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

1. Introduction

Consider a source with:

- Alphabet $\mathcal{A} = \{x_1, x_2, \ldots, x_m\}$
- Each symbol x_i has probability p_i, $0 \leq p_i \leq 1$ of occurring in the message.

$p_1 + p_2 + \ldots + p_m = 1$, $0 \leq p_i \leq 1$.

Example of a source - English language

- Alphabet size $m = 26$ (or $m = 27$)
- Probabilities of symbols are well known and tabulated.
- Eg. letter a has probability $p_1 = 0.064$

1. Introduction

Memoryless source:

- Each symbol x_i is an independent and identically distributed random variable (iid).
- Real life sources are seldom memoryless and are modeled as ergodic processes.

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

1. Source extension

Example: Let $\mathcal{A} = \{x_1, x_2\}$ with $p_1 = Pr(x_1) = 0.4$ and $p_2 = Pr(x_2) = 0.6$.

Second extension $\mathcal{A}^2 \Leftrightarrow \mathcal{A}^2 = \{x_1x_1, x_1x_2, x_2x_1, x_2x_2\}$

Probabilities 0.16, 0.24, 0.24 and 0.36.

Sometimes more efficient to encode blocks of consecutive source words rather than individual source words $\Leftrightarrow$ block coding.

By independence:

Entropy of an Extension

If Γ has alphabet $\mathcal{A}$, and Γ^s is the s'th order extension of Γ, then

$$H(\Gamma^s) = sH(\Gamma)$$

3. Encodings

Encoding f: maps source words from $\mathcal{A}$ to a string with symbols in alphabet Y.

Example of an encoding:

- $f \to$ ASCII
- Y might be the binary alphabet ($Y = \{0, 1\}$)
- $\mathcal{A}$ might be the upper-case English alphabet

3. Encodings

Condition for encoding: x_i, x_j with $i \neq j$, $f(x_i) \neq f(x_j)$.

Message $\to$ any string of source words from $\mathcal{A} = \{x_1, x_2, \ldots, x_m\}$.

Consider $M = x_3x_1x_3$.

Encoding $\to$ $f(M) = f(x_3)f(x_1)f(x_3)$

Code words $\to$ strings over Y of the form $f(x_i)$, $1 \leq i \leq m$

Code C $\to$ set of code words $f(x_i)$

Example:
$\mathcal{A}$ consist of the three source words u, v and w
$Pr(u) = 0.3$, $Pr(v) = 0.5$ and $Pr(w) = 0.2$

$$H(\mathcal{A}) \quad =$$

Example:
$\mathcal{A}$ consist of the three source words u, v and w
$Pr(u) = 0.3$, $Pr(v) = 0.5$ and $Pr(w) = 0.2$

$$\begin{aligned} H(\mathcal{A}) &= (0.3)\log_2(1/0.3) + (0.5)\log_2(1/0.5) + (0.2)\log_2(1/0.2) \\ &= 0.5211 + 0.5 + 0.4644 \\ &= 1.4855 \end{aligned}$$

3. Encodings

Encoding f from $\mathcal{A}$ to Y with $Y = \{0, 1\}$ is given as follows:

$$f(u) = 01,\ f(v) = 1 \text{ and } f(w) = 101$$

Then if $m = vu$, $f(m) = f(v)f(u) = 101$.

Average length of an encoded source word:

3. Encodings

Encoding f from $\mathcal{A}$ to Y with $Y = \{0,1\}$ is given as follows:

$$f(u) = 01,\ f(v) = 1 \text{ and } f(w) = 101$$

Then if $m = vu$, $f(m) = f(v)f(u) = 101$.

Average length of an encoded source word:

$(0.3)(2) + (0.5)(1) + (0.2)(3) = 1.7.$

3. Uniquely decipherable

Encoding $f \Leftrightarrow$ uniquely decipherable (u.d.) if there do not exist two different messages M_1 and M_2 with $f(M_1) = f(M_2)$.

Previous example f is not u.d. $\rightarrow f(vu) = f(w) = 101$.

Encoding f is an instantaneous code (or prefix code) if there do not exist two code words x_i and x_j such that $f(x_i)$ is a prefix of $f(x_j)$.

Thus, a prefix code can be uniquely decoded from left to right without "look ahead".

3. Encodings

Lemma: If f is instantaneous, then f is u.d. (Leave the proof)

Lemma: There exist u.d. codes which are not instantaneous. (Leave the proof)

Example: $\mathcal{A} = \{a, b\}$, $f(a) = 1$, $f(b) = 10$

$f(a)$ is a prefix of $f(b)$, but code is still u.d.

Prefix code can be decoded "on line" moving from left to right.

3. Kraft's inequality

Necessary and sufficient condition for the existence of an instantaneous code:

$$\sum_{i=1}^{n} 2^{-l_i} \leq 1$$

where l_i is the word-lengths.

Proof not for examination

3. Maximum information

Theorem: $H(x) \leq \log_2 n$ with equality if and only if $p_1 = p_2 = \ldots p_n = 1/n$ so that X is equiprobable.

In order to maximise the entropy, make the probabilities equal.

(Proof not for examination)

3. McMillan's inequality

Theorem: A necessary and sufficient condition for the existence of a u.d. code C with codewords of length $l_1, l_2, \ldots, l_n$ is

$$\sum_{i=1}^{n} 2^{-l_i} \leq 1$$

(Proof not for examination)

3. Noiseless coding Theorem

Theorem: If a memoryless source has entropy H then the average length of a binary, uniquely decipherable, encoding of that source is at least H.

Moreover, there exist a code having average word-length less than $1 + H$, on the assumption that the emission probability p_i of each source word is positive.

(Proof not for examination)

Block Coding, The Oracle, Yes-No Questions

Go through on own time

Optimal Codes

Not for Examination

Huffman Coding

Huffman Code C:

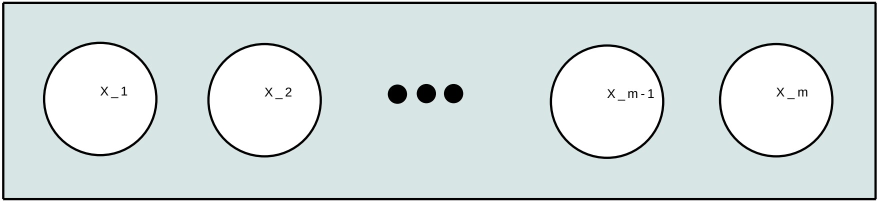
- Prefix code
- $L(C) \leq L(C_1)$ (C_1 any code that is u.d.)

Huffman Coding

Has a source $S = S_0$ with source words $\mathcal{A} = \mathcal{A}_0 = \{x_1, x_2, \ldots, x_m\}$

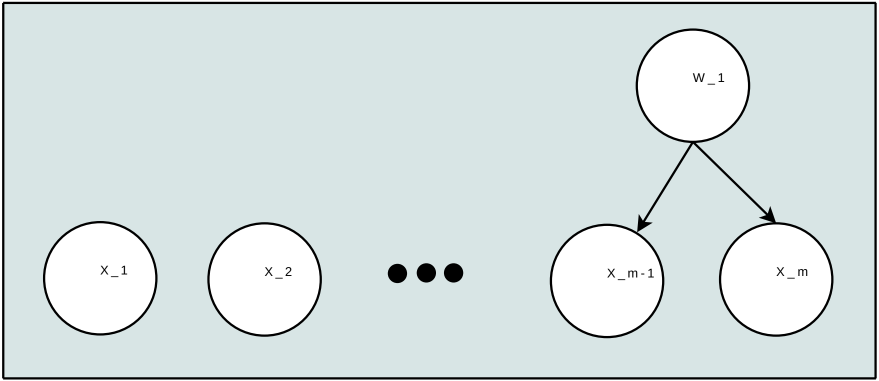
Have that $p_1 \geq p_2 \geq p_3 \ldots \geq p_m$

Huffman Coding



Step 1: Merge two source words with the smallest probability

Thus, merge x_{m-1} and x_m to form new "symbol" W_1 with probability $y = p_m + p_{m-1}$.



Huffman Coding

Algorithm

1. "Combine" two source words with smallest probability into a new source word
2. Construct the resulting graph
3. If number of source words > 1, Go to step 1.