信息熵与数据压缩

祝润天

复旦大学计算机科学技术学院

2024 年 10 月 11 日

信息的度量

抛硬币

抛一枚硬币，平均意义下，最少需要几个比特来表示得到的结果？

抛出反面记为 0，抛出正面记为 1。则两种情况均需要 1 个比特来表示结果。平均需要

$$E = \frac{1}{2} \times 1 + \frac{1}{2} \times 1 = 1$$

个比特。

信息的度量

投骰子

投一个骰子，平均意义下，最少需要几个比特来表示得到的结果？

将 6 种结果分别编码为

$$1 \to 000, 2 \to 001, 3 \to 010$$
$$4 \to 011, 5 \to 100, 6 \to 101$$

平均需要 3 个比特。
能否更少？

信息的度量

投骰子

投一个骰子，平均意义下，最少需要几个比特来表示得到的结果？

将 6 种结果分别编码为

$$1 \to 000, 2 \to 001, 3 \to 010$$
$$4 \to 011, 5 \to 10, \quad 6 \to 11$$

平均需要

$$E = \frac{4}{6} \times 3 + \frac{2}{6} \times 2 \approx 2.67$$

个比特。

文件数据压缩

例

你说的对，但是《原神》是由米哈游自主研发的一款全新开放世界冒险游戏。游戏发生在一个被称作「提瓦特」的幻想世界，在这里，被神选中的人将被授予「神之眼」，导引元素之力。你将扮演一位名为「旅行者」的神秘角色，在自由的旅行中邂逅性格各异、能力独特的同伴们，和他们一起击败强敌，找回失散的亲人。

汉字字频

的	一	是	$\cdots$
0.0575	0.0473	0.0429	$\cdots$

使用等长的 Unicode 码 $\rightarrow$ 为概率更高的字符分配更短的码字

唯一可译码

将一个文件编码后，能够确保恢复这个文件。

例

$c(x)$ 将 $\mathcal{X} = \{1,2,3,4,5,6\}$ 中的每个字符 x 编码为它的二进制表示，对应的码字分别为 $1,10,11,100,101,110$。则 c 不是一个唯一可译码，因为 110 可以被译成 6 或者 12。

- 令 $c^*(x_1,\ldots,x_n) = c(x_1)\cdots c(x_n)$。则 c 是一个唯一可译码当且仅当 c^* 是一个单射。

信息熵

信息熵

假设 X 是一个离散型随机变量，其可能的取值集合为$\mathcal{X}$。则 X 的信息熵定义为

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

符号码信源编码定理

若编码 c 将 x 编码为有限长的 01 字符串 $c(x)$，则

$$H(X) \le E(|c(x)|)$$

对 $\mathcal{X}$ 中的字符 x，c 将其编码为对应的 01 字符串 $c(x)$。令 $l_x = |c(x)|$ 是码字的长度。则 $\sum_{x\in\mathcal{X}} 2^{-l_x} \leq 1$。

证明.

令 $l_{min} = \min_{x\in\mathcal{X}}|c(x)|$, $l_{max} = \max_{x\in\mathcal{X}}|c(x)|$，$a(k)$ 为长度为 k 的码字个数。则

$$\left(\sum_{x\in\mathcal{X}} 2^{-|c(x)|}\right)^n = \sum_{k=nl_{min}}^{nl_{max}} a(k)2^{-k}$$

由唯一可译性质码的性质可知 $a(k) \leq 2^k$，则

$$\sum_{x\in\mathcal{X}} 2^{-|c(x)|} \leq (n(l_{max} - l_{min} + 1))^{\frac{1}{n}}$$

令 $n \to \infty$ 则得到结论。 □

符号码信源编码定理

$$\begin{aligned} \textit{minimize} \quad & \sum_{x\in\mathcal{X}} p(x) l_x \\ s.t. \quad & \sum_{x\in\mathcal{X}} 2^{-l_x} \le 1 \text{ and } l_x > 0 \end{aligned}$$

证明.

极值只会在 $\sum_{x\in\mathcal{X}} 2^{-l_x} = 1$ 时取到。利用拉格朗日乘数法，构造

$$L(l_{x_1}, \ldots, l_{x_n}, \lambda) = \sum_{x\in\mathcal{X}} p(x) l_x - \lambda \left(\sum_{x\in\mathcal{X}} 2^{-l_x} - 1 \right)$$

解得

$$l_x = -\log p(x)$$

□

符号码信源编码定理

因此，当 $l_x = -\log p(x)$ 时，$\sum_{x\in\mathcal{X}} p(x)l_x$ 取到极小值

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

从而

$$H(X) \le \sum_{x\in\mathcal{X}} p(x)l_x = E(|c(x)|)$$

这就是符号码信源编码定理。

信源编码定理（香农第一定理）

信源编码定理

设离散型随机变量 X 的取值集合为 $\mathcal{X}$。对任意 $\varepsilon > 0$，存在一个整数 n 和一个唯一可译码 $c : \mathcal{X}^n \to \{0,1\}^*$，使得

$$\frac{1}{n} E(|c(X_1, \ldots, X_n)|) \le H(X) + \varepsilon$$

- 典型序列：数量少，但出现概率大
- 给典型序列分配短编码，给其他序列分配长编码

典型序列

假设 $X_1, \ldots, X_n$ 为取值为 $\{0, 1\}$，独立同分布的随机变量，取到 0 的概率为 p，$q = p - 1$。则对于某个有 a 个 0，b 个 1 的序列（例如 $\underbrace{0 \cdots 0}_{a}\underbrace{1 \cdots 1}_{b}$），其出现的概率为

$$P((X_1, \ldots, X_n) = (x_1, \ldots, x_n)) = p^a q^b$$