信息论与编码

马啸
maxiao@mail.sysu.edu.cn

计算机学院
中山大学

2021 年春季学期

1. Shannon-Fano-Elias Coding

2. 算术码

3. LZ编码

Shannon-Fano-Elias Coding

Assume that for a source $\mathcal{X} = \{1, 2, \ldots, m\}$, $p(x) > 0$ for all x. The cumulative distribution function is defined as

$$F(x) = \sum_{a \leq x} p(a). \tag{1}$$

Let $\bar{F}(x)$ be a modified cumulative distribution function as

$$\bar{F}(x) = \sum_{a < x} p(a) + \frac{1}{2} p(x). \tag{2}$$

The value of $\bar{F}(x)$ can be used as codeword for x. Since $\bar{F}(x)$ is a real number expressible only for an infinite number of bits, we round off $\bar{F}(x)$ to $\ell(x)$ bits and denote it by $\lfloor \bar{F}(x) \rfloor_{\ell(x)}$.

What should the length $\ell(x)$ be?

If $\ell(x) = \lceil \log \frac{1}{p(x)} \rceil + 1$, then we have

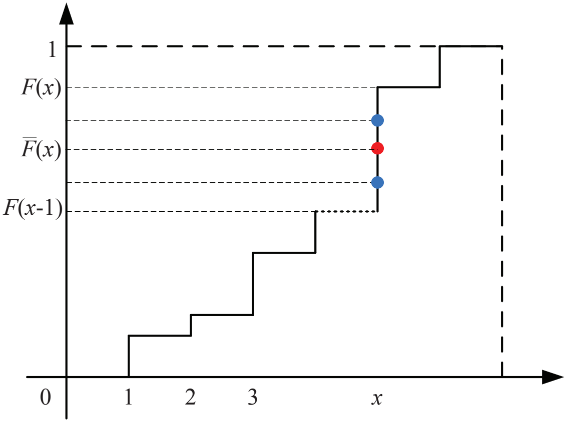
$$\bar{F}(x) - \lfloor \bar{F}(x) \rfloor_{\ell(x)} < 2^{-\ell(x)} < \frac{p(x)}{2} = \bar{F}(x) - F(x-1) \tag{3}$$

Then $\lfloor \bar{F}(x) \rfloor_{\ell(x)}$ lies within the lower-half step corresponding to x. Thus $\ell(x)$ bits suffices to describe x.
On the other hand,

$$\lfloor \bar{F}(x) \rfloor_{\ell(x)} + 2^{-\ell(x)} < \lfloor \bar{F}(x) \rfloor_{\ell(x)} + \frac{p(x)}{2} = \lfloor \bar{F}(x) \rfloor_{\ell(x)} + F(x) - \bar{F}(x) < F(x). \tag{4}$$

Let the bits corresponding to $\lfloor \bar{F}(x) \rfloor_{\ell(x)}$ be $z_1 z_2 \ldots z_\ell$. Then the interval corresponding to the codeword $z_1 z_2 \ldots z_\ell$ is $[0.z_1 z_2 \ldots z_\ell, 0.z_1 z_2 \ldots z_\ell + \frac{1}{2^\ell}]$. Such intervals are disjoint from the above two inequalities. Hence the code is prefix-free. And the average length is

$$L = \sum p(x) \left(\lceil \log \frac{1}{p(x)} \rceil + 1 \right) < \sum p(x) \left(\log \frac{1}{p(x)} + 2 \right) = H(X) + 2. \tag{5}$$


1
F(x)
F̄(x)
F(x-1)
0
1
2
3
x

算术码编码

算术码编码的主要思想如下：设信源符号集包含 N 个符号，对每个符号从 $1 \sim N$ 进行编号。设每个符号出现的概率为 p_i，此处 $1 \le i \le N$。在初始区间给每个符号分配一个初始子区间，其长度等于对应符号的概率。每个序列的首个信源符号概率确定本序列编码的初始区间，后续信源符号的编码过程是对选定区间进行再分割的过程。

算术码解码

算术码解码的过程与编码的过程相反，相当于编码的逆运算。解码前首先需要对区间 $[0, 1)$ 按照符号概率进行分割。解码时仅输入一个小数，观察输入的小数位于哪个子区间，输出对应的符号后，选定该子区间并在该子区间中继续下一轮的分割。不断地进行这个过程，直到所有的符号都被解码出来。

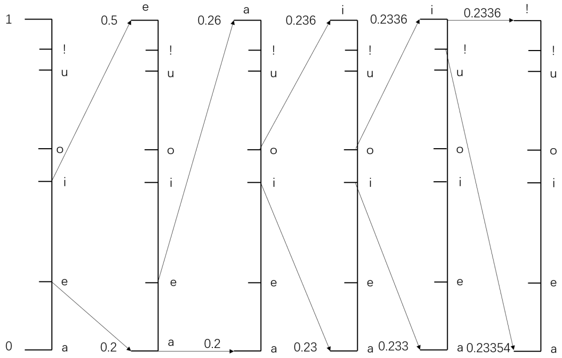
算术码实例：编码

下面举一个实例来说明算术码编码的过程。符号序列 *eaii*! 的概率分布如表所列。

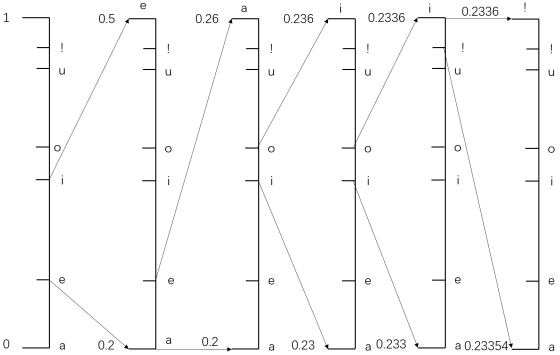
Table: 信源符号概率分布

符号	概率	区间
a	0.2	[0,0.2)
e	0.3	[0.2,0.5)
i	0.1	[0.5,0.6)
o	0.2	[0.6,0.8)
u	0.1	[0.8,0.9)
!	0.1	[0.9,1.0)

编码从区间 $[0,1)$ 开始。首先按照符号的概率分布将当前区间分割为多个子区间，然后根据当前输入的符号选择对应的子区间。



第一个编码的符号是 e.根据表查看符号 e 的概率区间是 $[0.2, 0.5)$，因此当前的编码区间更新为 $[0.2, 0.5)$。接下来根据信源符号的概率对区间 $[0.2, 0.5)$进行分割，下一个输入的符号是 a,其所对应的当前区间是 $[0.2, 0.26)$，因此在编码下一个符号 i 前，把当前区间更新为 $[0.2, 0.26)$，再按照信源符号的概率将其划分为五个子区间。依此类推，最后会得到一个最终区间，在这个区间中任意选一个数就是编码结果。



按照不断地更新区间、分割区间、选择子区间的方法，直到所有符号全部编码完成。最后得到的区间是 $[0.23354, 0.2336)$。输出这个区间内的某个小数，如 0.23358，那么序列 *eaii*! 经过算术编码后的编码结果就是 0.23358。

算术码实例：译码

编码的结果是 0.23358,亦即解码器的输入为 0.23358。可以发现， 0.23358 落在了区间 $[0.2, 0.5)$中，该区间所对应的符号是 e,于是可以解码出符号 e。接下来选择子区间 $[0.2, 0.5)$作为下一次解码的区间，并根据信源符号概率布对区间 $[0.2, 0.5)$ 进行分割。由于 0.23358 落在子区间 $[0.2, 0.26)$ 中，所对应的符号是 a，于是可以解码出符号 a，进而将子区间 $[0.2, 0.26)$ 作为下一次解码的区间。重复该过程直到解码出所有符号。

LZ78 算法：编码

设信源符号集 $\mathbf{A} = \{a_1, a_2, \cdots, a_k\}$ 共 K 个符号，设输入信源符号序列为 $\mathbf{u} = (u_1, u_2, \cdots, u_L)$ 。编码时将此序列分成不同的段。分段的规则为：尽可能取最少个相连的信源符号，并保证各段都不相同。

开始时，先取一个符号作为第一段，然后继续分段。若出现与前面相同的符号时，就再取紧跟后面的一个符号一起组成一个段，使之与前面的段不同。这些分段构成字典。当字典达到一定大小后，再分段时就应查看有否与字典中的短语相同，若有重复就添加符号，以便与字典中短语不同，直至信源符号序列结束。这样，不同的段内的信源符号可看成一短语，可得不同段所对应的短语字典表。

码字构成：前面字段所在的段号+末尾的一个符号对应的号。设 $\mathbf{u}$ 构成的字典中的短语共有 $M(\mathbf{u})$ 个。若编为二元码，段号所需码长 $n = \lceil \log M(\mathbf{u}) \rceil$，每个符号需要的码长为 $\lceil \log K \rceil$。单符号的码字段号为0。

LZ78 算法：译码

LZ78 编码的编码方法很便捷，译码也很简单，可以一边译码一边建立字典，只需要传输字典的大小，无需传输字典本身。当编码的信源序列较短时，LZ 算法性能似乎会变坏,但是当序列增长时，编码效率会提高，平均码长会逼近信源熵。

LZ78 算法实例

设 $U=\{a_1,a_2,a_3,a_4\}$，信源序列为 $a_1a_2a_1a_3a_2a_4a_2a_4a_3a_1a_1a_4\cdots$，按照分段规则，可以分为 $a_1,a_2,a_1a_3,a_2a_4,a_2a_4a_3,a_1a_1,a_4$，共 7 段，字典表如表所示。

段号	短语	编码
1	a_1	000 00
2	a_2	000 01
3	a_1a_3	001 10
4	a_2a_4	010 11
5	$a_2a_4a_3$	100 10
6	a_1a_1	001 00
7	a_4	000 11

其中每个符号编码如下

a_1	a_2	a_3	a_4
00	01	10	11

7 个短语使用 3 bit 就可以表示段号，每个信源符号用 2 bit 表示，因此，一个短语使用 5 bit。

LZ78 算法

将有 K 个符号，长为 L 的信源序列 $\mathbf{u}$ 分为 $M(\mathbf{u})$ 个码段后，设最长的段的长度为 ℓ_{max}，可以证明，每个源符号的平均码长有

$$H(U) + \frac{\log K}{\ell_{max}} < \bar{n} < H(U) + \frac{\log K + 2}{\ell_{max}}$$

将编码的信源序列趋于无穷时，ℓ_{max} 也趋于无穷，平均码长趋近于信源熵。

作业

Exercise 1.

Shannon 码需要已知概率质量函数 $p_i, 0 \le p_i \le 1, \sum_{i=1}^{M} p_i = 1$，其码长 $\ell_i = \lceil \log \frac{1}{p_i} \rceil$。若用了一个不匹配的概率质量函数 $q_i, 0 \le q_i \le 1$, $\sum_{i=1}^{M} q_i = 1$，则其码长是 $\tilde{\ell}_i = \lceil \log \frac{1}{q_i} \rceil$。

(1)求其平均码长，并计算不匹配概率对应的平均码长与匹配概率对应的平均码长之差。

(2)请设计一个程序例子，用 Matlab 程序仿真，体现上述计算的合理性。

作业

Exercise 2.

记 $\mathcal{X}=\{0,1\}$ 为二元集，$\mathcal{X}^n$ 为所有 n 重二元数组。$x^n \in \mathcal{X}^n$ 的汉明重量是 x^n 中 1 的个数。

(1) $\mathcal{X}^n$ 中有多少个二元数组？

(2) 重量是 w 的 n 重二元数组有几个？

(3) 以 $n=4$ 为例，重为 2 的数组有 $(0,0,1,1)$，$(0,1,0,1)$，$(1,0,0,1)$，$(0,1,1,0)$，$(1,0,1,0)$，$(1,1,0,0)$ 6 个，因此我们称其可以承载 2 bit信息。即可以建立一个单射：$00 \to (0,0,1,1)$，$01 \to (0,1,0,1)$，$10 \to (1,0,0,1)$，$11 \to (0,1,1,0)$。记 $\mathcal{X}^n$中重量为 w 的全体二元数组为 T_w，求最大的 k，使得存在 $\mathcal{X}^k \to T_w$ 的单射。

(4) **思考题(不必完成)：** 任意给定 n,w，如何建立 $\mathcal{X}^k \to T_w$ 之间的单射？此单射可以称为等重编码，那么应该如何译码？

谢谢！