DATE 25.9.17

dec 12: Huffman Coding

Recap

* Extended Kraft inequality
* Shannon code
  $$\ell(x) = \left\lceil \log \frac{1}{P(x)} \right\rceil$$
* $L \geq H_D(X)$
  $H_D(X) \leq L < H_D(X)+1$
* Wrong code

TODAY

* { KI for uniquely decodable codes
* { Huffmann Coding
* { Shannon-fano-Elias Code

Thm (Mc Milan) The codewords of any uniquely decodable (UD) code must satisfy the KI.

$$\sum_{x \in \mathcal{X}_0} D^{-\ell(x)} \leq 1 \quad \rightarrow (1) \qquad (|\mathcal{X}_0| < \infty)$$

Conversely, given $\{\ell(x), x \in \mathcal{X}_0\}$ satisfying (1), an UD can be found.

Pf: Consider $\left(\sum_{x \in \mathcal{X}_0} D^{-\ell(x)}\right)^k = \sum_{x_1} \cdots \sum_{x_k} D^{-(\ell(x_1)+\cdots+\ell(x_k))}$

$$= \sum_{\underline{x}} D^{-\ell(\underline{x})} = \sum_{j=1}^{k\ell_{max}} a_j D^{-j}$$

where a_j = # of strings of the k codewords from the source code with length j.

Note that no two strings of k ~~codewords~~ source symbols can result in the same coded string of length j

Hence $a_j \leq D^j$

$$\therefore \left(\sum_{x \in \mathcal{X}} D^{-\ell(x)}\right)^k = \sum_{j=1}^{k\ell_{max}} a_j D^{-j} \leq \sum_{j=1}^{k\ell_{max}} D^j D^{-j}$$

$$\therefore \sum_{x \in \mathcal{X}_0} D^{-\ell(x)} \leq \left(k\ell_{max}\right)^{1/k} = k\ell_{max}$$ for any $k \in \mathbb{N}$.

This must be true for any k and hence true as $k \to \infty$.

$$\therefore \sum_{x \in \mathcal{X}} D^{-l(x)} \leq \lim_{k \to \infty} (k l_{max})^{1/k} = 1$$

$$\Rightarrow \sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1$$

Converse holds, since one can find a prefix code if ① is satisfied and every prefix code is a uniquely decodable code.

Corollary: The above result also holds if $\mathcal{X}$ is countably infinite $|\mathcal{X}| = \infty$.

Pf: Let $|\mathcal{X}| = \infty$, Take a finite subset $S \subseteq \mathcal{X}$ $|S| < \infty$. Let $\mathcal{C}$ be a U.D code for $\mathcal{X}$, and $\mathcal{C}_S$ be the subset of $\mathcal{C}$ obtained by restricting $\mathcal{C}$ to the set S. The resultant code $\mathcal{C}_S$ also should be U.D.

$$\Rightarrow \sum_{x \in S} D^{-l(x)} \leq 1$$

Let $S = \{x_1, x_2, \ldots x_s\}$ ie, $|S| = s$.

$$\Rightarrow \lim_{s \to \infty} \sum_{i=1}^{s} D^{-l(x_i)} \leq 1$$

The converse holds true, cause one can construct a prefix free code if $\sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1$, $(|\mathcal{X}| = \infty)$ and prefix free code is UD.

DATE

The Huffmann code

Begin with some examples. $|\mathbb{D}| = 2$.

(a)

$C(x)$	$x \in \mathcal{X}_0$	$p(x)$	A	B	C	D.
01	1	0.25	0.3	0.45	0.55	1
10	2	0.25	0.25	0.3	0.45	
11	3	0.2	0.25	0.25		
000	4	0.15	0.2			
001	5	0.15				

$L(c) = (0.7)2 + (0.3)3$
$= 2.3$ bits

$|\mathbb{D}| = 3$.

(b)

$C(x_0)$	$x \in \mathcal{X}_0$	$p(x)$	A	B
1	1	0.25	0.5	1
2	2	0.25	0.25	
00	3	0.2	0.25	
01	4	0.15		
02	5	0.15		

$L(c) = (0.5)1 + (0.5)2$
$= 1.5$ ternary digits/ Source symbol

5 → 5 → 3 → 1.

(c) $|\mathbb{D}| = 3$

$C(x_0)$	$x \in \mathcal{X}_0$	$p(x)$	A		
1	1	0.25	0.25	0.5	1
2	2	0.25	0.25	0.25	
00	3	0.2	0.2	0.25	
02	4	0.1	0.2		
010	5	0.1	0.1		
011	6	0.1			
012	Dummy	0			

$L(c) = (0.5)1 + (0.3)2 + (0.2)3$
$= 1.7$ ternary digits/ Source symbol.

7 → 5 → 3 → 1

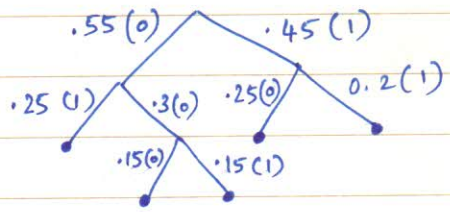
Added a dummy so that you end up with 3 nodes.

You would want $m = 1 \bmod (q-1)$.

where $q = |\mathbb{D}|$. and $m = |\mathcal{X}_0|$

DATE

Example ④



Optimality of the Huffmann code.

(focus on the binary, finite $\mathcal{D} = \{0,1\}$ case)

lemma: Given any $p(x)$, $\exists$ an optimal prefix code satisfying:

1) the lengths l_i satisfy
$$p_j > p_i \Rightarrow l_j \leq l_i$$
2) the two longest codewords have same length
3) 2 of the longest codewords differ only in the last bit and correspond to the 2 least likely symbols.

① Suppose C_m is optimal

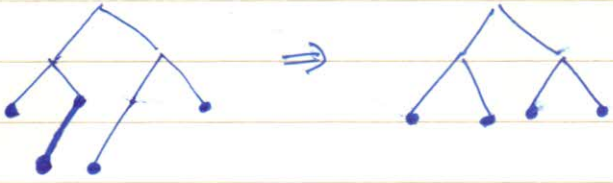
C_m	C_m'
$p_i > p_j$	p_i, p_j
(say) $l_i > l_j$	$l_i' = l_j$, $l_j' = l_i$
	rest remain

$$L(C_m) - L(C_m') = p_i l_i + p_j l_j - p_i l_j + p_j l_i$$
$$= (p_i - p_j)(l_i - l_j) > 0$$

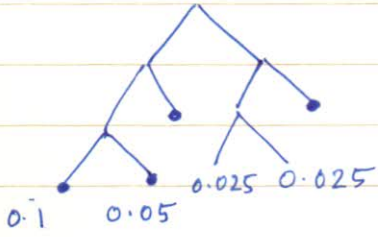
$\Rightarrow C_m'$ is strictly better code than C_m, $\therefore$ C_m cannot be optimal (Contradiction)

DATE

② If not true, the picture would look like



③



Clearly if there are 2s codewords of the same length, these corespond to the smallest 2s probabilities.

Shuffling these codewords will not change the average code length $L(C_m)$. Hence WLOG we can assume ③ holds.

Claim: The Huffmann procedure yields optimal code.

$$L(C_m) - L(C_{m-1}) = l_{max}(P_m + P_{m-1}) - (l_{max} - 1)(P_m + P_{m-1})$$

$$= P_m + P_{m-1} = \text{Constant}$$

$P_1, P_2 \cdots P_{m-1}\ P_m$

$P_1, P_2, \cdots P_{m-2}, (P_m + P_{m-1})$

Cause there exists a prefix free code that is optimal with properties ①, ②, ③.

But C_{m-1} is a code for $(m-1)$-symbol source.

∴ To optimize $L(C_m)$ it is necessary & sufficient that $L(C_{m-1})$ is optimal.

For C_2 there is only one optimal code, therefore the claim follows.