Information Theory

Mohamed Hamada

Software Engineering Lab
The University of Aizu

Email: hamada@u-aizu.ac.jp
URL: http://www.u-aizu.ac.jp/~hamada

Today's Topics

- Watching a Coding Video (50 mins.)
- What is Information Theory
- Information Source
- Introduction to Source Coding
- What is Information Theory

Information Theory

- What is Information Theory?

- What is the purpose of information theory?

- Why we need to study information theory?

What is information theory (IT) ?

IT is the science that deals with the concept " Information" :
Its measurement & its application

What is the purpose of IT?

Transmission of information in an efficient way : minimum time & space

Why we need to study information theory?

Because of the revolution of communication dealing efficiently with information and its transmission becomes a necessary requirement for a computer engineer.

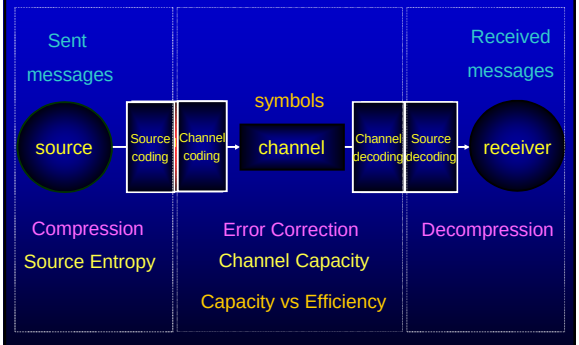
What is Information?

There are 3 types of Information

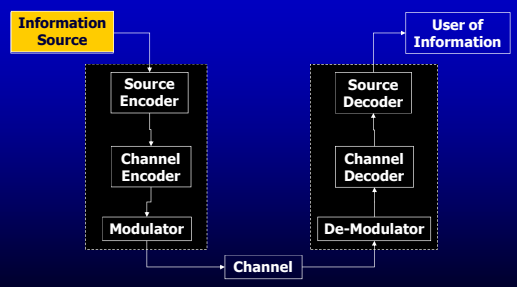
Syntactic Information	*Semantic information*	*Pragmatic information*
Related to the structure of the messages (characters) that forms the information	Related to the meaning of the messages	Related to the usage and effect of the messages

Example:
i. I eat sushi by hashi
ii. By hashi I eat sushi
iii. There is a typhoon in Japan
iv. There is a typhoon in Kanto area in Japan
i. and ii. are syntactically different but semantically & Pragmatically equal
iii. and iv. are syntactically, semantically & pragmatically different
(iv. gives more information than iii.)

INFORMATION TRANSFER ACROSS CHANNELS



Digital Communication Systems



Information Source

Examples:

1. English text
2. A man speaking
3. Photographs
4. Motion of films,
   ..... etc.

A chief aim of information theory is to study how such sequence of symbols (signals) can be most effectively encoded for transmission (by electrical means).

For Information source we have:

· Information expressed as a finite set of symbols : Source Alphabet

· A (discrete) information source : is a source that generates a sequence of symbols

· The symbols are denoted by a1, a2, …., am and the alphabet by A = { a1, a2, …, am }

· A finite sequence of symbols is called *word*. The set of all words is denoted by A*

Information Source

Memoryless

Memoryless means the generated symbols (of a source message) are independent.

Stationary

The idea of stationary of a source demands no change with time

Discrete

The source produces independent symbols in different unit times

Memoryless

Memoryless means the generated symbols (of a source message) are independent.

P(X = i | X1, X2, …, Xn) = P (X = i)

i.e. The probability of the output X is conditionally independent of previous channel inputs or outputs X1, .., Xn

Example:

Coin toss 8 times : the probability to get head
Each time is ½ x …x ½ = (½)^8 = 1/256 = 0.0039 = 0.4 %
Coin toss number 9 : The probability to get head is still ½

So it is independent of the previous 8 toss

Stationary

A process is called ***Stochastic*** if its output is associated with a probability distribution.

A stochastic process is said to be ***stationary*** when the process is (temporally homogeneous) remain invariant under every translation of the time scale

Example 1 : Assume a source produces an infinite sequence of the form:

i.e. What comes later is like what has gone before. Stationary is a designation of such source of characters

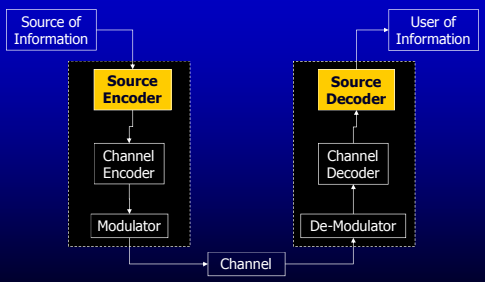
Example 2 : Assume the source that produces

such source is not stationary

The idea of stationary of a source demands no change with time

i.e P(X = i) = C (constant)

Digital Communication Systems



Source Coding

· We consider memoryless, stationary, discrete information source S ,

$$S = \begin{pmatrix} a_1\ a_2\ \ldots\ a_m \\ p_1\ p_2\ \ldots\ p_m \end{pmatrix}$$ where $p_i = p(X_n = a_i)$

so that

$$P(X_n = a_i) = C \text{ for a constant } C$$

i.e. probability doesn't depend on the trial (n)

Source Coding

-**Information source *encoding (enciphering)* :**

is a procedure for associating words constructed from a finite alphabet of a language with given words of another language in a one-to-one manner.

i.e. encoding is a procedure for mapping a given set of messages {m1, m2, ... , mi } onto a new set of encoded messages {c1, c2, ... , ci } in a one-to-one manner.

-**A measure of the encoded information is *entropy* which is the shortest mean encoding.**
The unit measure of entropy is *bit* : *bi*nary digi*t*.

Example: **The entropy of a fair coin toss is 1 bit**

Later on we will study entropy in more details.

Source Coding

-The goal of source coding is to make encoding as short as possible

- ***Code*** **:** For the alphabet A = { a1,…., am }
a code C is a nonempty set of words, i.e. C is a subset of A*

-***Code word*** : Is an element of the code

- ***Code word length*** **:** Is the number of symbols is the code word.
For a code word x, l(x) denote its length

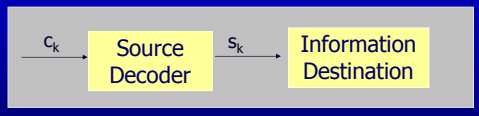
Source Coding

Example : for the source information { Red, Blue} and the code alphabet is {0,1 }.
Let C (Red) = 00 and C (Blue) = 11
Then C = { 00, 11 } is a subset of { 0, 1 }* and the length is l(00) = 2 and l(11) = 2

- ***Binary Source Coding*** **:** Is the code whose alphabet is { 0, 1 }

Decoding

Is the inverse process of encoding.
i.e. The operation of assigning code words to the corresponding words of the source information.



Decoding

Is the inverse process of encoding.
i.e. The operation of assigning code words to the corresponding words of the source information.

Various Coding Techniques :

- **Comma Code :** is a code C with the code words
1, 01, 001, 0001, ….., 0…01, 000…0
where the last two codes have n-1 and n 0's respectively.

-**Variable length Code** : is the code with codeword of different lengths.

- **Instantaneous code** : is the code in which each codeword in any string of codewords can be decoded (reading from left to right) as soon as it is received.

- **Singular Code :** A code is singular if some codewords are not different

Examples of Information Source Coding

Symbol	Pro	C1	C2	C3	C4	C5	C6
A	0.4	00	1	1	1	0	0
B	0.3	01	01	01	00	1	0
C	0.2	10	001	001	010	10	10
D	0.1	11	0001	000	011	01	01
MCL		2	2	1.9	1.9	1.3	1.3

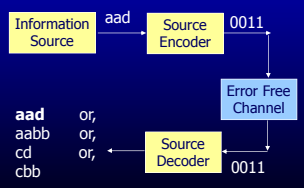
MCL = Mean code length

C1 : fragile (recovery impossible if letters drop out)
C2 : robust (recovery possible if letters drop out);
variable-length code (non-equal length code);
C3 : smarter comma code (incorporating maximum length)
C3, C4 : also (variable- length) instantaneous codes;
C5 : decoding impossible;
D A A C
0 1 0 ➔ 01 0 or 0 10
C6 : singular code (so decoding impossible)

Coding/Decoding Example

Bad Code

One input can be interpreted to many different outputs.



Example: Code I

Source Symbol	Binary Code Representation
a	0
b	1
c	00
d	11