Information, Codes and Ciphers

Hao Ren

November 7, 2020

3 Compression Coding

3.1 Variable Length Encoding

3.1.1 Definition

a source S	with q source symbols	$s_1, s_2, \cdots, s_q$,
	with probabilities	$p_1, p_2, \cdots, p_q$,
encoded by a code C	with q codewords	$c_1, c_2, \cdots, c_q$,
	of lengths	$l_1, l_2, \cdots, l_q$.

- with a radix r codewords,
- variable length codes,
- not channel noise for source coding.

3.1.2 UD and I-code

A code C is

UD uniquely decodable codes if it can always be decoded unambiguously,

I-code instantaneous if no codeword is the prefix of others.

3.1.3 Comma Codes

The standard comma code of length n is

- a code which every codeword has length $\leq n$,
- a code which every codeword contains at most one 0,
- and if a codeword contains 0 then 0 must be the final symbol in the codeword.

3.1.4 Decision Trees

3.1.5 The Kraft-McMillan Theorem

Theorem 3.1 (The Kraft-McMillan Theorem)
A UD-code of radix r with q codewords $c_1, c_2, \cdots, c_q$ of lengths $l_1 \leq l_2 \leq \cdots \leq l_q$ exists
if and only if *an I-code with the same parameters exists*
if and only if

$$K = \sum_{i=1}^{q} \frac{1}{r^{l_i}} \leq 1.$$

3.1.6 Length and Variance

The expected or **average length** of codewords is given by

$$L = \sum_{i=1}^{q} p_i l_i$$

and the **variance** is given by

$$V = \sum_{i=1}^{q} p_i l_i^2 - L^2.$$

Our aim is to minimise L for a given source S and, if more than one code C gives this value, to minimise V.

Theorem 3.2 (Minimal UD-codes)
Let C be a UD-code with minimal expected length L for the given source S. Then, after permuting codewords of equally likely symbols if necessary,

- $l_1 \leq l_2 \leq \cdots \leq l_q$ *and*
- $l_{q-1} = l_1$.

Furthermore, if C is instantaneous, then

- c_{q-1} *and* c_q *differ only in their last place.*

If C is binary, then

- $K = \sum_{i=1}^{q} 2^{-l_i} = 1$.

3.2 Huffman's Algorithm

3.2.1 Huffman Coding

To compute Huffman prefix-free code:

- Count character frequencies ps for each symbol s in file.
- Start with a forest of trees, each consisting of a single vertex corresponding to each symbol s with weight p_s.
- Repeat:
  - select two trees with min weight p_1 and p_2
  - merge into single tree with weight $p_1 + p_2$

Applications JPEG, MP3, MPEG, PKZIP.

Theorem 3.3 (Huffman Code Theorem)
For the given source S, the Huffman algorithm produces a minimum average length UD-code which is an instantaneous code.

Proposition 3.4 (Knuth)
For a Huffman code created by the given algorithm, the average code word length is sum of all the probabilities at child nodes.

3.2.2 Properties of Huffman Codes

1. The place high strategy always produces a minimum variance Huffman code .
2. If there are 2^n equally likely source symbols then the Huffman code is a block code of length n.
3. If for all j, $3p_j \geq 2\sum_{k=j+1}^{q} p_k$ then the Huffman code is a comma code.
4. Small changes in the pi can change the Huffman code substantially, but have little effect on the average length L. This effect is smaller with smaller variance.

Resources