信息论导论

第4讲 Huffman码，分组编码、熵率，Shannon第一定理

[信息论教材中页码范围] Huffman码：p118~p127，分组编码、Shannon第一定理：p113~p115，熵率：p74~p77

信息学部-信息科学与技术学院 吴绍华
hitwush@hit.edu.cn

关于最优信源编码码长上下界的几个问题

哈爾濱工業大學(深圳) HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

$$H(X) \le L^* \le L_s < H(X) + 1$$

- 对于Shannon编码，若使用错误的概率分布计算码字长度，有何代价?
- 两个 “=” 分别在什么情形下成立?
- 什么样的信源编码方法/算法是最优的?
- 从上界中可以看到，极端情形下即时是最优信源编码，期望码长也比下界长将近1bit（相对下界的额外开销），这种极端情形是什么?
- 如何减小（甚至消除）这将近1bit的额外开销?

最优编码应具有的性质

哈爾濱工業大學(深圳) HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

引理 5.8.1

最优的二元即时码（即有最小的期望码长）必须满足如下性质：

1. 概率越小的消息，码字长度越长（即：若 $p_j > p_k$，则 $l_j \leq l_k$）
2. 最长的两个码字应具有相同的长度（否则，可把较长码字的多余部分去掉）
3. 最长的两个码字应仅在最后一位不同（否则，可把两者多余的码字部分均去掉）

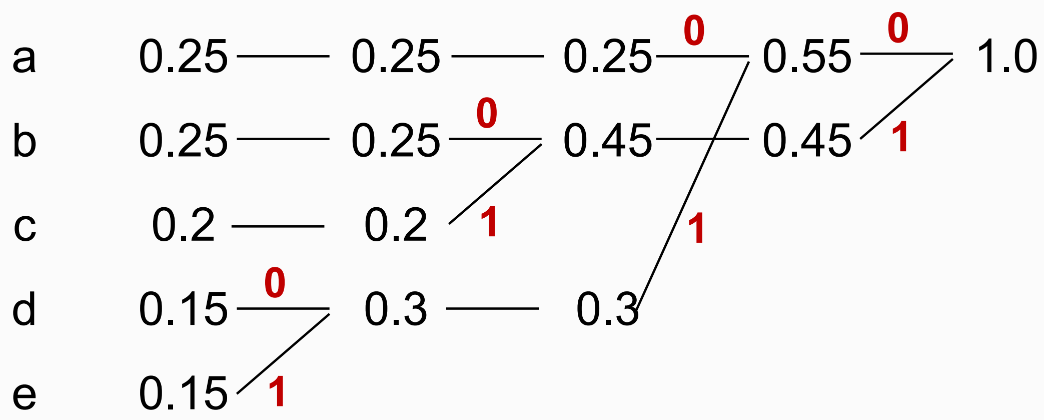
- 总结：若 $p_1 \geq p_2 \geq \ldots \geq p_m$，则最优信源编码应满足 $l_1 \leq l_2 \leq \ldots \leq l_{m-1} = l_m$，且码字 $C(x_{m-1})$ 与 $C(x_m)$ 仅在最后一位不同

Huffman码

- Huffman 编码（1952 年由 Huffman 发明的最优即时码编码算法）简要流程：
  1. 在信源分布中选择概率最小的两个 $p(x_i)$，为其对应的两个码字的最后一位分别分配 0 和 1，然后将这两者合并；
  2. 重复第一步，直到最后只剩一个。

例

信源符号 $\mathscr{X} = \{a; b; c; d; e\}$，其对应概率分布 $\mathbf{p}_X = \{0.25; 0.25; 0.2; 0.15; 0.15\}$

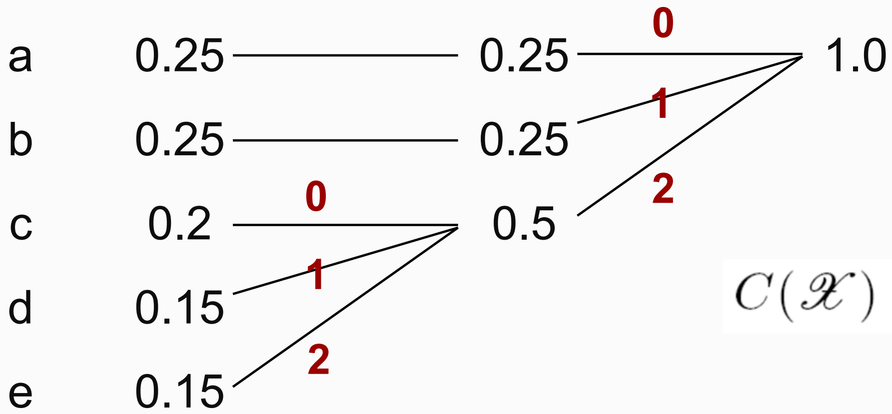


从码树根节点读到叶子节点，
即可得到各消息符号对应的码字
$C(\mathscr{X}) = \{00; 10; 11; 010; 011\}$
$L_C = 2.3, H(X) = 2.286$

Huffman码（D≥3）

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

- **对上一页中的信源进行三元编码：**

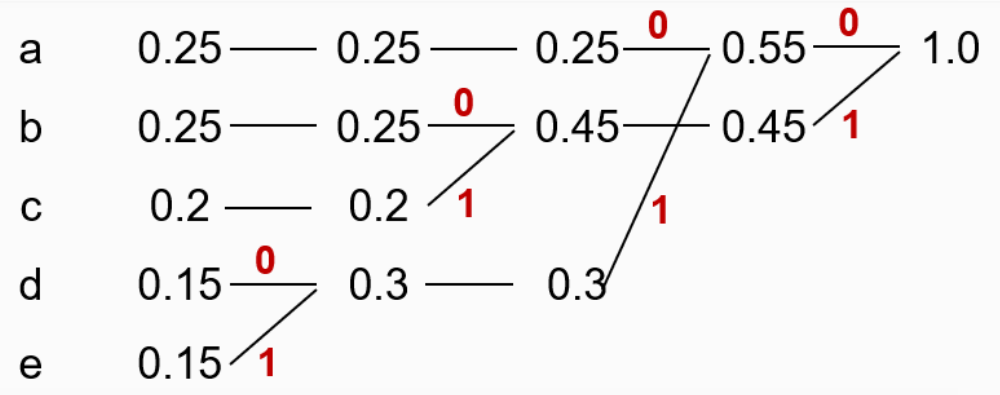


$$C(\mathscr{X}) = [0; 1; 20; 21; 22]$$

- **对于D≥3的Huffman编码，需先判断是否需要添加额外的辅助编码符号（举例说明）**

Huffman码是最优前缀码

- Huffman 编码结果中的每一级，都可看成是针对相应概率分布、给出了一种二元编码结果：
  - $\mathbf{p}_2 = [0.55; 0.45]$,
    $C_2 = [0; 1], L_2 = 1$
  - $\mathbf{p}_3 = [0.25; 0.45; 0.3]$,
    $C_3 = [00; 1; 01], L_3 = 1.55$
  - $\mathbf{p}_4 = [0.25; 0.25; 0.2; 0.3]$,
    $C_4 = [00; 10; 11; 01], L_4 = 2$
  - $\mathbf{p}_5 = [0.25; 0.25; 0.15; 0.15]$,
    $C_5 = [00; 10; 11; 010; 011]; L_5 = 2.3$
  - $\mathbf{p}_m = \ldots$



- 可以证明，每一级编码 C_m 均是最优的——等价于证明了 Huffman 码的最优性

Huffman码的最优性（证明）

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

定理

由 Huffman 编码生成的各级码字 C_m 均是最优的

证明（数学归纳法）.

- $m=2$：只有两个符号，C_2 显然是最优的
- $m=k-1$：假设C_{k-1} 是最优的
- $m=k$：（反证法）
    - 若 C_k 不是最优，则必定存在另一个编码 C_k^*，其具有最小期望码长 $L_{C_k^*}$，即 $L_{C_k^*} < L_{C_k}$
    - 根据最优二元编码的性质，C_k^* 中对应最小概率 p_i, p_j 的两个码字仅在最后一位不同
    - 采用与 Huffman 算法相同的操作：将 x_i, x_j 对应符号合并，得到新的编码 C_{k-1}^*，容易得到 $L_{C_{k-1}^*} = L_{C_k^*} - p_i - p_j$
    - 由于从 C_k 到 C_{k-1} 也是由 x_i, x_j 合并而成，故有 $L_{C_{k-1}} = L_{C_k} - p_i - p_j$
    - 从而有 $L_{C_{k-1}^*} < L_{C_{k-1}}$，与“$C_{k-1}$ 是最优的”矛盾，故 C_k 是最优的

□

Huffman码的最优性（定理）

定理 5.8.1

Huffman 编码是最优的，意即：假设 C^* 为 Huffman 编码而 C' 为其他任意唯一可译码，则有 $L(C^*) \leq L(C')$

- Huffman 编码构造出的码字只是众多可能的最优码字中的一种。
  - 交换具有相同长度的两个码字，可以获得另一个最优码。
  - 最优码的码字长度的集合并不唯一（存在具有相同期望长度的码字集合）例如，当概率分布为 $(1/3, 1/3, 1/4, 1/12)$

回顾：最少“二元问题”数量

第一讲 p20

定义 随机变量 X 的熵用 $H(X)$ 表示，定义为：

$$H(X) = E(-\log_2 p(x)) = -\sum_{x\in\mathscr{X}} p(x)\log_2 p(x)$$

对 $H(X)$ 的几点基本理解、拓展：

- 它表示信源 X 的平均香农信息量
- 对于随机变量 X，从未知其值到获知其值，这个过程所获得的平均信息量
- 对于随机变量 X，若允许我们用一系列“二元问题”去确定它的值，则所需要的平均问题数量在区间 $[H(X), H(X)+1)$ 内

本质：设计“二元问题”序列确定随机变量的值，

等价于针对该随机变量进行“二元信源编码”

最少问题数量 = 最优码长

Shannon码 vs. Huffman码

- 从平均意义上说，Huffman 码的码长比 Shannon 码短，其平均码长之差小于 1bit，这是因为：

$$H(X) \leq L^* \leq L_s < H(X) + 1$$

- 但具体到单个码字，无法确定到底是 Shannon 码还是 Huffman 码更短

例

$\mathbf{p}_X = [0.36; 0.34; 0.25; 0.05] \Rightarrow H(X) = 1.78\text{bits}$

Shannon 码：

$$-\log_2(\mathbf{p}_X) = [1.47; 1.56; 2; 4.32]$$

$$\mathbf{l}_s = -\lceil \log_2(\mathbf{p}_X) \rceil = [2; 2; 2; 5]$$

$$L_s = 2.15\text{bits}$$

Huffman 码：

$$\mathbf{l}^* = [1; 2; 3; 3]$$

$$L^* = 1.94\text{bits}$$

关于最优信源编码码长上下界的几个问题

$$H(X) \le L^* \le L_s < H(X) + 1$$

- 对于Shannon编码，若使用错误的概率分布计算码字长度，有何代价?
- 两个 “=” 分别在什么情形下成立?
- 什么样的信源编码方法/算法是最优的?
- **从上界中可以看到，极端情形下即时是最优信源编码，期望码长也比下界长将近1bit（相对下界的额外开销），这种极端情形是什么?**
- **如何减小（甚至消除）这将近1bit的额外开销?**

分组编码

$$\boxed{H(X) \leq L^* \leq L_s < H(X) + 1}$$

- 如何减少每个符号的 1bit 额外开销?
  可以使用分组编码的方式，将开销分担于 n 个符号上
- 定义 L_n 为单个符号的期望码字长度，即:

$$L_n = \frac{1}{n} \sum p(x_1, x_2, \cdots, x_n) l(x_1, x_2, \cdots, x_n) = \frac{1}{n} El(X_{1:n})$$

其中 $X_{1:n}$ 代表 $(X_1, X_2, ...X_n)$
此时有:

$$H(X_{1:n}) \leq El(X_{1:n}) < H(X_{1:n}) + \boxed{1}$$

两边同时除以 n，可以得到

$$\frac{1}{n} H(X_{1:n}) \leq L_n < \frac{1}{n} H(X_{1:n}) + \boxed{\frac{1}{n}}$$

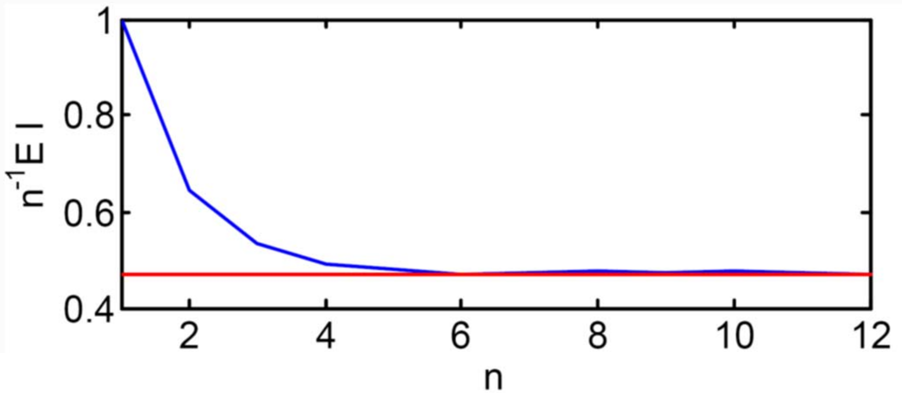
1bit的额外开销被分摊到了n个符号上

分组编码示例

- 若 $X_{1:n}$ 中各随机变量独立同分布（i.i.d.），则有
  $H(X_{1:n}) = \sum H(X_i) = nH(X)$，
  进而
  $$H(X) \leq L_n < H(X) + \frac{1}{n}$$
  当 $n \to \infty$ 时，有 $L_n \to H(X)$

例

$\mathscr{X} = [A; B], \mathbf{p}_X = [0.9; 0.1], H(X_i) = 0.469\text{bits}$



随机过程的熵率

- 如果 $X_{1:n}$ 不是独立同分布，我们期望 $\frac{1}{n}H(X_{1:n})$ 能够收敛到一个极限 $H(\mathscr{X})$

定义

$\{X_i\} = X_1, X_2, \ldots$ 是一个信源随机过程，当如下极限存在时，将其定义为随机过程 $\{X_i\}$ 的**熵率**：

$$H(\mathscr{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_{1:n})$$

- 熵率，即“熵的增长率”，是信源随机过程每出现一个新符号所带来的熵增加量
- 熵率给出了平均每个信源符号编码所需比特数的下界
- 若 X_i 独立同分布，则 $H(\mathscr{X}) = H(X)$
- 若 X_i 不是独立同分布的（如：独立但分布不相同、不独立……），用上述极限所定义的熵率不一定存在

平稳过程

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

定义

若对于任意 l, n 及 $x_i \in \mathscr{X}$，随机过程 $\{X_i\}$ 均满足

$$\Pr(X_{1:n} = x_{1:n}) = \Pr(X_{l+(1:n)} = x_{1:n})$$

则称 $\{X_i\}$ 是平稳随机过程。

定理 4.2.1

若 $\{X_i\}$ 平稳，则 $H(\mathscr{X})$ 存在，且

$$\boxed{H(\mathscr{X})} = \lim_{n\to\infty} \frac{1}{n} H(X_{1:n}) = \lim_{n\to\infty} H\left(X_n \middle| X_{1:(n-1)}\right)$$

每个符号的熵

其中 $\boxed{H'(\mathscr{X})} = \lim_{n\to\infty} H(X_n|X_{1:(n-1)})$ 是熵率的另一种定义形式。

给定过去条件下的末项条件熵

平稳过程（续）

定理 4.2.2

对于平稳随机过程 $\{X_i\}$，$H(X_n|X_{1:(n-1)})$ 随 n 增大而递减，且存在极限 $H'(\mathscr{X})$。

证明.

条件作用使熵减小

$$H(X_n|X_{1:(n-1)}) \boxed{\leq} H(X_n|X_{2:(n-1)})$$
$$\boxed{=} H(X_{n-1}|X_{1:(n-2)})$$

因此 $\{H(X_n|X_{1:(n-1)})\}$ 递减，又因为其非负，故其极限 $H'(\mathscr{X})$ 必然存在。 □

平稳性

Cesàro均值定理

定理 4.2.3 Cesàro 均值

$$a_n \to b \quad \Rightarrow \quad b_n = \frac{1}{n}\sum_{k=1}^{n} a_k \to b$$

证明.

- 因为 $a_n \to b$，则存在数 $N(\varepsilon)$，使得对于任意 $n > N(\varepsilon)$，有 $|a_n - b| \le \varepsilon$
- 因此

$$\begin{aligned}
|b_n - b| = |\frac{1}{n}\sum_{k=1}^{n}(a_k - b)| &\le \frac{1}{n}\sum_{k=1}^{n}|a_k - b| \\
&\le \frac{1}{n}\sum_{k=1}^{N(\varepsilon)}|a_k - b| + \frac{n - N(\varepsilon)}{n}\varepsilon \le \frac{1}{n}\sum_{k=1}^{N(\varepsilon)}|a_k - b| + \varepsilon
\end{aligned}$$

当 n 足够大时，$|b_n - b| \le 2\varepsilon$，因此当 $n \to \infty$ 时 $b_n \to b$

$\square$

定理4.2.1的证明

证明.

- 由熵的链式法则

$$\frac{1}{n}H(X_{1:n}) = \frac{1}{n}\sum_{i=1}^{n}H(X_i|X_{1:(i-1)})$$

- 由定理 4.2.2

当 $n \to \infty$ 时，有 $H(X_n|X_{1:(n-1)}) \to H'(\mathscr{X})$

- 由 Cesàro 均值定理

$$H(\mathscr{X}) = \lim_{n\to\infty}\frac{1}{n}H(X_{1:n}) = \lim_{n\to\infty}H(X_n|X_{1:(n-1)}) = H'(\mathscr{X})$$

□

信源编码定理（Shannon第一定理）

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

定理 5.4.2

针对 $\{X_i\}$ 的信源编码，每符号最小期望码长满足

$$\frac{1}{n}H\left(X_{1:n}\right) \leq L_n^* < \frac{1}{n}H\left(X_{1:n}\right) + \frac{1}{n}$$

若 $\{X_i\}$ 是平稳随机过程，则

$$L_n^* \to H(\mathscr{X})$$

其中 $H(\mathscr{X})$ 为 $\{X_i\}$ 的熵率

- 上述定理亦被称作香农第一定理
- 若 $\{X_i\}$ 独立同分布，则 $H(\mathscr{X}) = H(X)$
- 延伸思考：对于非独立同步分布 $\{X_i\}$，如何计算其熵率？如：Markov 过程，隐 Markov 过程……

阶段小结

- **Huffman 码**
  - 是一种 Bottom-up 编码算法
  - 具备最小期望码长意义下的最优性
- 通过分组编码:

$$\frac{1}{n}H\left(X_{1:n}\right) \leq L_n^* < \frac{1}{n}H\left(X_{1:n}\right) + \frac{1}{n}$$

- 若 $\{X_i\}$ 为平稳随机过程:

$$H(\mathscr{X}) = \lim_{n\to\infty}\frac{1}{n}H(X_{1:n}) = \lim_{n\to\infty}H(X_n|X_{1:(n-1)})$$

结束