Source Coding

Data and Information Management: ELEN 3015

School of Electrical and Information Engineering,
University of the Witwatersrand

Information Theory

"Cryptography, Information Theory and Error-Correction," Bruen A.A., Forcinito M.A.

Chapter 11

Overview

Lempel-Ziv Coding

Tuts

1. Lempel-Ziv Coding: Introduction

In large, replaced Huffman coding

For English text, LZ obtains 55 % compression, Huffman 43 %

Huffman doesn't exploit statistical dependencies as well as LZ.

Disadvantage of Huffman $\rightarrow$ need to know statistics a priori

Uses: ZIP, UNZIP, etc.

2. Lempel-Ziv Coding: Operation

Parse source stream into segments that are the shortest subsequences not yet encountered.

New subsequences are longer by one symbol than previously encountered sequences $\rightarrow$ compression by storing pointers to data

Each new subsequence not yet encountered will be equal to an old subsequence with a single letter added on at the end.

Lempel-Ziv Encoding: Example

Alphabet $\mathcal{A} = \{x, y\}$

Stream:

xyyyxxyxxxxyxyxxyxxxx

Lempel-Ziv Coding: Example

Proceeding from left, break up remaining stream into segments that represent the shortest subsequences not yet encountered.

Index subsequences

x	y	yy	xx	yx	xxx	yxy	xxy	xxxx
1	2	3	4	5	6	7	8	9

Lempel-Ziv Coding: Example

Format subsequences into $i \cdot a$, $i \to$ index, $a \in \mathcal{A}$

Label	0x	0y	2y	1x	2x	4x	5y	4y	6x
Slots	01	2	3	4	5	6	7	8	9

Empty string $\{\}$ corresponds to 0, also indicate start of text.

1. Source extension

Given a source Γ with source words chosen from $\mathcal{A}$ we can construct a new source, called the s$^{\text{th}}$ order extension of Γ, denoted by Γ^s.

Alphabet of $\Gamma^s \rightarrow$ all possible strings of length s chosen from the alphabet $\mathcal{A}$.

If Z is a word in Γ^s then $Z = y_1, y_2, \ldots, y_s$ with $y_1, y_2, \ldots, y_s$ in $\mathcal{A}$.

Probability of $Z = Pr(y_1) \cdots Pr(y_s)$.

Question Exam 2008

Consider the following string of data:

BDBGFBAGBDFGFABGGBGABFAABADBAA.

- Determine the entropy of the source based on the sample string. [3]
- Encode the text using a Huffman code. [7]

Tut Question 1

Carry out the Huffman encoding for the source with probabilities
0.45, 0.2, 0.15, 0.1, 0.1

Tut Question 2

Find a Huffman code for source probabilities 0.1, 0.15, 0.15, 0.2, 0.4

Tut Question 3

Let X be the source which emits heads with a probability 0.8 and tails with a probability 0.2. Find an optimal encoding for X^2, the second extension of X. What is the average word length?

Tut Question 4

Find an optimal encoding for X^3, the third extension of X. What is the average word length?

Tut Question 2

If a source with N source words is encoded as an instantaneous code and the code word lengths are $l_1, l_2, \ldots, l_N$, show that $l_1 + l_2 + \ldots + l_N \geq N \log_2(N)$

Tut Question 2