Huffman Codes

Problem 5.1

Given source symbols and their probabilities of occurence, how to design an optimal source code (prefix code and the shortest on average)?

Huffman Codes

- Merge the D symbols with the smallest probabilities, and generate one new symbol whose probability is the summation of the D smallest probabilities.
- Assign the D corresponding symbols with digits $0, 1, \ldots, D-1$, then go back to Step 1.

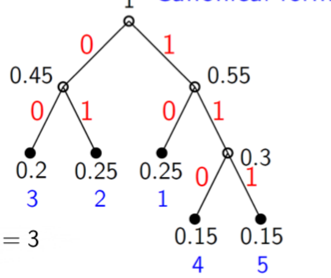
Repeat the above process until D probabilities are merged into probability 1.

Examples

Example 1

x	p(x)	C(x)
1	0.25	10
2	0.25	01
3	0.2	00
4	0.15	110
5	0.15	111

Canonical form



Validations:

$\ell(1) = \ell(2) = \ell(3) = 2, \ell(4) = \ell(5) = 3$

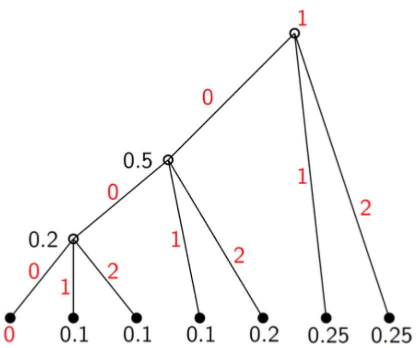
$L = \sum \ell(x)p(x) = 2.3\text{bits}$

$H_2(X) = -\sum p(x)\log_2 p(x) = 2.29\text{bits}$

$L \geq H_2(X)$

Example 2 ($D \geq 3$)

x	p(x)	C(x)
1	0.25	1
2	0.25	2
3	0.2	02
4	0.1	01
5	0.1	002
6	0.1	001
Dummy	0	000



Validations:

$L = \sum \ell(x)p(x) = 1.7$ ternary digits

$H_3(X) = -\sum p(x)\log_3 p(x) \approx 1.55$ ternary digits

Example 2

x	p(x)
1	0.25
2	0.25
3	0.2
4	0.1
5	0.1
6	0.1
Dummy	0

At one time, we merge D symbols, and at each stage of the reduction, the number of symbols is reduced by $D-1$. We want the total # of symbols to be $1 + k(D-1)$. If not, we add dummy symbols with probability 0.

$\mathcal{D} = \{0, 1, 2\}$

Optimality of Huffman Codes

Lemma 5.8.1

For any distribution, the optimal prefix codes (with minimum expected length) should satisfy the following properties:

1. *If $p_j > p_k$, then $\ell_j \leq \ell_k$.*
2. *The two longest codewords have the same length.*
3. *There exists an optimal prefix code, such that two of the longest codewords differ only in the last bit and correspond to the two least likely symbols.*

概率越大，编码长度越短
两个最长的编码同长
存在一种最优prefix code，最长的编码仅在最后一个bit不同.

$\Rightarrow$ If $p_1 \geq p_2 \geq \cdots p_m$, then there exists an optimal code with $\ell_1 \leq \ell_2 \leq \cdots \ell_{m-1} = \ell_m$, and codewords $C(x_{m-1})$ and $C(x_m)$ differ only in the last bit. (canonical codes)

- 1. If $p_j > p_k$, then $\ell_j \leq \ell_k$.

Proof.

Suppose that C_m is an optimal code. Consider C'_m, with the codewords j and k of C_m interchanged. Then

$$\underbrace{L(C'_m) - L(C_m)}_{\geq 0} = \sum p_i \ell'_i - \sum p_i \ell_i$$
$$= p_j \ell_k + p_k \ell_j - p_j \ell_j - p_k \ell_k$$
$$= \underbrace{(p_j - p_k)}_{>0} (\ell_k - \ell_j)$$

Thus, we must have $\ell_k \geq \ell_j$. □

- 2. The two longest codewords have the same length.

设 $C(j)$ 和 $C(k)$ 是最长的.

假设 $l_j < l_k$. 令 $m = l_j$, $m+n = l_k$, 则有:

$C(j): b_1 b_2 \dots b_m$

$C(k): b_1' b_2' \dots b_m' b_{m+1}' \dots b_{m+n}'$

构造 C': $\begin{cases} C'(i) = C(i), & i \neq k \\ C'(k) = b_1' b_2' \dots b_m' \end{cases}$,即第k个code砍了部分.

C'是prefix code 且 $L(C') < L(C)$, 即C不是optimal code. 矛盾

- 3. There exists an optimal prefix code, such that two of the longest codewords differ only in the last bit and correspond to the two least likely symbols.

Proof.

If there is a maximal-length codeword without a sibling, we can delete the last bit of the codeword and still preserve the prefix property. This reduces the average codeword length and contradicts the optimality of the code. Hence, every maximum-length codeword in any optimal code has a sibling. Now we can exchange the longest codewords s.t. the two lowest-probability source symbols are associated with two siblings on the tree, without changing the expected length. □

设C为optimal prefix code

$C(j): b_1 b_2 \dots b_m$

$C(k): b_1' b_2' \dots b_m'$

令 $C'(k): b_1 b_2 \dots \bar{b}_m$, 即最后一位与$C(j)$不同.

将 $C(k)$用$C'(k)$代替, 得到C'.

此时保持有 $L(C') = L(C)$

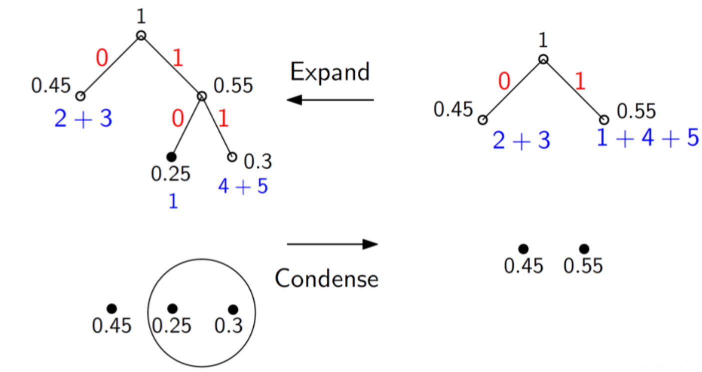
① 若 $C'(k) \in C$, 则C即为所需

② 若 $C'(k) \notin C$,

假如$\exists i$, $C(i)$是$C'(k)$的前缀, 那么$C(i)$也同时是$C(j)$的前缀, 与C为prefix code矛盾.

所以 C'是prefix code.

- We prove the optimality of Huffman codes by induction. Assume binary code in the proof.



Proof.

For $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ with $p_1 \geq p_2 \geq \cdots \geq p_m$, we define the Huffman reduction $\mathbf{p}' = (p_1, p_2, \ldots, p_{m-1+p_m})$ over an alphabet size of $m-1$. Let $C^*_{m-1}(\mathbf{P}')$ be an optimal Huffman code for $\mathbf{p}'$, and let $C^*_m(\mathbf{p})$ be the canonical optimal code for $\mathbf{p}$. □

Key idea.

expand C^*_{m-1} to $C_m(\mathbf{p}) \Rightarrow L(C_m) = L(C^*_m)$

$$\begin{array}{ccc|cc}
 & C^*_{m-1}(\mathbf{p}') & & C_m(\mathbf{p}) & \\
p_1 & w'_1 & l'_1 & w_1 = w'_1 & l_1 = l'_1 \\
p_2 & w'_2 & l'_2 & w_2 = w'_2 & l_2 = l'_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
p_{m-2} & w'_{m-2} & l'_{m-2} & w_{m-2} = w'_{m-2} & l_{m-2} = l'_{m-2} \\
p_{m-1}+p_m & w'_{m-1} & l'_{m-1} & w_{m-1} = w'_{m-1}0 & l_{m-1} = l'_{m-1}+1 \\
 & & & w_m = w'_{m-1}1 & l_m = l'_{m-1}+1
\end{array}$$

$$\begin{array}{ccc|cc}
 & C_{m-1}(\mathbf{p}') & & C^*_m(\mathbf{p}) & \\
p_1 & w'_1 & l'_1 & w_1 = w'_1 & l_1 = l'_1 \\
p_2 & w'_2 & l'_2 & w_2 = w'_2 & l_2 = l'_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
p_{m-2} & w'_{m-2} & l'_{m-2} & w_{m-2} = w'_{m-2} & l_{m-2} = l'_{m-2} \\
p_{m-1}+p_m & w'_{m-1} & l'_{m-1} & w_{m-1} = w'_{m-1}0 & l_{m-1} = l'_{m-1}+1 \\
 & & & w_m = w'_{m-1}1 & l_m = l'_{m-1}+1
\end{array}$$

expand $C^*_{m-1}(\mathbf{p}')$ to $C_m(\mathbf{p})$

$$L(\mathbf{p}) = L^*(\mathbf{p}') + p_{m-1} + p_m$$

condense $C^*_m(\mathbf{p})$ to $C_{m-1}(\mathbf{p}')$

$$L^*(\mathbf{p}) = L(\mathbf{p}') + p_{m-1} + p_m$$

$$L(\mathbf{p}) = L^*(\mathbf{p}') + p_{m-1} + p_m$$
$$L^*(\mathbf{p}) = L(\mathbf{p}') + p_{m-1} + p_m$$

$$\underbrace{(L(\mathbf{p}') - L^*(\mathbf{p}'))}_{\geq 0} + \underbrace{(L(\mathbf{p}) - L^*(\mathbf{p}))}_{\geq 0} = 0$$

Thus, $L(\mathbf{p}) = L^*(\mathbf{p})$. Minimizing the expected length $L(C_m)$ is equivalent to minimizing $L(C_{m-1})$. The problem is reduced to one with $m-1$ symbols and probability masses $(p_1, p_2, \ldots, p_{m-1} + p_m)$. Proceeding this way, we finally reduce the problem to two symbols, in which case the optimal code is obvious.

Expand: 将节点展开

Condense: 将节点合起来

① Huffman code 在 $|X|=2$ 时是最优的

② 假设在 $|X|=m$ 时最优,

则在 $|X|=m+1$ 时,重排概率 $p_1 \geq p_2 \geq \cdots \geq p_{m+1}$

进行condense,则概率分布为 $p_1, p_2, \ldots, p_{m-1}, p_m+p_{m+1}$

在canonical code的tree中,概率最低的两个节点在同一父亲节点下. condense后仍是Huffman code.

H-code for $(p_1, p_2, \ldots p_m, p_{m+1})$ optimal prefix code $\longrightarrow$ H-code for $(p_1, p_2, \ldots p_{m-1}, p_m+p_{m+1})$ prefix code

C_{m+1} : H-code for $(p_1, p_2, \ldots p_m, p_{m+1})$

C^*_m : H-code for $(p_1, p_2, \ldots, p_m+p_{m+1})$ (optimal)

C^*_{m+1} : Optimal code of C-form for $(p_1, \ldots, p_{m+1})$

C_m : condensed from C^*_{m+1}, prefix code for $(p_1, \ldots p_m+p_{m+1})$

$$L(C_{m+1}) = \sum_{i=1}^{m-1} p_i l_i + p_m l_m + p_{m+1} l_{m+1} \qquad (l_m = l_{m+1})$$

$$L(C^*_m) = \sum_{i=1}^{m-1} p_i l_i + (p_m + p_{m+1})(l_m - 1)$$

$\Rightarrow L(C^*_m) = L(C_{m+1}) - p_m - p_{m+1}$

同理: $L(C_m) = L(C^*_{m+1}) - p_m - p_{m+1}$

$$\Rightarrow \underbrace{L(C^*_m) - L(C_m)}_{\leq 0} = \underbrace{L(C_{m+1}) - L(C^*_{m+1})}_{\geq 0}$$

故 $L(C^*_m) = L(C_m)$, $L(C_{m+1}) = L(C^*_{m+1})$

Theorem 5.8.1

Huffman coding is optimal, that is, if C^* is a Huffman code and C' is any other uniquely decodable code, $L(C^*) \leq L(C')$.

Remark

Huffman coding is a *greedy algorithm* in which it merges the two least likely symbols at each step.

LOCAL OPT $\rightarrow$ GLOBAL OPT