Introduction to source coding

Dr. Yoann Altmann

B39AX – Fall 2023
Heriot-Watt University

Plan

- Types of compression
- Lossless compression
  - Expected code length
  - Prefix codes
  - Optimal codes
  - Shannon source coding theorem (symbol code)
  - Huffman code
- Based on the book:
  “Information Theory,. **Inference**, and **Learning Algorithms”**. David J.C. MacKay (Chap. 4-5)

Expected length of encoded symbol

$$L(X,C) = \sum_{x \in \mathcal{A}} p(x) l(x) = \sum_{i=1}^{I} p_i l_i$$

- $I = |\mathcal{A}|$
- $l_i = l(a_i)$
- $p_i = p(x = a_i)$

- To achieve optimal compression using a uniquely decodable code, we want to minimize $L(X,C)$ (shortest expected length)

Examples

- $H(X) = 1.75$
- $L(X, C_1) = 2 > H(X)$
- Uniquely decodable
- $L(X, C_2) = 1.25 < H(X)$

But not uniquely decodable

- $L(X, C_3) = 1.75 = H(X)$

Uniquely decodable

C_1

a_i	p_i	$c(a_i)$	l_i
a	1/2	00	2
b	1/4	01	2
c	1/8	10	2
d	1/8	11	2

C_2

a_i	p_i	$c(a_i)$	l_i
a	1/2	0	1
b	1/4	1	1
c	1/8	00	2
d	1/8	11	2

C_3

a_i	p_i	$c(a_i)$	l_i
a	1/2	0	1
b	1/4	01	2
c	1/8	011	3
d	1/8	111	3

Low bound for the expected length

- $L(X, C)$ for a uniquely decodable code is bounded below by $H(X)$

$$H(x) \leq L(X, C)$$

If the codelengths are equal to the Shannon information contents

$$l_i = -\log_2 p_i\,, \forall i,$$

the code is optimal.

More generally, an optimal code minimizes $L(X, C)$

Examples

- C_1 and C_3 are uniquely decodable

Are they optimal?

- C_1 is not optimal
- C_3 is optimal

C_1

a_i	p_i	$c(a_i)$	$-log_2(p$	l_i
a	1/2	00	1	2
b	1/4	01	2	2
c	1/8	10	3	2
d	1/8	11	3	2

C_3

a_i	p_i	$c(a_i)$	$-log_2(p$	l_i
a	1/2	0	1	1
b	1/4	01	2	2
c	1/8	011	3	3
d	1/8	111	3	3

Kraft inequality

- If a code is uniquely decodable, its lengths satisfy the following Kraft inequality

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

- This becomes an equality for an optimal code
- If a code does not satisfy this inequality, it is not uniquely decodable

Example

- $H(X) = 1.75$

C_2

a_i	p_i	$c(a_i)$	l_i
a	1/2	0	1
b	1/4	1	1
c	1/8	00	2
d	1/8	11	2

- $\sum_{i=1}^{I} 2^{-l_i} = \frac{1}{2} + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} = 1.5$
- This code is not uniquely decodable

Source coding theorem for symbol codes

- For a RV X, there exists a prefix code C whose expected length satisfies

$$H(x) \leq L(X, C) < H(x) + 1$$

- The expected length is bounded above by the entropy

Construction of an optimal code

- So far, we have proved the existence of “good” prefix codes
- We can assess if a code is uniquely decodable
- How can we construct an optimal code?
  - Huffman code