信息论

信号传输与处理的理论基础

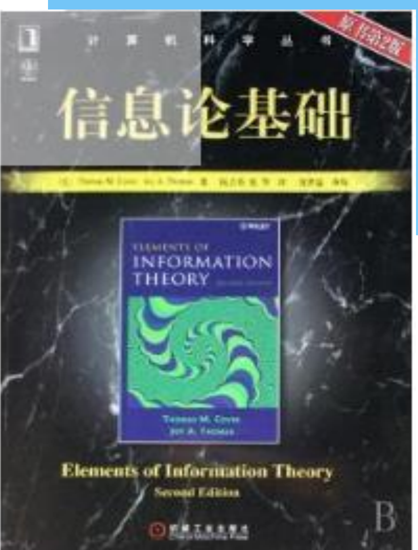
第二单元：基本概念

熵、互信息量、Markov过程的信息处理不等式、

典型集合的渐进均分性质

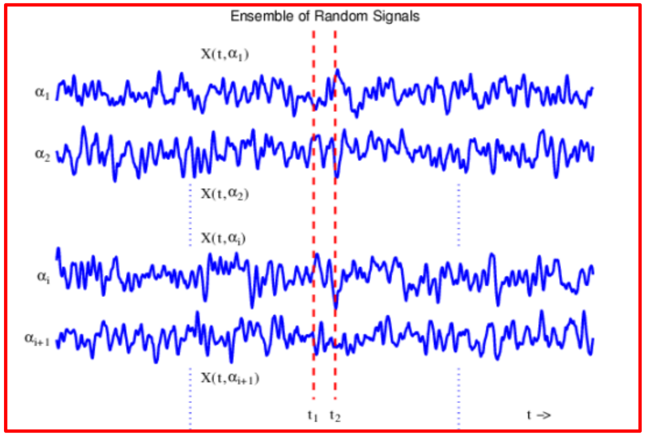
（教程/第一版 第2、3、8章；教程/第二版第2、3、9章）

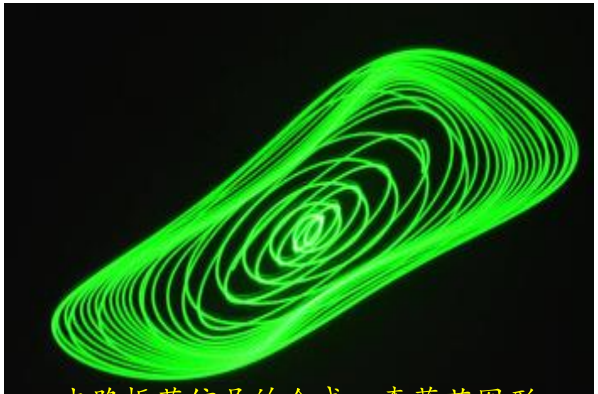
本节课教程阅读：2.1-2.7



信息熵（1）

* 基本概念（定性的表达）
* ⊙只有随机性对象或者随机过程，才蕴含信息。
* ⊙一个对象的随机性越强、随机过程的不可预期性越高，“信息量”越大。
* ⊙一个确定性的对象或过程，不论是多么复杂的函数，都不蕴含任何信息。
* ⊙等价地，一个完全确定性的对象、或完全可预期的过程，不蕴含任何信息（信息量为零！）。




电路振荡信号的合成：李萨茹图形

信息熵（2）

* 一个基本约定：
* 本单元和下一单元的大部分，“随机变量”均指离散型随机变量，即随机变量X的取值范围是有限集$\{x_1,\ldots,x_n\}$，对应的概率分布是$\{p_1,\ldots,p_n\}$，$p_j>0$, $p_1+\ldots+p_n=1$.

熵之父：Boltzman

* 基本概念（定量的表达：离散随机变量的情形）
* 随机变量X的Shannon熵
* $$H[X] = -\Sigma_j p_j \log p_j = -(p_1\log p_1 + \ldots + p_n \log p_n)$$

* ⊕ 熵是同随机变量相关联的一个非负数，更确切地说，是同概率分布相关联的一个非负数：两个相同的概率分布，无论其实际含义是否相同，具有相同的信息熵。
* ⊕ 从函数的观点看，信息熵H[X]总是n元变量（$p_1,\ldots,p_n$）的非负值函数。
* ⊕ 信息熵是对随机变量X所蕴含的“信息量“大小的度量。
* ⊕ 信息熵无量纲，在其定义中，log的底数可为任何大于1的正数，在任何公式中该底数保持一致即可。

信息熵（3）

* 几个实例：
* (1) 取值确定的“随机变量”X，其概率分布{P[X=a]=1.其他概率为0}
* 熵H[X] = - (1log1+0log0) = 0！ $(\lim_{x\to 0} x\log x=0)$
* (2) 二元随机变量X：P[X=a]=p, P[X=b]=1-p
* 熵H[X] = $-p\log p-(1-p)\log(1-p)$
* 注：以上表达式常记为H[p,1-p]或H[p].
* (3) 具有概率分布$\{p_1,\ldots,p_n\}$的随机变量的熵H[X] = $-\Sigma_j p_j\log p_j$ 何时最大？
* 解：熵H[X]是多变量（$p_1,\ldots,p_n$）的函数，用Lagrange乘子算法求解带约束的优化问题
* $max\ -\Sigma_j p_j\log p_j$ s.t. $p_1+\ldots+p_n-1=0$
* 即引进乘子β，对函数$-\Sigma_j p_j\log p_j+\beta(p_1+\ldots+p_n-1)$求偏导得到极值方程
* $0=\partial(-\Sigma_j p_j\log p_j+\beta(p_1+\ldots+p_n-1))/\partial p_i=-\log p_i-1+\beta$，$i=1$，…，$n$
* 对（$p_1,\ldots,p_n$）求解该方程，得最优解$p^*_1=p^*_2=\ldots=p^*_n$，因此$p^*_1=p^*_2=\ldots=p^*_n=1/n$。
* 结论：在有n个取值的随机变量中，具有均匀概率分布的随机变量具有最大的信息熵，
* 并且该最大信息熵的数值（最大信息量）= $\log n$。
* 【思考】思考以上结论的含义；$\log n$的含义。

信息熵（4）

* 熵的基本性质：
* 熵$H[X] = -\Sigma_j p_j \log p_j$是多变量（$p_1,\ldots,p_n$）的凹函数
* 证明：
* 第一步：用基于二阶微分矩阵的判定准则，首先计算二阶导数：
* $\partial(H[X]/\partial p_i = \partial(-\Sigma_k p_k \log p_k)/\partial p_i = -(1+\log p_i)$
* $\partial^2(H[X]/\partial p_i\,\partial p_j = -\delta_{ij}/p_i$
* 第二步：计算二次型并判定是否正定或负定：
* $\sum_{i,j=1}^{n} u_i\, u_j\, \partial^2(H[X]/\partial p_i\,\partial p_j = -\sum_{i,j=1}^{n} u_i\, u_j\, \delta_{ij}/p_i = -\sum_{i=1}^{n} u_i{}^2/p_i \leqq 0$
* 因此熵是概率分布参数的凹函数。

* 思考题（1）以上论证有问题吗？答案：有！
* （2）针对存在的问题，如何改进上述的论证？
* （3）你如何保持以上论证、以最简洁（完全依据概念而非计算性）的方式完成改进？

信息熵（5）

* 联合熵
* 随机变量X与Y的联合熵
* $$H[X,Y] = -\Sigma_{x,y}\, p(x,y)\log p(x,y)$$
* ⊕ 联合熵是同联合随机变量(X,Y)相关联的一个非负数，更确切地说，是同联合概率分布P(X,Y)相关联的一个非负数：两个相同的概率分布，无论其实际含义是否相同，具有相同的信息熵。
* ⊕ 从函数的观点看，联合熵H是mn元变量（$p_{x1,y1}, p_{x1,y2},\ldots, p_{xm,y,n-1},p_{xm,yn}$）的非负值函数。
* ⊕ 联合熵是对联合随机变量(X,Y)所蕴含的"信息量"大小的度量。
* ⊕ 联合熵无量纲。
* **特殊的情形**
* 当X和Y是概率独立的随机变量，即P[X,Y] = P[X]P[Y]，联合熵
* $$H[X,Y] = H[X] + H[Y]$$

【思考】以上公式的含义

信息熵（6）

- 联合熵：一般的情形
- 根据条件概率和联合概率的关系P[X,Y]=P[Y|X]P[X]重写联合熵的表达式，有
- $H[X,Y] \equiv -\Sigma_{x,y}\, p(x,y)\log p(x,y)$
- $= -\Sigma_{x,y}\, p(x,y)\log(p(y|x)p(x))$
- $= -\Sigma_{x,y} p(x,y)\log p(y|x) - \Sigma_x\{\Sigma_y\, p(x,y)\}\log p(x)$
- $= -\Sigma_{x,y} p(x,y)\log p(y|x) - \Sigma_x p(x)\log p(x)$
- $\equiv H[Y|X] + H[X]$
- 以上推导引出定义：随机变量(X,Y)的条件熵
- $H[Y|X] \equiv -\Sigma_{x,y} p(x,y)\log p(y|x)$
- 以上计算同时也导出了联合熵的递归关系（链公式）：
- $\boxed{H[X,Y] = H[Y|X] + H[X] = H[X|Y] + H[Y]}$

信息熵（7）

* $(X_1,\dots,X_n)$的联合熵
*
* $H[X_1,\dots,X_n] \equiv -\Sigma_{x1,\dots,xn}\, p(x_1,\dots,x_n) \log p(x_1,\dots,x_n)$
*
* 反复应用前述二元联合变量的递归公式，就导出更普遍的递归公式
* $H[X_1,\dots,X_n] = H[X_1,\dots,X_{n-1} \mid X_n] + H[X_n]$
* $= H[X_1,\dots,X_{n-2} \mid X_{n-1}, X_n] + H[X_{n-1} \mid X_n] + H[X_n]$
* $= \dots\dots$
* $= H[X_1 \mid X_2,\dots X_{n-1}, X_n] + H[X_2 \mid X_3,\dots X_{n-1}, X_n] + \dots + H[X_{n-1} \mid X_n] + H[X_n]$

（参见教程2.5节）

如果任何一对随机变量X_i、X_j彼此独立，则

$$H[X_1,\dots,X_n] = H[X_1] + \dots + H[X_n]$$

互信息量（1）

- 随机变量X和Y的互信息量

$$I(X;Y) \equiv \Sigma_{x,y}\, p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

- 互信息量仅针对两个随机变量：单一一个随机变量不存在互信息量，多于两个的随机变量，目前尚未发现好的互信息量的定义。
- 互信息量的性质
- （1）对称性 $I(X;Y) = I(Y;X)$
- (2) 非负性 $I(X;Y) \geq 0$
- 证明 第一步：验证函数$f(t) = t\log t$是t的凸函数（例如用二阶微分判别法）。
- 第二步：证明对概率分布$P\{p_1,\ldots, p_N\}$和$Q\{q_1,\ldots, q_N\}$，$D(P||Q) \equiv \Sigma_i\, p_i\log(p_i/q_i) \geq 0$
- 并且仅当P=Q时D(P||Q) =0：
- 这是因为 $\Sigma_i\, p_i\log(p_i/q_i) = \Sigma_i\, q_i(p_i/q_i)\log(p_i/q_i) \geq \log(\Sigma_i\, q_i(p_i/q_i))$（为什么？
- 提示：想想如何在此运用第一步的结论*）$= \log(\Sigma_i\, p_i) = \log 1 = 0$.
- 第三步：根据第二部的结论，得到 $I(X;Y) \geq 0$。
- * 关键的在于：凸函数的定义等价于 $f(\Sigma_i t_i x_i) \leqq \Sigma_i\, t_i\, f(x_i)$,其中 $t_1,\ldots,t_n$非负且$\Sigma_i t_i x_i=1$，并
- 注意任何这样一组实数都可等价于一个概率分布！
- 其他的推导：参阅2.6节。
- 思考：为什么仅当P=Q时D(P||Q) =0？这对互信息量意味着什么？

互信息量（2）

* 随机变量X和Y的互信息量
* $$I(X;Y) \equiv \Sigma_{x,y}\, p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$
* 互信息量的性质
* （3）互信息量和熵的关系 $I(X;Y) = H(X)+H(Y) - H(X,Y)$
* 证明： $I(X;Y) \equiv \Sigma_{x,y}\, p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$
* $= \Sigma_{x,y}\, p(x,y)\log p(x,y) - \Sigma_{x,y}\, p(x,y)\log p(x) - \Sigma_{x,y}\, p(x,y)\log p(y)$
* $= \Sigma_{x,y}\, p(x,y)\log p(x,y) - \Sigma_{x,y}\, p(x)\log p(x) - \Sigma_{x,y}\, p(y)\log p(y) = -H(X,Y)+H(X)+H(Y)$
* (4)互信息量和熵的其他关系
* $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
* 证明：将前面H(X,Y)的递归公式代入这里。
* (5) X和Y的联合熵在X和Y概率独立时最大 $H(X,Y) \leqq H(X)+H(Y)$
* 证明：将 $I(X;Y) \geq 0$代入(3)。
* (6)条件熵总不会超过熵 $H(X) \geq H(X|Y)$，$H(Y) \geq H(Y|X)$
* 证明：将 $I(X;Y) \geq 0$代入(4)。
*
* 习题一：思考以上结论的合理性 习题二：X和Y的互信息量在什么情况下等于零?

互信息量（3）

* 随机变量X和Y的互信息量

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

* 根据I(X;Y)的性质深入理解互信息量的含义
* （1）性质 I(X;Y) = H(X)+H(Y) – H(X,Y) 和 H(X,Y) ≦ H(X)+H(Y) 表明：
* 注意到 H(X,Y) = (X,Y)的实际不确定性：H(X)+H(Y) = (X,Y)的最大不确定性，
* 因此互信息量I(X;Y)度量的是(X,Y)的实际不确定性相对于最大不确定性的亏损。
* 进一步的思考：
* 什么因素导致(X,Y)的不确定性下降？是因为X和Y存在概率性的相关性，
* 即P(X,Y)可能不等于P(X)P(Y)，因此：
* 互信息量I(X;Y)度量X的信息量在多大程度上蕴含Y的信息量。

注意由于对称性I(X;Y)=I(Y;X), 互信息量I(X;Y)同样度量Y的信息量在多大程度上蕴含X的信息量。

通信领域的例子：

X: 发射机信号；Y：接收机信号

Y和X不独立，但由于存在噪声干扰，Y和X存在概率性的关联。

* （2） I(X;Y) = H(X) – H(X|Y) = H(Y) – H(Y|X)
* 习题：基于上面这个关系，从另一个角度思考互信息量的含义。

下一讲

* 互信息量函数的凸性与凹性
* Markov过程与数据处理不等式
* 小结