04104-106通信原理II第一讲

Feb. 27, 2007

1 信源的描述

信源或者是连续时间的，或者是离散时间的。如果是连续时间的，则总可以通过抽样使其变成离散时间的，因此以下只考虑离散时间的信源。

信源的输出当然是随机的，故可以把信源输出建模为一个随机序列$\{X_i\}$。

随机序列有千千万万，我们只考虑简单的一种：广义平稳随机序列。即序列中的任何一个X_i,不论时间坐标i是多少，它们的概率分布都相同，而且自相关函数$E[X_iX_j]$只与$i-j$有关。

一个信源所表达的信息当然是由整个序列贡献的，正如我这个讲稿所要表达的意思是由全体字符串构成的。先来考虑简单的情形，即序列$\{X_i\}$中的某一个单个符号。$\{X_i\}$是平稳序列，具体选哪一个都是一样的。为了省事，我们略去代表时间的下标i，这样，要研究的就是单个随机变量X。

随机变量的取值范围是实数或者实数构成的集合。例如Gaussian随机变量的取值范围是$(-\infty,\infty)$, Rayleigh随机变量的取值范围是$[0,\infty)$, 载波相位的取值范围是$(0,2\pi)$，《通信原理II》成绩的取值范围是整数$\{0,1,\cdots,99\}$－也是实数的子集。考虑简单的情形，假设X的取值x的范围是M个实数：

$$X=x\in\{x_1,x_2,\cdots,x_j,\cdots,x_M\}$$

实际信源的输出也许是连续取值的实数，但我们总可以通过足够精细的量化使它只有有限个取值。

现在，我们要研究的就是一个离散随机变量X，其取值范围是$\{x_1,x_2,\cdots,x_M\}$。X取值等于x_1，或者x_2，或者其他的概率很可能是不一样的，记其取$x_j,j=1,\cdots,M$的概率为P_j，即$P_j=P\{X=x_j\}$。注意，有时我们也用$P(X)$或者$P(x_j)$这样的记号。

对于离散随机变量，只要枚举出所有可能结果的出现概率，就完全描述了其概率统计特性。因此可以用下面的方法来表述这个事情：

$$\begin{pmatrix} X \\ P(X=x_j) \end{pmatrix}=\begin{pmatrix} x_1 & x_2 & \cdots & x_M \\ P_1 & P_2 & \cdots & P_M \end{pmatrix} \tag{1}$$

当然了，如果X是连续随机变量（有无限多种取值），我们也可以将其用类似的方法表述为

$$\begin{pmatrix} X \\ p\,(x) \end{pmatrix}=\begin{pmatrix} x\in(a,b) \\ p\,(x) \end{pmatrix} \tag{2}$$

再来看随机序列。随机序列是无限多个随机变量。先从两个随机变量X_1和X_2说起。掷两个筛子的结果可以当作是掷一个有$6\times6=36$面的筛子。由此类推，

两个随机变量X_1和X_2可以理解为一个随机的Y，它的结果是向量(x_i, x_j), 其中x_i是X_1的结果，x_j是X_2的结果。因而，仿照式(1), 可以将其描述为

$$\begin{pmatrix} X^2 \\ P(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} (x_1,x_1) & (x_1,x_2) & \cdots & (x_M,x_M) \\ P_{11} & P_{12} & \cdots & P_{MM} \end{pmatrix} \tag{3}$$

这个结果显然可以推广到任意L个随机变量构成的序列$\{X_1, X_2, \cdots, X_L\}$。这个序列的每一种出现结果是一个向量$(x_{i_1}, x_{i_2}, \cdots, x_{i_L})$，其中
$x_{i_k} \in \{x_1, x_2, \cdots, x_M\}$。这样的向量有$M^L$个可能结果。于是，整个序列可以描述为

$$\begin{pmatrix} X^L \\ P(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{M^L} \\ P_1 & P_2 & \cdots & P_{M^L} \end{pmatrix} \tag{4}$$

要点:把L长的序列看成一个更大的单个符号。

当我们用式(4)来描述随机序列时，一个关键就是如何写出每一种总结果的出现概率。如果这L个随机变量是相互独立的，那么总结果的出现概率就是各个随机变量出现概率的乘积。比如课本上的式(7.2.10)。这种情况称为无记忆信源。是说，信源每次输出的结果都与先前的输出无关，其随机性完全是单独的。

如果序列是相关的，那么写出第i种向量结果$\mathbf{x}_i$的出现概率就不能用各自概率的乘积，而要用链式规则:

$$\begin{aligned} P(X^L = \mathbf{x}) &= P(X_1 = x_{i_1}, X_2 = x_{i_2}, \cdots, X_L = x_{i_L}) \\ &= P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)\cdots P(X_L|X_{L-1}\cdots X_1) \end{aligned} \tag{5}$$

其中$P(X_1)$是$P(X_1 = x_{i_1})$的简记，其他类似。
对式(5)的理解是：X_2的结果是随机的，但其结果和X_1是有关系的。如果已知X_1的具体结果之后再来看X_2，虽然它还是不确定的，但此时的不确定性（随机性）是X_1之外的原因造成的，这个原因和造成X_1随机性的原因是独立的，因此就可以进行概率相乘了，即$P(X_1, X_2) = P(X_1)P(X_2|X_1)$。$P(X_2|X_1)$是已经知道$X_1$的条件下，$X_2$还具有的随机性。

2 信息量

信源发送随机变量X给信宿。因为信宿不能预见到X具体是多少，所以在信宿看来，X是随机的。而信宿所获得的信息量也是和这种随机性关联在一起的。假如信宿早就能料到X会是什么，那么信源发送X给信宿这个行为没有使信宿得到任何信息。

X有M种不同的取值，每种的可能性都不同。就是说，信宿未见到发送内容之前，它对信源发送的X的预料是：有P_1的机会是x_1、有P_2的机会是x_2，……。当信宿确实观察到X之后，事先的种种猜测已经变成现实。

若有一个事情，你不确定其具体结果是什么，只能猜测的话，这个事情对你来说就存在着“不确定性”。有人告诉了你结果，将使你消除不确定性，也就是说你获得了信息。因此，所谓“信息”其实就是“不确定性”。

如果X的内容越是预料之外，信息量也就越大。完全在料定之中的事情如果发生，和别人不告诉你结果是一样的，因而是没有信息量的（别人不告诉你结果，并没有改变你对这个事情的认识）。举例来说，某生参加《通信原理II》

考试，依其自我判断，他认为考试成绩X取值于$(99,100)$范围的概率大于0.99，考试成绩在$[60,99)$的概率是0.009。如果同学查分后告诉他考了99分，这个消息对他没有多少信息量。但如果同学告诉他说他考了12分，那么这可是个重大消息，足以使他产生多种常见的狂躁情绪。

以西洋科学为特征的“现代科学”的基本原则是要把所研究的问题数学化，一般就是数量化。为了研究“信息”这个东西，我们先试图将其量化。若信息的量是I, 那么依前面的讨论，I的大小是和X的具体实现有关系的，不同的X的实现，其随机性（概率）不同，因此信息量也不同，因此可记为$I[P(x_j)]$, 即I是概率$P(X=x_j)$的函数。这个函数应该满足: I是概率的减函数，概率越大，信息量越小（概率越大，就越在预料之中）。若概率为0则I是无穷；若概率为1，则I应该是0。另外，两个独立事件组成的事件（比如通原II考试得99，食堂吃饭发现蟑螂肉也很好吃。没有迹象表明这两件事情有因果联系，故此可认为它们独立）的信息量应该是各自信息量之和。你同学告诉你这两件事情，你所获得的信息的数量应该就是相加的关系。但在概率论中，这两个事件的概率是相乘的关系。这就是说I作为概率的函数，乘积的函数必须是函数的和：$I(p_1\cdot p_2)=I(p_1)+I(p_2)$。Shannon等先驱经过苦苦思索后，发现这个函数应该是

$$I(p)=-\log(p) \tag{6}$$

对数取什么底无关紧要。如果以2为底，则I的单位叫bit；如果是以e为第，则称奈特(nat); 若以10为底，则称为Det（笛特）。

如果随机变量X有两种可能，机会各占一半，则无论哪一个发生，信息量都是$-\log_2(1/2)=1bit$。因此，“1bit”的意思就是，在两个等可能的不确定的结果中，确定了其一发生，便是得到了1bit的信息量。

要点：信息的多少就是不可预料性的大小，它是概率的函数，取值于$[0,\infty)$,满足可加性和递减这两个关键特征。这样的函数只能是log函数。

信息量是概率的函数，条件概率和联合概率自然也有相应的信息量。条件概率$P(Y=y_j|X=x_i)$所算出的信息量是指，已经知道$X=x_i$的情况下，继续告诉信宿$Y=y_j$所能带给信宿的信息量。如果X与Y不独立，那么$I[P(Y|X)]$将不见得会等于$I[P(Y)]$。比如说，若X代表升高，Y代表体重。那么X的随机性和Y的随机性是高度相关的。尽管高度相关，即使已经知道了身高，你还是不能断定体重是多少。$Y|X$表示给定身高条件下Y的随机结果，相同身高条件下，造成体重随机性的因素已经和造成身高的随机性因素无关了。所以相应概率，例如$P(Y=10kg|X=180m)$的信息量$I[P(Y=10kg|X=180m)]$是在你已知身高为180m的条件下，继续告诉你身高只有10kg, 这个消息带给你的信息量。

注意，条件信息可能比无条件的信息量大或小。（信息量的大小就是你的意外程度）。

联合概率$P(X,Y)$的信息量容易理解。我们前面已经说过，可以把(X,Y)整体当作一个事件，因此联合概率的信息量就是这个整体事件的发生结果所带来的信息量。

3 熵

把随机变量X的每一种结果x_j通过函数$y=f(x)$映射为$y_j=f(x_j)$，我们就定义了另外一个随机变量$Y=f(X)$。

现在，X的每个结果x_j都通过其发生概率P_j映射为一个实数$I_j=-\log(P_j)$, 因此I也是一个随机变量。也就是说，因为发生哪个是不确定的，所以信宿得到

的信息量也是不确定的。

信源X的平均信息量称为其熵(entropy)，记为$H(X)$：

$$H(X) = E[I] = \sum_{j=1}^{M}(P_j \times I_j) = -\sum_{j=1}^{M}(P_j \log P_j) \tag{7}$$

熵作为信息量的数学期望，其单位自然也和信息量一样，是bit或者nat，取决于对数的底是多少。如果我们谈论的是随机序列$\{X_i\}$中某一个符号的熵，则习惯把单位写成“bit/symbol”。

二元熵函数：如果X是二进制随机变量，其可能取值只是x_1, x_2, 对应的概率是$p, 1-p$。那么X的熵是

$$h_2(p) \triangleq H(X) = -p\log(p) - (1-p)\log(1-p) \tag{8}$$

其图见课本296页。

联合熵: 两个随机变量X、Y有联和概率，因此有联合的信息$I[P(X,Y)]$，这个I的大小与具体的(X,Y)取值有关，是依赖于X、Y的随机量，对它求数学期望就是联合熵：

$$\begin{aligned} H(X,Y) &= E[-\log P(X,Y)] \\ &= -\sum_{i=1}^{M}\sum_{j=1}^{M} P(X=x_i, Y=y_j)\log P(X=x_i, Y=y_j) \end{aligned} \tag{9}$$

条件熵: 条件概率$P(Y|X)$对应有条件信息$I[P(X,Y)]$，它也是与具体的(X,Y)取值有关的，对其求数学期望就是条件熵：

$$\begin{aligned} H(Y|X) &= E[-\log P(Y|X)] \\ &= -\sum_{i=1}^{M}\sum_{j=1}^{M} P(X=x_i, Y=y_j)\log P(Y=y_j|X=x_i) \end{aligned} \tag{10}$$

利用前面提到的的链式法则：$P(X,Y) = P(X)P(Y|X)$，可得

$$H(X,Y) = E[-\log P(X,Y)] = E[-\log P(X) - \log P(Y|X)] = H(X) + H(Y|X) \tag{11}$$

同样可以得到

$$H(X,Y) = H(Y) + H(X|Y) \tag{12}$$

这些结果是很好理解的：注意熵就是平均的信息量。得知X，Y这两个结果平均获得的信息量是：先知道X所获得的信息量，再加上又知道Y所新增加的信息量。条件熵$H(Y|X)$就是已知X之后，继续知道Y所新增加的平均信息量。

关于熵有下面的不等式：

$$H(X) \le \log M \tag{13}$$

等号在X等概出现（$P(X=X_i) = \frac{1}{M}$）时成立，此即“等概的信源熵最大”。如果信源等概，我们对它是什么最没有把握。假如我们知道某种结果要比另外

一种出现机会大一些，那么我们实际上已经有了一点信息，在出现可能性判断上已经能有所偏向。因此，等概熵最大是很合理的事情。
证明: $H(X) - \ln M = E[-\ln P(X) - \ln M] = E[\ln \frac{1}{MP(X)}]$，由$\ln(x) \leq x - 1$得

$$H(X) - \ln M \leq E[\frac{1}{MP(X)} - 1] = E\left[\frac{1}{MP(X)}\right] - 1 = \sum_{i=1}^{M} P(x_i)\frac{1}{MP(x_i)} - 1 = 0$$

#

用类似方法还可以证明

$$H(X|Y) \leq H(X) \tag{14}$$

等式在独立时成立。这个结果的意思是说，如果两个随机变量有关系，那么已知第一个随机变量的结果时，再告诉你第二个随机变量的结果所带给你的平均信息量将比单纯告诉你第二个结果要来的小一些。（但不排除个别情况下，已知第一个事情后可，第二个事情的结果更让你吃惊，条件信息量$I(P(Y|X))$可能比$I[P(Y)]$大或者小，条件熵$H(Y|X)$一定不会比$H(Y)$更大）。