （11.100）$$



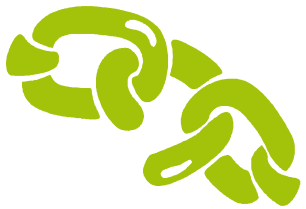
11.4.5 信息论方法应用展望

上面所介绍的内容仅仅是信息理论方法应用的很少一部分，实际上，由于篇幅所限，很多应用领域还没有提及。不过我们应认识以下几点：

（1）香农信息论研究的是语法信息，主要解决语法层次的信息处理的问题，但也可以解决某些语义信息的问题，如机器翻译等；

（2）在信息论的发展中不断提出新的信息度量方法，例如Golmogolov熵、Ronyi熵和Tsallis熵等，丰富了信息熵内容；

（3）香农信息论在其他领域仍有广阔的应用前景。



本章小结

1．信源熵的估计方法主要有：

（1）插入法，（2）无损压缩编码法，（3）模板匹配法

2．最大熵原理是一个利用部分信息确定随机变量集合概率分布的方法。

它的基本思想是，求满足某些约束的信源事件概率分布时，应使得信源的熵最大。

几种重要的最大熵分布：

- 满足均值约束的离散分布是几何分布；
- 满足均值约束的连续分布是指数分布；
- 满足均值和均方值约束的分布是高斯分布；
- 满足几何平均值约束的分布是幂律分布。

本章小结

3. 最小交叉熵原理是最大熵原理的推广。

当推断一个具有先验分布随机变量X的分布密度时，最小交叉熵原理选择在满足X的已知约束下使交叉熵最小的分布密度。

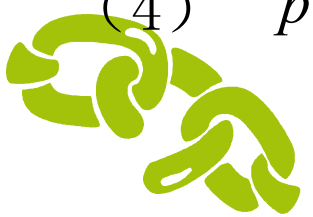
4. 最大熵建模的解p*满足：

（1） $p^* \in P \cap Q$ ；

（2） $p^* = \arg\max_{p \in P} H(p)$ ；

（3） $p^* = \arg\max_{p \in Q} L_{\tilde{p}}(p)$ ；

（4） p^* 是惟一的。



本章小结

5．信息理论方法的应用

- DNA序列的熵估计；
- 最大熵谱估计；
- 最小交叉熵谱估计；
- 最大熵原理建模在自然语言处理中的应用；
- 最大熵原理在经济学中的应用 。

谢谢！