上周作业解答

Exercise 1.

设一个随机变量的分布律是 $p_1 \geq p_2 \geq \cdots \geq p_m$，其最优编码具有码长 $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_m$。试问，能否得出对于所有$i$，均有$\ell_i \leq \lceil \log \frac{1}{p_i} \rceil$？

上周作业解答

解答：

对于最优码，它的码字码长并不一定总是小于等于 $\lceil \log \frac{1}{p_i} \rceil$。现举一反例说明：

考虑一个随机变量 X，其概率分布满足 $(\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12})$。对其进行Huffman编码，结果可以得到码长为 $(2, 2, 2, 2)$ 或 $(1, 2, 3, 3)$（取决于构造树时合并的顺序）。这两组码都具有相同的平均码长，但是第 2 组码中，第 3 个符号具有长度为 3 的码长，大于 $\lceil \log \frac{1}{p_3} \rceil = 2$。

因此，尽管从平均意义而言，Huffman编码是最优码，其平均码长会比Shannon码更小，但对于具体某个符号而言，Shannon码构造出的码字有可能更短，两者并不矛盾。

上周作业解答

Exercise 2.[王育民(2013)]

令离散无记忆信源

$$U = \left\{ \begin{array}{cc} a_1 & a_2 \\ 0.6 & 0.4 \end{array} \right\}$$

(a)求 U 的最佳二元码、平均码长及编码效率。

(b)求 U^2 的最佳二元码、平均码长及编码效率。

(c)求 U^3 的最佳二元码、平均码长及编码效率。

(d)求 U^4 的最佳二元码、平均码长及编码效率。

上周作业解答

解答：

(a)U：列出 U 条件下各符号的概率分布

U	P
a_1	0.6
a_2	0.4

有了概率分布后，根据Huffman编码构造出码字。

U	codeword
a_1	0
a_2	1

计算得编码速率 $R = \sum_{i=1}^{2} P_i \ell_i = 1$ bit。
熵 $H(X) = -\sum_{i=1}^{2} P(X_i) \log P(X_i) = 0.9709$
编码效率 $\eta = H(X)/R = 0.9709$。

上周作业解答

(b)U^2：列出 U^2 条件下各符号的概率分布

U^2	P	U^2	P
a_1a_1	0.36	a_2a_1	0.24
a_1a_2	0.24	a_2a_2	0.16

有了概率分布后，根据Huffman编码构造出码字。

U^2	codeword	U^2	codeword
a_1a_1	11	a_2a_1	01
a_1a_2	10	a_2a_2	00

计算得编码速率 $R = \frac{1}{2}\sum_{i=1}^{2} P_i\ell_i = 1$ bit。
熵 $H(X) = 0.9709$ bit
编码效率 $\eta = H(X)/R = 0.9709$。

上周作业解答

(c)U^3：列出 U^3 条件下各符号的概率分布

U^3	P	U^3	P
$a_1a_1a_1$	0.216	$a_1a_2a_2$	0.096
$a_1a_1a_2$	0.144	$a_2a_1a_2$	0.096
$a_1a_2a_1$	0.144	$a_2a_2a_1$	0.096
$a_2a_1a_1$	0.144	$a_2a_2a_2$	0.064

有了概率分布后，根据Huffman编码构造出码字。

U^3	codeword	U^3	codeword
$a_1a_1a_1$	01	$a_1a_2a_2$	000
$a_1a_1a_2$	110	$a_2a_1a_2$	001
$a_1a_2a_1$	100	$a_2a_2a_1$	1111
$a_2a_1a_1$	101	$a_2a_2a_2$	1110

计算得编码速率 $R = \frac{1}{3}\sum_{i=1}^{2} P_i\ell_i = 0.9813$ bit。

熵 $H(X) = 0.9709$ bit

编码效率 $\eta = H(X)/R = 0.9894$。

上周作业解答

(d)U^4：列出 U^4 条件下各符号的概率分布

U^4	P	U^4	P
$a_1a_1a_1a_1$	0.1296	$a_1a_2a_2a_1$	0.0576
$a_1a_1a_1a_2$	0.0864	$a_2a_1a_2a_1$	0.0576
$a_1a_1a_2a_1$	0.0864	$a_2a_2a_1a_1$	0.0576
$a_1a_2a_1a_1$	0.0864	$a_1a_2a_2a_2$	0.0384
$a_2a_1a_1a_1$	0.0864	$a_2a_1a_2a_2$	0.0384
$a_1a_1a_2a_2$	0.0576	$a_2a_2a_1a_2$	0.0384
$a_1a_2a_1a_2$	0.0576	$a_2a_2a_2a_1$	0.0384
$a_2a_1a_1a_2$	0.0576	$a_2a_2a_2a_2$	0.0256

上周作业解答

有了概率分布后，根据Huffman编码构造出码字

U^4	codeword	U^4	codeword
$a_1a_1a_1a_1$	000	$a_1a_2a_2a_1$	1010
$a_1a_1a_1a_2$	111	$a_2a_1a_2a_1$	1011
$a_1a_1a_2a_1$	0001	$a_2a_2a_1a_1$	1100
$a_1a_2a_1a_1$	0000	$a_1a_2a_2a_2$	1101
$a_2a_1a_1a_1$	0010	$a_2a_1a_2a_2$	00110
$a_1a_1a_2a_2$	0011	$a_2a_2a_1a_2$	00111
$a_1a_2a_1a_2$	1000	$a_2a_2a_2a_1$	00100
$a_2a_1a_1a_2$	1001	$a_2a_2a_2a_2$	00101

计算得编码速率 $R = \frac{1}{4}\sum_{i=1}^{2} P_i\ell_i = 0.9812$ bit。
熵 $H(X) = 0.9709$ bit
编码效率 $\eta = H(X)/R = 0.9895$。

Exercise 3.

1) 三元最优编码树应该满足什么条件？

2) **[王育民(2013)]**设离散无记忆信源

$$U = \left\{ \begin{array}{cccccc} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ 0.3 & 0.2 & 0.15 & 0.15 & 0.1 & 0.1 \end{array} \right\}$$

试求其二元和三元 Huffman 编码。

上周作业解答

(1)

1. 概率大的消息对应的叶子节点不应该比概率小的对应的叶子节点更深；
2. 最深一层的叶子数目模 $D=3$ 余数不能余 1（最深一层靠左排）；
3. 每个内点（除了次深层的），都应该有 $D=3$ 个儿子节点。

(2) 根据概率分布，根据二元Huffman编码构造出码字

U	codeword
a_1	11
a_2	01
a_3	100
a_4	101
a_5	000
a_6	001

上周作业解答

D 元Huffman编码，达到满树时能够使得平均码长最短。

对于 D 元Huffman编码，从第 i 阶的1个节点到第 $i+1$ 阶的节点，增加的数目为$D-1$。显然，达到满树的总叶子数目为 $s=1+(D-1)*k$，其中 k 是一个非负整数。信源符号数 n 必须满足上式，编码才能达到满树。否则，利用上式计算出大于 n 的最小正整数，然后给信源增补零概率符号。本题中，通过计算 s 的最小值为7，则需要增补 1 个零概率符号。

U	codeword
a_1	1
a_2	22
a_3	20
a_4	21
a_5	00
a_6	01