信息熵

冯裕祺　李海豹　李越　东北大学理学院

2021 年 5 月 19 日

前言

前言

信息是个很抽象的概念。人们常常说信息很多，或者信息较少，但却很难说清楚信息到底有多少。比如一本五十万字的中文书到底有多少信息量。直到 1948 年，香农提出了“信息熵”的概念，才解决了对信息的量化度量问题。信息熵这个词是 *C.E.Shannon*（香农）从热力学中借用过来的。热力学中的热熵是表示分子状态混乱程度的物理量。香农用信息熵的概念来描述信源的不确定度。信息论之父克劳德 · 艾尔伍德 · 香农第一次用数学语言阐明了概率与信息冗余度的关系。

熵的引出

在数学中，确定性过程是十分平常的，例如一个因变量的值在回归方程确定之后是可以唯一的求解出来的，例如：$y = b + \beta x$，在 b 和 β 给定之后，每一个 y 对每一个 x 是唯一确定的。

与确定性事件相反，随机事件在生活中也是很常见的，最简单的例如投掷硬币的正反面，这就是一个不确定事件。

不确定性作为一种自然的属性，应当如何使用数学的语言去描述刻画这一性质？这就引出了我们的熵的概念。

熵的历史

熵的历史

熵就是关于不确定性的一个极好的数学描述。历史上的熵概念起源于热力学。凡是学过热力学、统计物理或物理化学的人对“熵”这一术语都不陌生，但是这一概念发展的初始阶段却跟混沌思想并无任何历史瓜葛。实际上，当熵的名词诞生之时，混沌之祖庞加莱（Henri Poincare, 1854-1912）还只是一个乳臭未干的少年。当熵的触角从宏观的热力学伸展到微观的统计力学之后，才逐渐拉近它和混沌概念的距离。二十世纪中叶的一场信息论革命，无意中在古典熵的旧作坊内又酿造出醇香的新酒。

信息熵的出现

直到 1948 年，**香农**发表了具有历史意义的论文（A mathematical theory of communications,1948），开创了现代信息理论的先河。提出的**信息熵**在 ** 数学上量化了通讯过程中的“信息缺失”的统计本质，具有划时代的意义”

熵的本质

香农的信息熵本质上是对我们司空见惯的“不确定现象”的数学化度量。譬如说，如果天气预报说“今天中午下雨的可能性是百分之九十”，我们就会不约而同想到出门带伞；如果预报说“有百分之五十的可能性下雨”，我们就会犹豫是否带伞，因为雨伞无用时确是累赘之物。显然，第一则天气预报中，下雨这件事的不确定性程度较小，而第二则关于下雨的不确定度就大多了。

熵的推导

设有 n 个基本事件，各自出现的概率是 $(p_1, p_2, \ldots, p_n)$，则他们构成了一个样本空间。例如抛硬币的样本空间是 $(\frac{1}{2}, \frac{1}{2})$，我们用符号 H 来表示样本空间的不确定度。如果这时候这个硬币不是质量均匀的，他的正反面概率分别是 $(\frac{7}{10}, \frac{3}{10})$。我们会很明显的通过直觉判断出 $H(\frac{1}{2}, \frac{1}{2}) > H(\frac{7}{10}, \frac{3}{10})$

更一般的，如果用 $H(p_1, p_2, ..., p_n)$ 记为样本空间所对对应的不确定性，通过直觉我们可以知道当所有事件等可能时，即 $P(p_n) = \frac{1}{n}$，这个时候其不确定性最大。因此满足基本不等式

1

$$\mathrm{H(p1, p2, \ldots, pn)} \leq \mathrm{H(1/n, 1/n, \ldots, 1/n)}$$

如果我们不抛硬币，而是进行掷骰子，假设骰子是均匀的，那么每一面向上的概率都是 $\frac{1}{6}$，我们稍加思索便知掷骰子的不确定性是大于扔硬币的，因此我们退出不确定性函数应该满足单调性的要求

2

$$H(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$$

是自然数 n 严格递增函数。

假设物理系赵教授、数学系钱教授和孙教授竞争理学院的一笔科研基金，他们每人申请成功的概率分别为 1/2、1/3、1/6。院长为求公平，让每个系得此奖励的机会均等。若物理系拿到资助，就到了赵教授的名下。如数学系得到了它，钱教授有 2/3 的概率拿到，孙教授则有 1/3 的机会到手。通过分析“条件概率”，我们能得出不确定度 H(1/2, 1/3, 1/6) 的数值：这三个教授获得基金的不确定度，等于物理系或数学系拿到这笔基金的不确定度，加上数学系赢得该基金的概率与在数学系拿到基金的条件之下，钱教授或孙教授得到它的不确定度之乘积。换言之，$H(1/2, 1/3, 1/6) = H(1/2, 1/2) + ½ H(2/3, 1/3)$。推而广之，可以得出不确定度与条件概率有关的“**加权和**”性质：

3 如果一个不确定事件分解成几个持续事件，则原先事件的不确定度等于持续事件不确定度的加权和。

既然我们想用一个漂亮的数学公式来表达不确定度这一样本空间概率值函数，我们自然希望这个函数表达式和几乎所有的物理公式一样连续依赖于公式中的所有变元。这样，第四个条件就自然而然地加在了不确定度函数的头上：

4. 对固定的自然数 n，不确定度函数 H 是 $(p_1, p_2, ..., p_n)$ 的一个连续函数。

香农无需什么高深的数学，甚至连微积分都可不要，就证明了：任何在所有样本空间上都有定义的函数 H，只要它满足以上的“三项基本原则 (2)(3)(4)”，就非如下的表达式莫属：

$$H(p_1, p_2, \ldots, p_n) = -C(p_1 \ln p_1 + p_2 \ln p_2 + \ldots + p_n \ln p_n)$$

香农将常数 C 取-1，得到了我们的香农熵。

$$H(p_1, p_2, ..., p_n) = -(p_1 \ln p_1 + p_2 \ln p_2 + ... + p_n \ln p_n)$$

按照冯 · 诺伊曼的建议，**该函数被定义为样本空间** $(p_1, p_2, ...p_n)$ **所对应的信息熵**。

公式证明及推导

公式证明及推导

第一步：

把 H(1/n, 1/n, ..., 1/n) 记为 A(n)。设 n = 8。我们屡次应用上述条件（3）来论证公式 A(2^3) = 3A(2):

A(2^3) = A(2) + [2^(-1)A(2) + 2^(-1)A(2)] + [4^(-1)A(2) + 4^(-1)A(2) + 4^(-1)A(2) + 4^(-1)A(2)] = A(2) + A(2) + A(2) = 3A(2)。

运用数学归纳法就得到

A(s^m) = A(s)+s[s^(-1)A(s)]+...+s^(-m+1)[s^(-(-m+1))A(s)] = m A(s)。(a)

现在假设四个正整数 t,s,n,m 满足不等式 s^m <= t^n < s^(m+1) 。

求对数，有 m ln s <= n ln t < (m+1)ln s，即

m/n <= ln t / ln s < m/n + 1/n。

因而我们得到不等式 | m/n – ln t / ln s | < 1/n。(b) 由熵的条件 (2)，A(k) 是 k 的递增函数。故条件（a）推出 m A(s) <= n A(t) < (m+1)A(s)，继而有 | m/n – A(t) / A(s) | < 1/n。(c) (b) 和 (c) 保证了 | A(t) / A(s) – ln t / ln s | < 2/n。既然 n 是任意的，就有等式 A(t) / A(s) = ln t / ln s，或言之，A(t) / ln t A(s) / ln s。故存在常数 C（取为 1）使得对所有正整数 t，A(t) = C ln t = ln t。因此

$$H(\frac{1}{n}, \dots \frac{1}{n}) = -\sum \frac{1}{n} ln(\frac{1}{n})$$

即熵公式（H）在 p1 = p2 = ... = pn = 1/n 时成立。

第二步：我们现在证明公式（H）对所有和为 1 的正有理数 pi 都对。我们先用 p1 = 1/2, p2 = 1/3, p3 = 1/6 来阐述证明的思想

根据熵条件（3），

H(1/6, ..., 1/6) = H(1/2, 1/3, 1/6) + 2^(-1)H(1/3, 1/3, 1/3) + 3^(-1)H(1/2, 1/2) + 6^(-1)H(1)。

所以，

H(1/2, 1/3, 1/6) = H(1/6, ..., 1/6) – 2^(-1)H(1/3, 1/3, 1/3) – 3^(-1)H(1/2, 1/2) - 6^(-1)H(1)。

如此的分解是为了用到第一步的结果。如果注意到有理数的分数形式

p1 = 1/2 = 3/(3+2+1) = n1/(n1+n2+n3),

p2 = 1/3 = 2/(3+2+1) = n2/(n1+n2+n3),

p3 = 1/6 = 1/(3+2+1) = n3/(n1+n2+n3),

上述的分解就能写成

H(p1, p2, p3) = A(n1+n2+n3) − [p1A(n1) + p2A(n2) + p3A(n3)]。

同样的道理用到一般情形 p1, p2, ..., pk： **设**

pi = ni / (n1 + ... + nk), i = 1, 2, ..., k,

则有等式

H(p1, p2, ..., pk) = A(n1+...+nk) - [p1A(n1) + ... + pkA(nk)]。

由上面的第一步， A(n) = ln n。**代入到上式，给出**

H(p1, p2, ..., pk) = ln(n1 + ... + nk) − (p1ln n1+...+pkln nk) = (p1+...+pk) ln(n1+...+nk) − (p1ln n1+...+pkln nk) = −[p1ln(n1 / (n1+...+nk))+...+pkln(nk/(n1+...+nk))]= −(p1ln p1+...+pkln pk)。

第三步

既然熵公式（H）对所有和为 1 的所有正有理数成立，连续性条件（4）推出它对所有和为 1 的非负实数成立。这就完成了证明。

香农熵的变形

香农熵的变形

联合熵（*joint entropy*），如果 X, Y 是一对离散型随机变量，$X, Y - p(x, y)$,,X, Y 的联合熵 $H(X, Y)$ 为：

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y)$$

联合熵实际上就是描述一对随机变量平均所需要的信息量。

相对熵（K-L 距离）

考虑某个未知的分布 p(x)，假定我们已经使用一个**近似的分布 *q(x)*** 对它进行了建模。如果我们使用 q(x) 来建立一个编码体系，用来把 x 的值传给接收者，那么，由于我们使用了 q(x) 而不是**真实分布 *p(x)***，因此在具体化 x 的值（假定我们选择了一个高效的编码系统）时，我们需要一些附加的信息。我们需要的平均的附加信息量（单位是 nat ）为

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x})\ln q(\mathbf{x})\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x})\ln p(\mathbf{x})\mathrm{d}\mathbf{x}\right)\\ &= -\int p(\mathbf{x})\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x}.\end{aligned}$$

两个概率分布的相对熵：

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

可以把 Kullback-Leibler 散度（KL 散度之所以不说距离，是因为不满足对称性和三角形法则）。看做两个分布 p(x) 和 q(x) 之间不相似程度的度量。相对熵常被用以衡量两个随机分布的差距。当两个随机分布相同时，其相对熵为 0。当两个随机分布的差别增加时，其相对熵也增加。当 q=p 时，该度量的结果是 0，而其它度量的结果为正值。直观上，它度量了使用 q 而不是 p 的压缩损失（以二进制）的程度。

假设数据通过未知分布 p(x) 生成，我们想要对 p(x) 建模。我们可以试着使用一些参数分布 $q(x \mid \theta)$ 来近似这个分布。$q(x \mid \theta)$ 由可调节的参数 θ 控制 (例如一个多元高斯分布)。一种确定 θ 的方式是最小化 p(x) 和 $q(x \mid \theta)$ 之间关于 θ 的 Kullback-Leibler 散度。我们不能直接这么做，因为我们不知道 p(x) 。但是，假设我们已经观察到了服从分布 p(x) 的有限数量的训练点 x n ，其中 $n = 1, \dots, N$ 。那么，关于 p(x) 的期望就可以通过这些点的有限加和，

$$\mathrm{KL}(p \| q) \simeq \frac{1}{N} \sum_{n=1}^{N} \{-\ln q(\boldsymbol{x}_n \mid \boldsymbol{\theta}) + \ln p(\boldsymbol{x}_n)\}$$

公式右侧的第二项与 θ 无关，第一项是使用训练集估计的分布 $q(x \mid \theta)$ 下的　的负对数似然函数。因此我们看到，最小化 Kullback-Leibler 散度等价于最大化似然函数。

交叉熵 (cross entropy)

如果一个随机变量 X $p(x)$,$q(x)$ 为用于近似 $p(x)$ 的概率分布，那么随机变量 X 和模型 q 之间的交叉熵定义为:

$$\begin{aligned} H(X,q) &= H(X) + D(p\|q) \\ &= -\sum_{x} p(x)\log q(x) \end{aligned}$$

交叉熵的概念用以衡量估计模型与真实概率分布之间的差异。

互信息

两个随机变量 X, Y 的互信息，定义为 X, Y 的联合分布和独立分布乘积的相对熵。

$$\mathrm{I(X,Y) = D(P(X,Y) \| P(X)P(Y))}$$
$$I(X,Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

证明过程

$$
\begin{aligned}
& H(X) - I(X, Y) \\
& = -\sum_x p(x) \log p(x) - \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} \\
& = -\sum_x \left(\sum_y p(x, y)\right) \log p(x) - \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} \\
& = -\sum_{x,y} p(x, y) \log p(x) - \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} \\
& = \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(y)} \\
& = \sum_{x,y} p(x, y) \log p(x \mid y) \\
& = H(X \mid Y)
\end{aligned}
$$

参考文献

参考文献

1. "Entropy - an introduction," Jiu Ding and Tien-Yien Li, NankaiSeries in Pure and Applied Mathematics and Theoretical Physics, Volume 4, WorldScientific, 26-53, 1993.
2. Information theory and statistical physics, Physics Review 106(4), 620-630, 1957; Information theory and statistical physics, Physics Review 108(2), 171-190, 1957
3. L.R. Mead and N. Papanicolaou, Maximum entropy in the problem of moments, J. Math. Phys. 25, 2404–2417, 1984.
4. J. Ding, C. Jin, N. Rhee, and A. Zhou, "A maximum entropy method based on piecewise linear functions for the recovery of a stationary density of interval mappings,'' J. Stat. Phys. 145, 1620-1639, 2011.
5. 丁玖，信息熵是怎样炼成的｜纪念信息论之父香农，返朴，2019.4.30