熵

2020年3月13日 21:08

1. 随机变量的自信息

信息是对不确定性的消除，不确定性越高，消除后则包含的信息量越大，概率越小则信息量就越大。概率越大则信息量就越小。同样，如果事件的概率为1，那么则是无信息量，如果概率趋近于0，那么信息量则无穷大。

那么，假如有两个独立事件，设为 a_1, a_2。

(1) 若 $P(a_1) > P(a_2)$，则 $f(a_1) < f(a_2)$

(2) 若 $P(a_1) = 1$，则 $f(a_1) = 0$

(3) 若 $P(a_1) = 0$，则 $f(a_1) = \infty$

(4) 若 a_1, a_2 相互独立，则 $f(a_1, a_2) = f(a_1) + f(a_2)$

通过上述数学逻辑的分析，可以发现，对数函数可以满足上述要求，可以记为 $I(a_i) = -\log P(a_i) = \log \frac{1}{P(a_i)}$

我们可以从数学上解释一下，设底数为k。

那么

$$k^{\frac{1}{P(a_i)}} = I(a_i)$$

可以认为是，用k进制的符号表示，共需要 $\frac{1}{P(a_i)}$ 位。那么这则与进制相对应，例如k取2时，则可以理解为bit。

以e为底，单位则为 nat (nature unit)，以10为底，单位为 Hart (Hartley)

$1\ \text{nat} = 1.44\ \text{bit}$，$1\ \text{Hart} = 3.32\ \text{bit}$，

$= \log_2 e\ \text{bit}$　　$= \log_2 10\ \text{bit}$　　$\log_a b = \frac{\log_c b}{\log_c a}$，$\log_e e = \frac{\log_2 e}{\log_2 e}$，

那么，我们现在还需要证明他的唯一性，即仅有这种形式可以满足需求。

2. 唯一性的证明

对于单一变量，并不存在信息量的准确数学形式，但对于一个随机变量X，它的信息量的度量却提出了额外的条件。这也是我们对信息论模型建立所需求的条件。

(1) 连续性条件：$f(p_1, p_2, \dots, p_M)$ 应是 p_n 的连续函数，当 p_n 稍微变化时，$f(p_1, \dots, p_M)$ 也应稍微变化。

(2) 等概时为单调增函数：$f(\frac{1}{N}, \dots, \frac{1}{N}) = g(N)$ 应是N的增函数

(3) 可加性条件：当随机变量的取值不是通过一次试验而是若干试验才最后得到时，随机变量在各次实验中的不确定度应该可加，且不影响最后的结果。即

$$f(p_1, p_2, \dots, p_M) = f((p_1+p_2+p_3+\dots+p_K), p_{K+1}, \dots, p_M) + (p_1+p_2+\dots+p_K) f(p_1', p_2', \dots, p_K')$$

其中

$$p_i' = \frac{p_i}{p_1+p_2+\dots+p_K}$$

，即归一化后的 p_i'，这个等式也是符合我们的直觉的。举一个便于理解的例子。

掷骰子，可以将1,2,3看做一个整体，那么整个掷骰子的信息量为 $g(\frac{1}{N})$，就等于 $f(\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$ 外加"1,2,3"的不确定度以及他们作为整体所发生的概率，这个概率可以理解为他们归一化的代价，那么就有 $\frac{1}{2} f(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$。

下面我们开始证明唯一性，首先，假设一个MN项的均匀分布。

$$g(MN) = f(\frac{1}{MN}, \frac{1}{MN}, \dots, \frac{1}{MN}),$$

我们利用可加性

$$g(MN) = f(\underbrace{\frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M}}_{M个}) + \sum_{i=1}^{M} \frac{1}{M} f(\underbrace{\frac{1}{N}, \frac{1}{N}, \dots, \frac{1}{N}}_{N个}) = g(M) + g(N)$$

$$g(s^m) = g(s \cdot s^{m-1}) = g(s) + g(s^{m-1}) \Rightarrow m\, g(s)$$

那么我们继续假设 取 $s, m, t, n \in N$，使 $s^m \le t^n < s^{m+1}$

由上述的条件②，我们可得

$$g(s^m) \le g(t^n) < g(s^{m+1}),$$

再根据上述的推导

$$m\, g(s) \le n\, g(t) < (m+1)\, g(s)$$

同除 $n \cdot g(s)$ 得

$$\frac{m}{n} \le \frac{g(t)}{g(s)} < \frac{m+1}{n}$$

上述公式变形（可根据数轴理解 $\frac{m}{n}$　$\frac{g(t)}{g(s)}$　$\frac{m}{n}+\frac{1}{n}$ ）

$$\left| \frac{m}{n} - \frac{g(t)}{g(s)} \right| < \frac{1}{n} \Longrightarrow 0 \le \frac{g(t)}{g(s)} - \frac{m}{n} < \frac{1}{n}$$

$$\left|\frac{m}{n}-\frac{g(t)}{g(s)}\right|<\frac{1}{n} \Longrightarrow 0\le \frac{g(t)}{g(s)}-\frac{m}{n}<\frac{1}{n}$$

对于底数大于1的对数函数，我们有

$$m\log s \le n\log t < (m+1)\log s,$$

那么同理，我们也可得到不等式

$$\left|\frac{m}{n}-\frac{\log t}{\log s}\right|<\frac{1}{n} \Longrightarrow 0\le\frac{\log t}{\log s}-\frac{m}{n}<\frac{1}{n}$$

将上式缩放，

$$\left|\frac{m}{n}-\frac{g(t)}{g(s)}\right|+\left|\frac{m}{n}-\frac{\log t}{\log s}\right|<\frac{2}{n} \Longrightarrow \left|\frac{\log t}{\log s}-\frac{g(t)}{g(s)}\right|<\frac{2}{n}.$$

这里利用了 $|a\pm b|\le|a|+|b|$。由几何图形我们还可以得到 < 号的结论

考虑到n可以任意大，那么n为无穷大时，上面的不等式也需满足，

$$\lim_{n\to\infty}\left|\frac{g(t)}{g(s)}-\frac{\log t}{\log s}\right|=0 \Rightarrow g(t)=C\log t.$$

上面我们考虑了均匀分布，下面考察X是非均匀分布，但是是有理数取值的情况。即 $X=P_X(x)$，这里可以简写为 $P(x)$

令 $P(n)=\frac{m_n}{\sum_{i=1}^{N} m_i}=\frac{m_n}{M}$（有理数可以用分数表示），那么 $g(M)=f(\frac{1}{M},\dots,\frac{1}{M})$，M个。

因为是有理数，可以将 $g(M)$ 由下分组 $g(M)=f(\underbrace{\frac{1}{M},\dots,\frac{1}{M}}_{m_1个},\underbrace{\frac{1}{M},\dots,\frac{1}{M}}_{m_2个},\dots,\underbrace{\frac{1}{M},\dots,\frac{1}{M}}_{m_N个})$

根据可加性的条件

$$g(M)=f\left(\frac{m_1}{M},\frac{m_2}{M},\dots,\frac{m_N}{M}\right)+\sum_{i=1}^{N}\frac{m_i}{M}g(m_i)$$

$$\Rightarrow f\left(\frac{m_1}{M},\dots,\frac{m_N}{M}\right)=g(M)-\sum_{i=1}^{N}\frac{m_i}{M}g(m_i)$$

$$=C\log M-\sum_{i=1}^{N}\frac{m_i}{M}C\log m_i$$

$$=C\left[\sum_{i=1}^{N}\frac{m_i}{M}\left(\log\frac{M}{m_i}\right)\right]=-C\sum_{i=1}^{N}p_i\log p_i,$$

我们可以发现，在均匀分布中也符合这个结论。
最后，我们考虑无理数的情况，无理数的证明可以利用有理数逼近，然后利用连续性条件可得证。
于是，我们已经讨论了所有情况下的数学形式，从而证明了不确定度的唯一性，即

$$f(P_X(x))=-\sum_{i=1}^{N}p_i\log p_i$$

3. 信息熵

不确定度的数学形式和热力学熵的数学形式是一致的，因为不确定度也被定义成为信息熵。在推导F-D分布中（固体物理初步的最后一部分），我们简单提到了熵的概念，也就是从有序到无序化，也就是熵增的。而我们所处世界的信息也是自然界的一部分，从无噪声逐渐变为有噪声，这也符合熵增的理论。所以用熵来定义是很合理的，数学上我们也证明了这一点。

在热力学中有一个著名的谬论即麦克斯韦妖谬论，Landauer 提出："麦克斯韦妖在减少绝热系统熵的时候所需要的信息熵不少于热力学熵的减少量"，从而解决了这个谬论。这个例子也说明了信息熵和热力学熵的一致性。

3.1 定义

我们这里用较为严格的数学语言定义他：

离散随机变量X的信息熵H(X)定义为

$$H(X)=-\sum_{x\in X}p(x)\log p(x)$$

对于 $0\log 0$，$\lim_{x\to 0}x\log x=0$，零概率事件不影响信息熵。而均匀分布时，信息熵达最大值。

3.2 联合熵

对于一维的随机变量X，我们定义了H(X)，很自然的我们可以推广到联合的随机事件，那么就得到了联合熵

例如对于一对离散随机变量(X,Y)的联合熵定义为

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

例如X,Y，对于联合分布概率$p(x,y)$的联合熵定义为

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

联合熵的大小与两个随机变量的信息熵之和不一定是相等的。如果两个事件相互独立，那么则相等。即$H(X,Y)=H(X)+H(Y)$，这个在数学上很好证明，只需要将$P(x,y)=P(x)+P(y)$代入即可。

如果两个事件不独立的话，则存在一定的相关性（需要注意的是这与相关系数没有关系，相关系数表示的是线性相关），那么$H(X,Y)<H(X)+H(Y)$，因为当获取到X时，就已经获取到一部分Y，相当于减小了Y的信息量，所以联合熵会偏小。

而如果我们想让等号成立的话，$P(x,y)=P(x)\cdot P(y|x)$或$P(x,y)=P(y)\cdot P(x|y)$则一直成立。这时我们可以引出条件熵的概念。

3.3 条件熵

我们希望得到$H(X,Y)=H(X)+H(Y|X)=H(Y)+H(X|Y)$这样的结果，那么我们可以根据条件概率来获得条件熵的数学形式。

$$H(X,Y)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(x,y)\quad,\quad \text{代入}\ P(x,y)=P(x)\cdot P(y|x)$$

$$=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\left[\log p(x)+\log p(y|x)\right]$$

$$=-\sum_{x\in X}\sum_{y\in Y}p(x)\cdot p(y|x)\cdot\log p(x)-\sum_{x\in X}\sum_{y\in Y}p(x)\cdot p(y|x)\log p(y|x)$$

$$=-\sum_{x\in X}p(x)\log p(x)\left[\sum_{y\in Y}p(y|x)\right]-\sum_{x\in X}\sum_{y\in Y}p(x)p(y|x)\log p(y|x)\quad,\quad\sum_{y\in Y}p(y|x)=1.$$

那么第一项则为$H(X)$，第二项则为$H(Y|X)$。

我们给出他们的精确定义。

若$(X,Y)\sim p(x,y)$，则条件熵定义为$H(Y|X)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(y|x)=-\sum_{x\in X}p(x)\sum_{y\in Y}p(y|x)\log p(y|x)$。

我们可以把$-\sum_{y\in Y}p(y|x)\log p(y|x)$记做$H(Y|X=x)$，

而$H(Y|X)$并不是$H(Y|X=x)$，需要注意。

有了最基础的条件熵的概念，我们可以将结论进行推广。

① $H(X,Y|Z)=H(X|Z)+H(Y|X,Z)$

简单证明一下，$H(X)=-E(\log p(x))$

$P(x,y|z)=P(x|z)\cdot P(y|x,z)$

$\log P(x,y|z)=\log P(x|z)+\log P(y|x,z)$

$\Rightarrow H(X,Y|Z)=H(X|Z)+H(Y|X,Z)$

[说明]通过对数函数，则使概率分布的数学结论从乘法变为加法。而信息熵则可以认为是对其的加权平均。这个加权平均都可以认为是联合概率的加权平均，所以类似的结论都是成立的，这样则满足了可加性。联合观察必须等于逐次观察。

② 设随机变量$X_1,X_2,\cdots,X_n$满足分布$p(x_1,x_2,\cdots,x_n)$，则

$H(X_1,X_2,\cdots,X_n)=\sum_{i=1}^{n}H(X_i|X_{i-1},X_{i-2},\cdots,X_1)$

这也则是可加性。

在通信中，$X_1,\cdots,X_n$通常也是依次到达的，我们也可以根据可加性做次序性的检测，每次只观察最新到达的信息即可。

③ 若X,Y相互独立，那么$P(x,y)=P(x)P(y)$，就有$H(X,Y)=H(X)+H(Y)$，之前也证明过。

3.4 信息熵的性质

1). 对称性

$H(P_1,P_2,\cdots,P_N)=H(P_{k(1)},P_{k(2)},\cdots,P_{k(n)})$

信息熵只关乎随机分布的概率，并不关于信息意义本身，这也是经典信息论的核心思路

2). 非负性

熵的最小值为0，即为概率为1的时候

3) 可加性

$H(X,Y)=H(X)+H(Y|X)$，即为上述的"③"

4). 条件减小熵

$H(X|Y)\le H(X)$，

这也告诉我们了有些……条件对相关……变量……也可以减小不确定性……统计……减小不确定……

4). 条件减小熵

$H(X|Y) \le H(X)$，

这样会符合我们的直觉，知道了统计相关性的变量，则可以减小不确定性。统计平均意义上减少不确定性但是针对某一个Y取值不一定。从哲学上来说，事物是广泛关联的。但是不意味着 $H(X|Y=y) \le H(X)$

因为对于具体的取值来说，有可能使 $P(X|y)$ 的概率分布减小从而使信息量增大。例如不可能事件。

5) 离散随机变量X在等概率分布时，熵取得最大值，也就是熵增的。之所以叫做熵。用pm选一定满足这个结论的。

我们这里要给出数学上的证明。即

$$H(p_1, p_2, \cdots, p_N) \le H(\tfrac{1}{N}, \tfrac{1}{N}, \cdots, \tfrac{1}{N}) = \log N = \log |X| = G(N)$$

我们可以借助数学的概念——凸集，这部分内容会放在数学部分。这里就直接应用其数学结论了。

证明过程如下：

我们已经推导了概率分布是具有凸性的，那么仅需证明 $H(x)$ 是下凸函数即可。

$$H(\lambda\vec{p_1} + (1-\lambda)\vec{p_2}) - \lambda H(\vec{p_1}) - (1-\lambda)H(\vec{p_2})$$

$$= -\sum_{n=1}^{N}[\lambda p_{1n} + (1-\lambda)p_{2n}]\log[\lambda p_{1n} + (1-\lambda)p_{2n}] + \lambda\sum_{n=1}^{N} p_{1n}\log p_{1n} + (1-\lambda)\sum_{n=1}^{N} p_{2n}\log p_{2n}$$

$$= \sum_{n=1}^{N}\lambda p_{1n}\log\frac{p_{1n}}{\lambda p_{1n} + (1-\lambda)p_{2n}} + \sum_{n=1}^{N}(1-\lambda)p_{2n}\log\frac{p_{2n}}{\lambda p_{1n} + (1-\lambda)p_{2n}}$$

由 $\lim_{x\to 0}\frac{\log_a(1+x)}{x} = \lim_{x\to 0}\log_a(1+x)^{\frac{1}{x}} = \log_a e = 1/\ln a \Rightarrow \log x \sim \frac{1}{\ln 2}(x-1)$ 或 $\log x \le \frac{1}{\ln 2}(x-1)$

代入上式　　　　$\log\frac{1}{x} \ge \frac{1}{\ln 2}(1-x)$

$$\ge \sum_{n=1}^{N}\lambda p_{1n}\cdot\frac{1}{\ln 2}\left[1 - \frac{\lambda p_{1n} + (1-\lambda)p_{2n}}{p_{1n}}\right] + \sum_{n=1}^{N}(1-\lambda)p_{2n}\cdot\frac{1}{\ln 2}\left[1 - \frac{\lambda p_{1n} + (1-\lambda)p_{2n}}{p_{2n}}\right]$$

$$\ge \frac{1}{\ln 2}\sum_{n=1}^{N}\lambda p_{1n} + (1-\lambda)p_{2n} - (\lambda + 1-\lambda)[\lambda p_{1n} + (1-\lambda)p_{2n}]$$

$$\ge 0.$$

于是我们证明了熵函数具有凸性，即 $H(p)$ 是上凸函数，那么我们可以直接利用上凸函数的 Jensen 不等式

$$H(E(X)) \ge E[H(X)]$$

等号在等概时取到最大值

3.1 从离散到连续

这件事我们的概率论与数理统计已经帮我们完成了，我这里用黎曼积分的方式去求解，即所谓的定积分

$$H(X) = -\sum_k p(x_k)\Delta x\log p(x_k)\Delta x = -\sum_k p(x_k)\log p(x_k)\cdot\Delta x - \sum_k[p(x_k)\log\Delta x]\cdot\Delta x.$$

$$= -\int p(x)\log p(x)\,dx - \int p(x)\lim_{\Delta x\to 0}\log\Delta x\,dx = h(X) + \infty$$

这也与直觉一致，例如实数的信息量无穷大，无论用多少个比特我们都无法表示清楚。

香农针对这一问题提出了微分熵的概念

微分熵　　$h(x) = \int_{-\infty}^{+\infty} -p(x)\log p(x)\,dx.$

从这个定义还可以推广得到：

联合熵　　$h(X,Y) = -\iint p(x,y)\log p(x,y)\,dx\,dy.$

条件熵　　$h(X|Y) = -\iint p(x,y)\log p(x|y)\,dx\,dy = -\int p(y)\int p(x|y)\log p(x|y)\,dx\,dy.$

以及一系列关系

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

$$h(X|Y) \le h(X), \quad h(Y|X) \le h(Y), \quad h(X,Y) \le h(X) + h(Y)$$

但是微分熵可能为负值，因为他并不是信息熵。例如.

$$h(X|Y) \le h(X), \quad h(Y|X) \le h(Y), \quad h(X,Y) \le h(X) + h(Y)$$

但是微分熵可能为负值，因为他并不是信息熵，例如：

$$X: p(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b, \quad b-a<1 \\ 0, & x>b, \quad x<a \end{cases}$$

$$\text{则 } h(X) = -\int_a^b \frac{1}{b-a}\log\frac{1}{b-a}dx = \log(b-a) < 0.$$

我们可以发现连续随机变量 $p(x)$ 并不一定小于等于1.

还有一个重要的分布是高斯分布

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right].$$

那么我们直接代入公式计算

$$h(X) = -\int_{-\infty}^{+\infty} p(x)\log\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]\right\}dx.$$

$$= -\int_{-\infty}^{+\infty} p(x)\log\frac{1}{\sqrt{2\pi\sigma^2}}dx - \int_{-\infty}^{+\infty} p(x)\log\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]dx$$

$$= \log\sqrt{2\pi\sigma^2} + \log^e\int_{-\infty}^{+\infty} p(x)\frac{(x-m)^2}{2\sigma^2}dx.$$

$$= \log\sqrt{2\pi\sigma^2} + \frac{1}{2}\log^e = \frac{1}{2}\log 2\pi e\sigma^2$$

这个公式的重要性会在信道容量中体现，连续变量若是高斯分布，那么它所带来的信息量与均值无关，只与方差有关。在通信系统中，方差和功率谱密度是紧密相关的，那么一个连续的高斯分布的随机变量只与他的功率有关。

由此我们还有一个定理

若给定连续随机变量 X 的均值 m 与方差 σ^2，则当其服从高斯分布时，微分熵最大。

证明如下：

设 $p(x)$ 为满足均值为 m，方差为 σ^2 的高斯分布概率密度函数（PDF），设为随机变量 Y.

$q(x)$ 为满足均值为 m，方差为 σ^2 的任意分布的 PDF，设为随机变量 Z.

我们直接做差，这实际上也体现了信息熵之间的距离的思想，各类读者能多体现其思路

$$H(Z) = -\int_{-\infty}^{\infty} q(x)\log q(x)\,dx = H(Y) - H(Y) + H(Z)$$

$$= H(Y) - [H(Y) - H(Z)]$$

$$H(Y) - H(Z) = -\int_{-\infty}^{+\infty} q(x)\log q(x)\,dx - \int_{-\infty}^{+\infty} p(x)\log p(x)\,dx.$$

对于高斯分布，我们可以证明

$$-\int_{-\infty}^{+\infty} q(x)\log p(x)\,dx = -\int_{-\infty}^{+\infty} q(x)\log\frac{1}{\sqrt{2\pi\sigma^2}}dx + \log e\int_{-\infty}^{+\infty} q(x)\frac{(x-m)^2}{2\sigma^2}dx = \frac{1}{2}\log 2\pi e\sigma^2 = H(Y)$$

于是

$$H(Y) - H(Z) = \int_{-\infty}^{+\infty} q(x)\log\frac{p(x)}{q(x)} \le \int_{-\infty}^{+\infty} q(x)\left(\frac{p(x)}{q(x)} - 1\right)dx = 0.$$

于是得证。

我们将 $\int^{+\infty} q(x)\log\frac{p(x)}{q(x)}$ 记为 $-D(q(x)\|p(x))$，即为相对熵的负数.

得证。

我们将 $\int_{-\infty}^{+\infty} q(x)\log\frac{p(x)}{q(x)}$ 记为 $-D(q(x)\|p(x))$，即为距离的负数.

这也说明了、$H(X)=\log|X|-D(p\|u)$，对于连续变量的推广。这个定理也与热力学第二定律有相同的地方. 值得深思