CS258: Information Theory

Fan Cheng

Spring, 2018. chengfan@sjtu.edu.cn

Recap: Lossy Source Coding

- Shannon source coding theorem
- Probability inequalities
- Typicality: Application of Law of large numbers

- Introduction
- Kraft inequality
- Optimality of symbol codes
- Huffman Coding

Symbol codes

A (binary) symbol code $\mathcal{C}$ for an ensemble X is a mapping from the range of x, $A_{\mathcal{X}} = \{a_1, ..., a_I\}$, to $\{0, 1\}^+$. $c(x)$ will denote the codeword corresponding to x, and $l(x)$ will denote its length, with $l_i = l(a_i)$.

Symbol codes

A (binary) symbol code $\mathcal{C}$ for an ensemble X is a mapping from the range of x, $A_\mathcal{X} = \{a_1, ..., a_I\}$, to $\{0,1\}^+$. $c(x)$ will denote the codeword corresponding to x, and $l(x)$ will denote its length, with $l_i = l(a_i)$.

$$A_\mathcal{X} = \{a, b, c, d\}, P_\mathcal{X} = \{1/2, 1/4, 1/8, 1/8\}$$

a_i	$c(a_i)$	l_i
a	1000	4
b	0100	4
c	0010	4
d	0001	4

Symbol codes

A (binary) symbol code $\mathcal{C}$ for an ensemble X is a mapping from the range of x, $A_{\mathcal{X}} = \{a_1, ..., a_I\}$, to $\{0,1\}^+$. $c(x)$ will denote the codeword corresponding to x, and $l(x)$ will denote its length, with $l_i = l(a_i)$.

$$A_{\mathcal{X}} = \{a, b, c, d\}, P_{\mathcal{X}} = \{1/2, 1/4, 1/8, 1/8\}$$

a_i	$c(a_i)$	l_i
a	1000	4
b	0100	4
c	0010	4
d	0001	4

The extended code C^+ is a mapping from $\mathcal{A}_{\mathcal{X}}^+$ to $\{0,1\}^+$ obtained by concatenation, without punctuation, of the corresponding codewords:

$$c^+(x_1, x_2, ..., x_N) = c(x_1)c(x_2)...c(x_N)$$

$$c^+(acdbac) = 100000100001010010000010$$

Unique decoding

The decoding result should be unique

Unique decoding

The decoding result should be unique

Prefix code

A symbol code is called a prefix code if no codeword is a prefix of any other codeword.

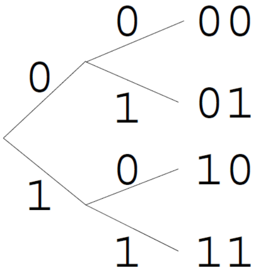
$\{0, 101\}$, $\{1, 110\}$

The decoding result should be unique

Prefix code

A symbol code is called a prefix code if no codeword is a prefix of any other codeword.

$\{0, 101\}$, $\{1, 110\}$



Unique decodeability

Question: Given a list of positive integers $\{l_i\}$, does there exist a uniquely decodeable code with those integers as its codeword lengths?

Unique decodeability

Question: Given a list of positive integers $\{l_i\}$, does there exist a uniquely decodeable code with those integers as its codeword lengths?

Kraft inequality

For any uniquely decodeable code $C(X)$ over the binary alphabet $\{0, 1\}$, the codeword lengths must satisfy:

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

where $I = |\mathcal{A}_{\mathcal{X}}|$

Unique decodeability

Question: Given a list of positive integers $\{l_i\}$, does there exist a uniquely decodeable code with those integers as its codeword lengths?

Kraft inequality

For any uniquely decodeable code $C(X)$ over the binary alphabet $\{0, 1\}$, the codeword lengths must satisfy:

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

where $I = |\mathcal{A}_{\mathcal{X}}|$

Completeness. If a uniquely decodeable code satisfies the Kraft inequality with equality then it is called a complete code.

Unique decodeability

Question: Given a list of positive integers $\{l_i\}$, does there exist a uniquely decodeable code with those integers as its codeword lengths?

Kraft inequality

For any uniquely decodeable code $C(X)$ over the binary alphabet $\{0, 1\}$, the codeword lengths must satisfy:

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

where $I = |\mathcal{A}_{\mathcal{X}}|$

Completeness. If a uniquely decodeable code satisfies the Kraft inequality with equality then it is called a complete code.

Kraft inequality and prefix codes. Given a set of codeword lengths that satisfy the Kraft inequality, there exists a uniquely decodeable prefix code with these codeword lengths.

Expected Length

The expected length $L(C, X)$ of a symbol code C for ensemble X is

$$L(C, X) = \sum_{x \in \mathcal{A}_\mathcal{X}} P(x) l(x)$$

We may also write this quantity as

$$L(C, X) = \sum_{x \in \mathcal{A}_\mathcal{X}} p_i l_i$$

Expected Length

The expected length $L(C, X)$ of a symbol code C for ensemble X is

$$L(C, X) = \sum_{x \in \mathcal{A}_{\mathcal{X}}} P(x) l(x)$$

We may also write this quantity as

$$L(C, X) = \sum_{x \in \mathcal{A}_{\mathcal{X}}} p_i l_i$$

Source coding theorem for symbol codes

For an ensemble X there exists a prefix code C with expected length satisfying

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

Expected Length

The expected length $L(C, X)$ of a symbol code C for ensemble X is

$$L(C, X) = \sum_{x \in \mathcal{A}_\mathcal{X}} P(x) l(x)$$

We may also write this quantity as

$$L(C, X) = \sum_{x \in \mathcal{A}_\mathcal{X}} p_i l_i$$

Source coding theorem for symbol codes

For an ensemble X there exists a prefix code C with expected length satisfying

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

$l_i = \log_2(1/p_i)$

Huffman coding

Reading and Exercise

Reading: Ch. 5 (David MacKay)

Exercise

5.14, 5.19, 5.20, 5.21, 5.27