信源编码

- 信源编码就是用最少的数字或符号串表示所期望的信息。
- 编码规则：
  1 从码字序列到原始消息的对应有明确的规则
  2 在码字不发生错误的情况下，由码字序列可以正确恢复原始数据。
  3 如果还考虑性能，则要求表示同样的信息，不能用更短的数据表示，即得到的码是最优的。

- 例子　要表达信源的3种不同的消息　“晴天”　“阴天”　“下雨”

方法（1）：　“晴天”　“阴天”　“下雨”

存储需要16*6=96 bits

方法（2）：　“晴”　“阴”　“雨”

存储需要16*3=48 bits

方法（3）：01　10　11

定长信源编码

- 假定信源有n个符号，依次赋予序号 0,1,2,…,n-1，

然后用等长的二进制表示，则这就是一种编码方案。

等长信源编码：将所有信源消息都编码为固定相同长度的码字。 $k = \lceil \log_2 n \rceil$

定理　如果上述二进制表示长度为 ，则是一种最优的等长信源编码。

- 证明： 需要证明上述定理中给出的编码规则满足上述最优编码的性质。 首先不难验证，译码规则是将所获得的码字序列按照k比特截断，然后将每一段对应到原始消息。这种明确的译码规则将每一个原始信源信息对应到长度为k 的一个二元数组，在码字不发生错误的情况下，显然逆过程也正确。另外显然，不能用少于k的二元数组来表示信源消息，因为所有这样的固定长度的二元数组的个数少于信源消息的个数，即对所有t<k, 有 $2^t < n$ 。这就证明了上述方案的最优性。

变长信源编码

- 例　信源消息共有三个字符： 晴，阴， 雨

编码方式 (1)： 00　01　10　等长编码

编码方式 (2)：　0　　1　10　　(10)

编码方式 (3)：　0　　01　10　　(01010)

编码方式 (4)：　00　01　1

前缀码：没有一个码字是另一个码字前缀的信源码。

Kraft不等式

- 对于一个前缀码，假设各信源符号的编码长度依次为 $L_1 \le L_2 \le \ldots \le L_n$，则有

$$\sum_{i=1}^{n} 2^{-L_i} \le 1$$

变长信源编码可以将信源消息的概率分布充分利用，使得信源输出消息对应的码字序列最短。

Huffman信源编码

- 算法

假设信源符号xi的出现概率为 $P(x_i), i = 1,2,\ldots,n.$ 则编码步骤如下：

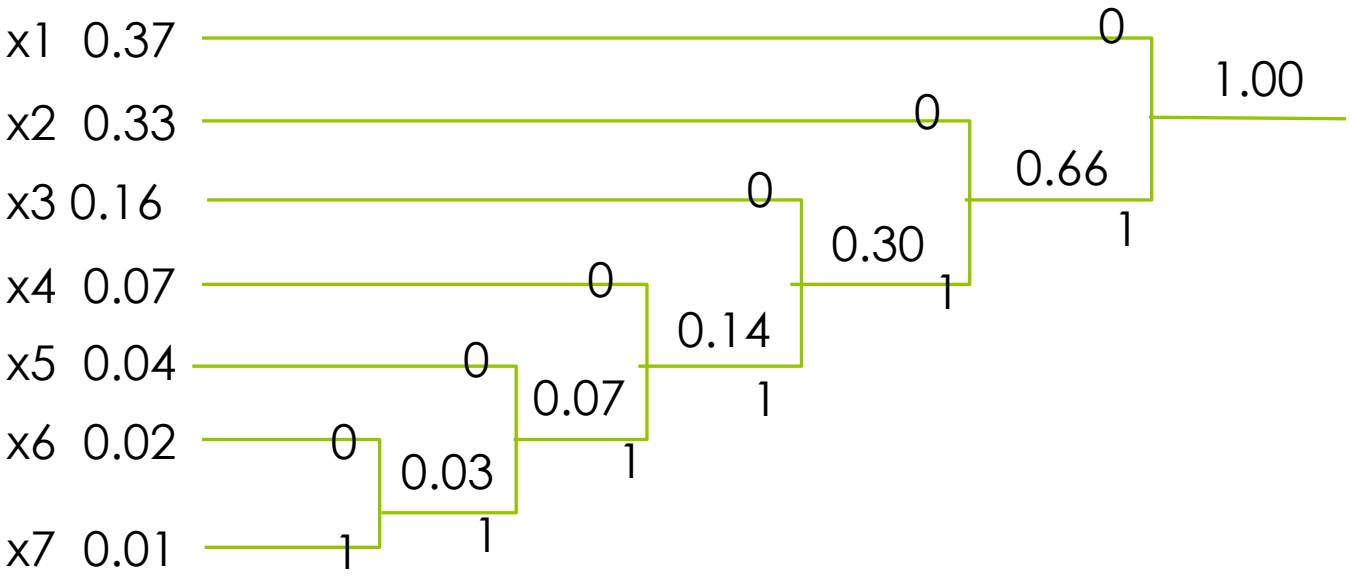
(1) 将信源符号按概率递减次序由上至下排列。

(2) 将最下面的两个符号连接在一起，并分别对应两个分支标记为0和1。将它们的概率和作为一个新符号的概率，重新选取最小两个概率的符号，重复上述连接编码方法。

(3) 将此过程进行下去，直到只剩下一个概率，这就完成了Huffman 树的构造。

例　著名的变长信源编码算法 - - Huffman编码

- 假设信源X有七个符号x1, x2, …, x7,它们的概率分别为0.37, 0.33, 0.16, 0.07, 0.04, 0.02, 0.01。首先根据算法9.3.1建立Huffman树，如图所示



根据Huffman编码算法的第(2)步， 沿Huffman树的最后节点到每个符号的路径的标号就是该符号的码字， 于是得到信源符号$x_1, x_2, ..., x_7$对应的码字分别为0, 10,110,1110, 11110,111110,11111.

这个看上去不怎么好的编码，对于这个特定概率分布的信源，却是最优的。

不唯一，前缀码

信源编码定理

- 假定信源符号 x_i 的出现概率为 $P(x_i)$，该符号的码字长度为 $L(x_i)$，则信源符号平均所占的比特数为

$$\bar{R} = \sum_{i=1}^{n} L(x_i)P(x_i).$$

Shannon从理论上给出了离散无记忆信道的变长编码的平均码长所能达到的理论区间。

信源编码定理

- 设X为离散无记忆信源的字母集合，其有限熵为H(X)，而其输出符号 x_i 的出现概率为 $P(x_i), i=1,2,...,n,$ 则存在满足前缀条件的码，其平均长度满足不等式

$$H(X) \le \bar{R} \le H(X)+1.$$

注： 用来表示信源符号的平均最小比特数必须至少等于该信源的熵。同时也说明，一个高熵的信源必须用更多的比特数来表示信源符号。

- 考虑前面Huffman编码，则信源熵

$$H(X) = -\sum_{i=1}^{7} P(x_i)\log_2 P(x_i) = 2.1151(bit)$$

而通过Huffman编码后的平均每个符号的码字长度为

$$\bar{R} = \sum_{i=1}^{7} L(x_i)P(x_i) = 1*0.37 + 2*0.33 + 3*0.16 + 4*0.07 + 5*0.04 + 6*0.02 + 6*0.01 = 2.1700(bit)$$

- 信源编码定理说明，任何有效编码的平均码字长度不小于信源熵。平均码字长度越接近信源熵，说明编码效率越高。

- 定义编码效率

$$\eta = \frac{H(X)}{\overline{R}}$$

上例的编码效率为 $\eta = 2.1152 / 2.17 = 0.9747$