Lecture 10 The Huffman Code

Corresponding to section 5.6-5.8 of the textbook

September 30th, 2022

Outline

Binary Huffman code

1. Take the two least probable symbols in the alphabet. These two symbols will be given the longest codewords, which will have equal length, and differ only in the last digit.
2. Combine these two symbols into a single symbol, and repeat.

Example

Consider a random variable X *taking values in the set* $\mathcal{X} = \{1, 2, 3, 4, 5\}$ *with probabilities* 0.25, 0.25, 0.2, 0.15, 0.15.

Codeword Length	Codeword	X	Probability				
2	01	1	0.25	0.3	0.45	0.55	1
2	10	2	0.25	0.25	0.3	0.45	
2	11	3	0.2	0.25	0.25		
3	000	4	0.15	0.2			
3	001	5	0.15				

This code has average length 2.3 bits.

Example

Consider a ternary code for the same random variable.

Codeword	X	Probability		
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		

This code has average length 1.5 ternary bits.

Example

If $D \geq 3$, we may not have a sufficient number of symbols so that we can combine them D at a time. In such case, we add dummy symbols to the end of the set of symbols. The dummy symbols have probability 0 and are inserted to fill the tree. Since at each stage of reduction, the number of symbols is reduced by $D - 1$, we want the total number of the symbols to be $1 + k(D_1)$, where k is the number of merges. Hence, we add enough dummy symbols so that the total number of symbols is of the form. For example:

Codeword Length	Codeword	X	Probability				
2	01	1	0.25	0.3	0.45	0.55	1
2	10	2	0.25	0.25	0.3	0.45	
2	11	3	0.2	0.25	0.25		
3	000	4	0.15	0.2			
3	001	5	0.15				

This code has average length 1.7 tenary bits.

Lemma

Let $\mathcal{C}$ be an optimal binary prefix code with codeword lengths l_i, $i = 1, \ldots, M$, for a source with alphabet $\mathcal{X} = \{a_1, \ldots, a_M\}$ and symbol probabilities $p_1, \ldots, p_M$. We assume, without loss of generality, that

$$p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_M,$$

and that any group of source symbols with identical probability is listed in order of increasing codeword length (i.e., if $p_i = p_{i+1} = \cdots = p_{i+s}$, then $l_i \leq L_{i+1} \leq \cdots \leq l_{i+s}$). Then the following properties hold.

1. *Higher probability source symbols have shorter codewords: $p_i > p_j$ implies $l_i \leq l_j$, for $i, j = 1, \ldots, M$.*
2. *The two least probable source symbols have codewords of equal length: $l_{M-1} = l_M$.*
3. *Among the codewords of length l_M, two of the code words are identical except in the last digit.*

Proof.

(1) If $p_i > p_j$ and $l_i > l_j$, then it is possible to construct a better code $\mathcal{C}'$ by interchanging ("swapping") codewords i and j of $\mathcal{C}$, since

$$\begin{aligned} L(\mathcal{C}') - L(\mathcal{C}) &= p_i l_j + p_j l_i - (p_i l_i + p_j l_j) \\ &= (p_i - p_j)(l_j - l_i) \\ &< 0. \end{aligned}$$

Hence code $\mathcal{C}'$ is better than code $\mathcal{C}$, contracting the fact that $\mathcal{C}$ is optimal.

Proof.

(2) We first know that $l_{M-1} \leq l_M$, since

- If $p_{M-1} > p_M$, then $l_{M-1} \leq l_M$ by (1) above.
- If $p_{M-1} = p_M$, then $l_{M-1} \leq l_M$ by our assumption about the ordering of codewords for source symbols with identical probability.

Now, if $l_{M-1} < l_M$, we may delete the last digit of codeword M, and the deletion cannot result in another codeword sinvr $\mathcal{C}$ is a prefix code. Thus, the deletion forms a new prefix code with a better average codeword length than $\mathcal{C}$, contradicting the fact that $\mathcal{C}$ is optimal. Hence, we must have that $l_{M-1} = l_M$.

Proof.

(3) Among the codewords of l_M, if no two codewords agree in all digits except the last, then we may delete the last digit in all such codewords to obtain a better codeword.

Summarizing, we have shown that if $p_1 \geq p_2 \geq \cdots \geq p_m$, there exists an optimal code with $l_1 \leq l_2 \leq \cdots \leq l_{m-1} = l_m$, and codewords $C(x_{m-1})$ and $C(x_m)$ differ only in the last bit. Such codes are called canonical codes. □

Theorem

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

Example

An instantaneous code has word lengths $l_1, \cdots, l_m$, which satisfy the strict inequality

$$\sum_{i=1}^{m} D^{-l_i} < 1.$$

The code alphabet is $\mathcal{D} = \{0, 1, \cdots, D-1\}$. Show that there exist arbitrarily long sequences of code symbols in $\mathcal{D}^*$ which cannot be decoded into sequences of codewords.

Example

Assume that a sequence of symbols from the random variable X as below using the code C_3. Imagine picking one bit at random from the binary encoded sequence $\mathbf{x} = c(x_1)c(x_2)c(x_3)\cdots$. What is the probability that this bit is a 1?

C_3:

a_i	$c(a_i)$	p_i	$h(p_i)$	l_i
a	0	$1/2$	1.0	1
b	10	$1/4$	2.0	2
c	110	$1/8$	3.0	3
d	111	$1/8$	3.0	3