Huffman Codes

Let $\Sigma = \{0,1\}$, $\mathcal{X}$, $p(\cdot)$.

Huffman algorithm: Number of leaves equalts $|\mathcal{X}| = N$. leaves have prefix free labels

add pic

Build tree from bottom up.

1. label $\mathcal{X}$: $x_1, \dots, x_n$ such that

$$p(x_1) \geq p(x_2) \geq \cdots \geq p(x_N)$$

2. (If $N = 1$ stop.) Otherwise build

$$\begin{array}{cc} x_1^{\text{new}} = x_1^{\text{old}}, \dots, & x_{N-2}^{\text{new}} = x_{N-2}^{\text{old}}, x_{N-1}^{\text{new}} = \{x_{N-1}^{\text{old}}, x_N^{\text{old}}\} \\ p(x_1) & p(x_{N-2}) p(x_{N-1}^{\text{new}}) = p(x_{N-1}^{\text{old}}) + p(x_N^{\text{old}}) \end{array}$$

wit noew collection continue at 1 ($N \leftrightarrow N-1$)

Example. $N = 10$

draw the tree step by step

i	1	2	3	4	5	6	7	8	9	10
$p(x_i)$	0.2	0.2	0.15	0.15	0.1	0.05	0.05	0.04	0.03	0.03

i	1	2	3	4	5	(9,10)	6	7	8
$p(x_i)$	0.2	0.2	0.15	0.15	0.1	0.06	0.05	0.05	0.04

i	1	2	3	4	5	(7,8)	(9,10)	6
$p(x_i)$	0.2	0.2	0.15	0.15	0.1	0.09	0.06	0.05

i	1	2	3	4	(6,9,10)	5	(7,8)
$p(x_i)$	0.2	0.2	0.15	0.15	0.11	0.1	0.09

i	1	2	(5,7,8)	3	4	(6,9,10)
$p(x_i)$	0.2	0.2	0.19	0.15	0.15	0.11

i	(4,6,9,10)	1	2	(5,7,8)	3
$p(x_i)$	0.26	0.2	0.2	0.19	0.15

i	(3,5,7,8)	(4,6,9,10)	1	2
$p(x_i)$	0.34	0.26	0.2	0.2

i	(1,2)	(3,5,7,8)	(4,6,9,10)
$p(x_i)$	0.4	0.34	0.26

i	(3,4,5,6,7,8,9,10)	(1,2)
$p(x_i)$	0.6	0.4

i	(1,2,3,4,5,6,7,8,9,10)
$p(x_i)$	1

Lemma 0.1. *Let $C : \mathcal{X} \to \{0,1\}^+$ be an optimal prefix code. $W = \{w \in C(\mathcal{X}) : \ell(w) = \ell_{\max}\}$. Then*

1. *If $p(x) > p(y)$, then $\ell(C(x)) \leq \ell(C(y))$.*
2. *If $v, w \in C(\mathcal{X})$ are longest codewords that is $\ell(w') \leq \ell(v) \leq \ell(w)$ for all $w' \in C(\mathcal{X}) \setminus \{v, w\}$, then $\ell(v) = \ell(w)$. $|W| \geq 2$.*

3. *C can be modiefied into another optimal prefix code $\tilde{C}$ which has the additonal property that among the codewords of maximal length [there are at least 2 by (2)] [$\equiv$ the elements of W] there are two which are siblings (differ only in the last bit) and corresopnd to two least likely symbols (elements of $\mathcal{X}$).*

Proof. 1. $\ell_x = \ell(C(x))$ exchange codewords new

$$\begin{aligned} C'(z) &= C(z),\ z \neq x, y \\ C'(x) &= C(y) \\ C'(y) &= C(x) \end{aligned}$$

$$L_{C'} \geq L_C$$

$$\begin{aligned} 0 \leq L_{C'} - L_C &= p(x)\ell_y - p(x)\ell_x + p(y)\ell_x - p(y)\ell_y \\ &= (p(x) - p(y))(\ell_y - \ell_x) \implies \ell_y - \ell_x \geq 0 \end{aligned}$$

2. Suppose $\ell(v) < \ell(w)$. Reduce w to the length of v: (tape prefix of w with length $\ell(v)$). Get new prefix code with $<$ expected length $\lightning$

3. Claim: If $w \in W$ and $\tilde{w}$ its sibling [siblings: $v0, v1$], then $\tilde{w} \in W$.

   Proof: Assume there exists a $w \in W$ such that $\tilde{w} \notin W$. Let v be their parent. Replace w by v: new collection of codewords. [If $C(x) = w$ then $C'(x) = v$, $C'(z) = C(z)$ for all $z \in \mathcal{X}\backslash\{x\}$.] again prefix code, $L_{C'} < L_C$ $\lightning$

   By (1), there exist $x, y \in \mathcal{X}$ least likely with $C(x), C(y) \in W$. Let $C(x) = v$, $C(y) = w$, $v, w \in W$. $\tilde{v}, \tilde{w}$ their siblings in W. If $\tilde{v} = w$ OK, otherwise there exists $\tilde{x} \in \S$: $C(\tilde{x}) = \tilde{v}$ $\tilde{x} \neq v, w$

$$\begin{aligned} \tilde{C}(y) &= \tilde{v} \\ \tilde{C}(\tilde{x}) &= w \\ \tilde{C}(z) &= C(z) \end{aligned}$$

   for all $z \in \mathcal{X} \setminus \{y, \tilde{x}\}$. Now prefix code, $L_{\tilde{C}} = L_C$

□

Definition. *Codes which are optimal and have properties (1),(2),(3) [siblings at the bottom] are canonical.*

Theorem 0.2. *If $C^* : \mathcal{X} \to \{0,1\}^+$ is a Huffman-code for $\mathcal{X}$, p, then it is optimal*

Proof. $|\mathcal{X}| = N$, induction on N

Algorithm: If C_N^* is a Huffman code for $\mathcal{X}^N = \{x_1, \ldots, x_N\}$, then it is abtained from a Huffman-code for $\{x'_1, \ldots, x'_{N-2}, x'_{N-1}\}$ where

$$p(x_1) \geq p(x_2) \geq \cdots \geq p(x_N)$$

and $x'_k = x_k$ for $k = 1, \ldots, N-2$ and $x'_{N-1} = \{x_{N-1}, x_N\}$ with probabilites $p(x'_k) = p(x_k)$ and $p(x'_{N-1}) = p(x_{N-1}) + p(x_N)$ respectively. $C_N^*(x_k) = C_{N-1}^*(x'_k)$ and $C_N^*(x_{N-1}) = C_{N-1}^*(x'_{N-1})0$ and $C_N^*(x_N) = C_{N-1}^*(x'_{N-1})1$

Induction on $N \geq 2$: $N = 2$: $C_2^*(x_1) = 0$, $C_2^*(x_2) = 1$ add pic

$N-1 \to N$: Let $p^{(N)} = (p_1, \ldots, p_N)$ ordered $\geq$. (Note: $p_i = p(x_i)$) Let $p^{(N)'} = (p_1, \ldots, p_{N-2}, p_{N-1} + p_N)$. C_{N-1}^* Huffman code for $p^{(N)'}$ optimal by assumption construct C_N^* from C_{N-1}^*.

$$L_{C_N^*} = L_{C_{N-1}^*} + p_{N-1} + p_N$$

Now let $\bar{C_N}$ be a canonical code for $p^{(N)}$ $L_{\bar{C}_N} \leq L_{C_N^*}$

$$\begin{aligned} \bar{C}_{N-1}(x'_k) &= \bar{C}_N(x_k),\ k = 1, \ldots, N-2 \\ \bar{C}_{N-1}(x'_{N-1}) &= \text{ parent of the siblings } \bar{C}_N(x_{N-1}), \bar{C}_N(x_N) \end{aligned}$$

$\bar{C}_{N-1}$ is a prefix code for $p^{(N)'}$

$$L_{\bar{C}_{N-1}} = L_{\bar{C}_N} - p_{N-1} + p_N$$

$\square$