数据科学基础
Foundations of Data Science

6.1 数据热力学-熵

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

热力学

- 熵的概念最早起源于物理学，用于度量一个热力学系统的无序程度。
- 热力学系统从一个平衡态到另一平衡态的过程中，其熵永不减少：若过程可逆，则熵不变；若不可逆，则熵增加。
- 热传导过程不可逆：孤立系统自发地朝着热力学平衡方向（最大熵状态）演化。
- 玻尔兹曼关系是对熵的微观（统计意义的）解释，表述为：系统的熵S与其微观状态数Ω存在函数关系 $S = k \ln \Omega$，其中 k为玻尔兹曼常数。
- 玻尔兹曼关系给出了熵的微观解释：系统微观粒子的无序程度的度量。对熵这一概念引入信息论、生态学等其他领域具有深远意义。

数据热力学

- 1948年，C. E. Shannon（香农）发表信息论奠基性论文《A Mathematical Theory of Communication》。
- 香农：“信息是用来消除随机不确定性的东西。”
- 关于热力学熵与信息论熵的关联，曾是一个长期争论的问题。
- 热力学熵用来度量热力系统中微观状态的无序性。
- 信息论熵用来度量随机系统中的消息不确定性。

信息熵

信息论之父香农给出的信息熵的三个性质：

1. 单调性：确定性越高的事件的信息量越低；
2. 非负性：非负性是从随机性引入信息度量的必然；
3. 可加性：事件总不确定性可表示为各事件不确定性和。

香农从数学上严格证明了满足上述三个条件的随机变量不确定性度量函数具有唯一形式：

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

其中的 C为常数，将其归一化为 $C = 1$ 即得到了信息熵公式。

信息熵性质

- 考虑到信息熵的定义涉及到了事件发生的概率，我们可以假设信息熵是事件发生概率的函数：

$$H(X) = H\big(p(x)\big)$$

- 对于两个相互独立的事件 $X = A, Y = B$ 来说，其发生的概率为：

$$p(X = A, Y = B) = p(X = A) \cdot (Y = B)$$

- 两事件的信息熵根据可加性得：

$$H(p(X = A, Y = B)) = H(p(X = A) \cdot (Y = B)) = H(p(X = A)) + H(p(Y = B))$$

信息熵性质

熵的非负性：

$$H(X) \geq 0$$

熵的期望性：

$$H(X) = E(\log(1/p(X)))$$

熵的对数底可换性：

$$H_b(X) = (\log_b \mathrm{a})\, H_\mathrm{a}(X)$$

信息熵示例

赌马比赛里，有4匹马$\{A, B, C, D\}$，获胜概率分别为$\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\}$。

接下来，让我们将哪一匹马获胜视为一个随机变量 $X \in \{A, B, C, D\}$ 。

假定我们需要用尽可能少的二元问题来确定随机变量X 的取值。

例如：问题1：A获胜了吗？问题2：B获胜了吗？问题3：C获胜了吗？

最后我们可以通过最多3个二元问题，来确定 的取值，即哪一匹马赢了比赛。

信息熵示例

如果$X=A$，那么需要问1次（问题1： 是不是A？），概率为$\frac{1}{2}$

如果$X=B$，那么需要问2次（问题1： 是不是A？ 问题2： 是不是B？），概率为$\frac{1}{4}$

如果$X=C$，那么需要问3次（问题1，问题2，问题3），概率为$\frac{1}{8}$

如果$X=D$，那么同样需要问3次（问题1，问题2，问题3），概率为$\frac{1}{8}$

那么很容易计算，在这种问法下，为确定X取值的二元问题数量为：

$$E(N)=\frac{1}{2}*1+\frac{1}{4}*2+\frac{1}{8}*3+\frac{1}{8}*3=\frac{7}{4}$$

信息熵示例

那么我们回到信息熵的定义，会发现通过之前的信息熵公式，神奇地得到了：

$$H(X)=\frac{1}{2}\log(2)+\frac{1}{4}\log(4)+\frac{1}{8}\log(8)+\frac{1}{8}\log(8)=\frac{1}{2}+\frac{1}{2}+\frac{3}{8}+\frac{3}{8}=\frac{7}{4}bits$$

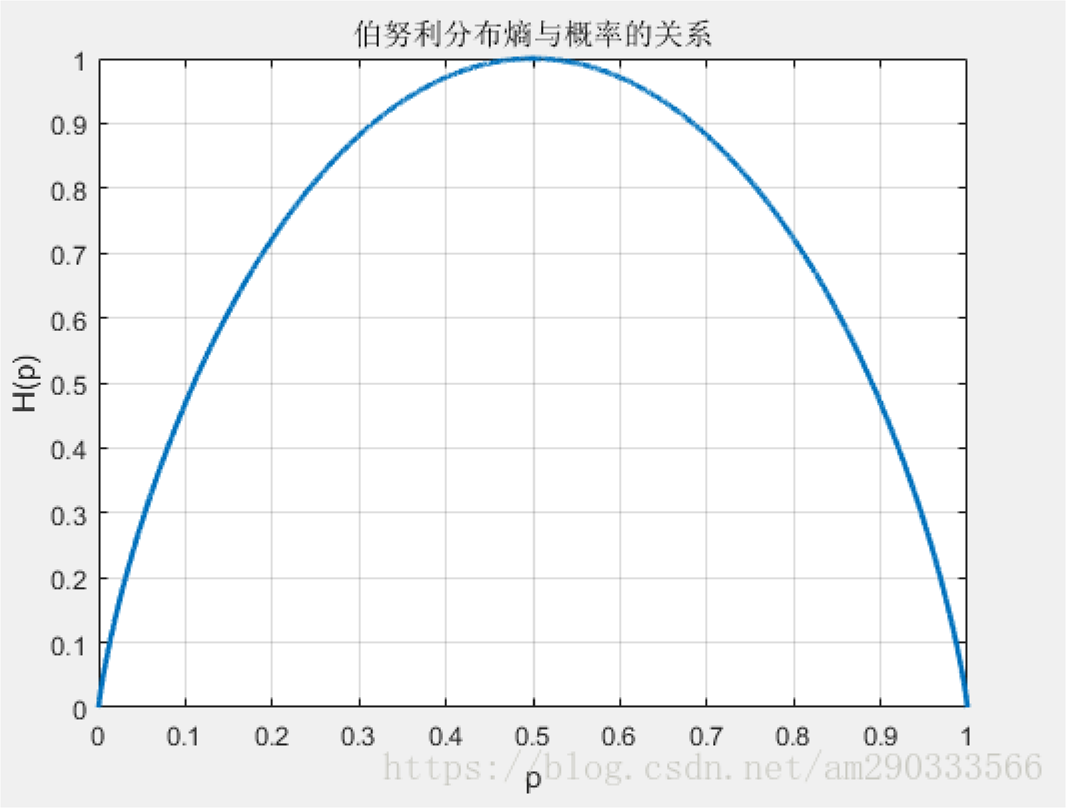
在二进制计算机中，一个比特为0或1，其实就代表了一个二元问题的回答。也就是说，在计算机中，我们给哪一匹马夺冠这个事件进行编码，所需要的平均码长为1.75个比特。

平均码长定义为：

$$L(C)=\sum_{x\in\chi}p(x)l(x)$$

信息熵示例

X 为伯努利（0-1）分布，记H(p)=-(plogp+qlogq)， p+q=1



信息熵示例

很显然，为了尽可能减少码长，我们要给发生概率$p(x)$较大的事件，分配较短的码长$l(x)$。这是霍夫曼编码的基本概念。

那么$\{A, B, C, D\}$四个实践，可以分别由$\{0,10,110,111\}$表示，那么很显然，我们要把最短的码0分配给发生概率最高的事件 A，以此类推。而且得到的平均码长为1.75比特。如果我们硬要反其道而行之，给事件A分配最长的码111，那么平均码长就会变成2.625比特。