DCSP-6: Information Theory

Jianfeng Feng

Department of Computer Science Warwick Univ., UK

Jianfeng.feng@warwick.ac.uk

http://www.dcs.warwick.ac.uk/~feng/dcsp.html
Assignment: 2016

- Q1: you should be able to do it after last week seminar
- Q2: need a bit reading (my lecture notes)
- Q3: standard
- Q4: standard
Assignment: 2016

• Q5: standard

• Q6: standard

• Q7: after today’s lecture

• Q8: load jazz, plot soundscape
  load TuneJazz
  load NoiseJazz
  plot
Correlation or covariance

- \( \text{Cov}(X,Y) = E(X-EX)(Y-EY) \)

- correlation coefficient is normalized covariance
  \( \text{Coef}(X,Y) = E(X-EX)(Y-EY) / [\sigma(X)\sigma(Y)] \)

- Positive correlation, Negative correlation

- No correlation (independent)
Stochastic process = signal

• A stochastic process is a collection of random variables $x[n]$, for each fixed $[n]$, it is a random variable

• Signal is a typical stochastic process

• To understand how $x[n]$ evolves with $n$, we will look at auto-correlation function (ACF)

$$g(k) = E(x[n+k] - Ex[n+k])(x[n] - Ex[n])$$

• ACF is the correlation between $k$ steps
Stochastic process

\[ (k) = \frac{E(x[n+k] \cdot Ex[n+k]) \cdot (x[n] \cdot Ex[n])}{(x[n+k]) \cdot (x[n])} \]

• two signals are generated: y (red) is simply randn(1,200)
  x (blue) is generated \( x[i+10] = 0.8x[i] + y[i+10] \)

• For y, we have \( \gamma(0)=1, \gamma(n)=0, \) if \( n \) is not 0 : having no memory
• For x, we have \( \gamma(0)=1, \) and \( \gamma(n) \) is not zero, for some \( n \): having memory
White noise $w[n]$

- **White noise** is a random process we can not predict at all (independent of history)

$$\gamma(k) = \frac{E(x[n + k] - Ex[n + k])(x[n] - Ex[n])}{\sigma(x[n + k])\sigma(x[n])} = 0, \text{ if } k \neq 0$$

- In other words, it is the most ‘violent’ noise

- White noise draws its name from white light which will become clear in the next few lectures
The most ‘noisy’ noise is a white noise since its autocorrelation is zero, i.e.
\[
corr(w[n], w[m]) = 0 \text{ when } n \neq m
\]

Otherwise, we called it colour noise since we can predict some outcome of \( w[n] \), given \( w[m] \), \( m < n \)
Sweety Gaussian
A linear combination of two Gaussian random variables is Gaussian again.

For example, given two independent Gaussian variable X and Y with mean zero:

- \( aX + bY \) is a Gaussian variable with mean zero and variance \( a^2 \sigma(X) + b^2 \sigma(Y) \).

This is very rare (the only one in continuous distribution) but extremely useful: panda in the family of all distributions.
Data Transmission

source → modulator → transmitter → transmission channel → receiver → demodulator → sink

(noise) → noise, distortion, attenuation → (noise)
Data Transmission

How to deal with noise?
How to transmit signals?
Data Transmission

Clock

Data

Manchester
Data Transmission

- Fourier Transform
- ASK (AM), FSK(FM), and PSK (skipped, but common knowledge)
- Noise
- Signal Transmission
Today

- Data transmission:
- Shannon Information and Coding:
Information and coding theory

Information theory is concerned with

- description of *information sources*
- representation of the information from a source (coding),
- transmission of this information over channel.
The best example
how a deep mathematical theory
could be successfully applied to
solving engineering problems.
Information and coding theory

Information theory is a discipline in applied mathematics involving the quantification of data with the goal of enabling as much data as possible to be reliably stored on a medium and/or communicated over a channel.
Information and coding theory

The measure of data, known as

information entropy,

is usually expressed by the average number of bits needed for storage or communication.
The field is at the crossroads of

- mathematics,
- statistics,
- computer science,
- physics,
- neurobiology,
- electrical engineering.
Information and coding theory

Impact has been crucial to success of

• voyager missions to deep space,

• invention of the CD,

• feasibility of mobile phones,

• development of the Internet,

• the study of linguistics and of human perception,

• understanding of black holes,

and numerous other fields.
Founded in 1948 by Claude Shannon in his seminal work

A Mathematical Theory of Communication
A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist and Hartley on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which actually will be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:
1. Shannon's source coding theorem

the number of bits needed to represent the result of an uncertain event is given by its entropy;

2. Shannon's noisy-channel coding theorem

reliable communication is possible over noisy channels if the rate of communication is below a certain threshold called the channel capacity.

The channel capacity can be approached by using appropriate encoding and decoding systems.
Information and coding theory

1. Shannon's source coding theorem
   
   the number of bits needed to represent the result of an uncertain event is given by its *entropy*;

2. Shannon's noisy-channel coding theorem

   The channel capacity can be approached by using appropriate encoding and decoding systems.
Information and coding theory

Consider to predict the activity of Prime minister tomorrow. This prediction is an information source.
Consider to predict the activity of Prime Minister tomorrow.

This prediction is an information source $X$.

The information source $X = \{O, R\}$ has two outcomes:
- He will be in his office (O),
- he will be naked and run 10 miles in London (R).
Clearly, the outcome of 'in office' contains little information; it is a highly probable outcome.

The outcome 'naked run', however contains considerable information; it is a highly improbable event.
An information source is a probability distribution, i.e. a set of probabilities assigned to a set of outcomes (events).

This reflects the fact that the information contained in an outcome is determined not only by the outcome, but by how uncertain it is.

An almost certain outcome contains little information.

A measure of the information contained in an outcome was introduced by Hartley in 1927.
Defined the information contained in an outcome \( x_i \) in \( x = \{x_1, x_2, \ldots, x_n\} \)

\[
l(x_i) = - \log_2 p(x_i)
\]
Information

The definition above also satisfies the requirement that the total information in independent events should add.

Clearly, our prime minister prediction for two days contain twice as much information as for one day.
The definition above also satisfies the requirement that the total information in independent events should add.

Clearly, our prime minister prediction for two days contain twice as much information as for one day $X=\{OO, OR, RO, RR\}$.

For two independent outcomes $x_i$ and $x_j$,

$$I(x_i \text{ and } x_j) = - \log_2 P(x_i \text{ and } x_j)$$

$$= - \log_2 P(x_i) P(x_j)$$

$$= - \log_2 P(x_i) - \log_2 P(x_j)$$
The measure entropy $H(X)$ defines the information content of the source $X$ as a whole. It is the mean information provided by the source.

We have

$$H(X) = \sum_i P(x_i) \log_2 P(x_i) = - \sum_i P(x_i) \log_2 P(x_i)$$

A binary symmetric source (BSS) is a source with two outputs whose probabilities are $p$ and $1-p$ respectively.
The prime minister discussed is a BSS.

The entropy of the BBS source is

$$H(X) = -p \log_2 p - (1-p) \log_2 (1-p)$$
Entropy

When one outcome is certain, so is the other, and the entropy is zero.

As $p$ increases, so too does the entropy, until it reaches a maximum when $p = 1 - p = 0.5$.

When $p$ is greater than 0.5, the curve declines symmetrically to zero, reached when $p=1$. 
Tomorrow

- Application of Entropy in coding
- Minimal length coding
We conclude that the average information in BSS is maximised when both outcomes are equally likely.

Entropy is measuring the average uncertainty of the source.

(The term entropy is borrowed from thermodynamics. There too it is a measure of the uncertainty of disorder of a system).

Shannon:
- My greatest concern was what to call it.
- I thought of calling it ‘information’, but the word was overly used, so I decided to call it ‘uncertainty’.
- When I discussed it with John von Neumann, he had a better idea.
- Von Neumann told me, ‘You should call it entropy, for two reasons.
  - In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name.
  - In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage.
Entropy

In Physics: thermodynamics

The arrow of time (Wiki)

- Entropy is the only quantity in the physical sciences that seems to imply a particular direction of progress, sometimes called an arrow of time.

- As time progresses, the second law of thermodynamics states that the entropy of an isolated systems never decreases

- Hence, from this perspective, entropy measurement is thought of as a kind of clock
Figure 3: Functional Entropy vs. Age.

From
The Increase of the Functional Entropy of the Human Brain with Age
Y. Yao, W. L. Lu, B. Xu, C. B. Li, C. P. Lin, D. Waxman & J. F. Feng
Scientific Reports 3, Article number: 2853 | doi:10.1038/srep02853
Received 30 May 2013 | Accepted 10 September 2013 | Published 09 October 2013

(a) Combined

Entropy = 0.0013 × Age + 3.5444

(b) Age=50

Males
Females

(c) Males

Entropy = 0.0015 × Age + 3.5336

(d) Females

Entropy = 0.0011 × Age + 3.5547
DCSP-7: Minimal Length Coding

Jianfeng Feng
Department of Computer Science Warwick Univ., UK

Jianfeng.feng@warwick.ac.uk

http://www.dcs.warwick.ac.uk/~feng/dcsp.html
Automatic Image Caption (better than human)
This Week’s Summary

- Information theory
- Huffman coding: code events as economic as possible
Information sources

- \( X = \{x_1, x_2, \ldots, x_N\} \) with a known probability
- \( P(x_i) = p_i, \quad i=1,2,\ldots,N \)

Example 1:

- \( X = (x_1, x_2, x_3) \)
  - \( x_1 = \) lie on bed at 12 noon today
  - \( x_2 = \) in university at 12 noon today
  - \( x_3 = \) attend a lecture at 12 noon today
- \( X = (B, U, L) \)
- \( p = (1/2, 1/4, 1/4) \)

- \( H(X) = .5*1 + 2*1/4 + 2*1/4 = 1.5 \) (Entropy)
- \( B=0, \quad U=1, \quad L=01 \) (coding)
- \( L_s = 0.5*1 + 0.25*1 + 0.25*2 = 1.25 \) (average coding length)
Information sources

• Example 2.  Left: information source \( p(x_i), \ i = 1, \ldots, 27 \)
  right: codes

To be, or not to be, that is the question— Whether 'tis Nobler in the mind to suffer The Slings and Arrows of outrageous Fortune, Or to take Arms against a Sea of troubles, And by opposing end them? To die, to sleep—

As short as possible

01110 0000111111
1111100000000000
111100000011000
100010000010000
1011111111100000
Information source coding

• Replacement of the symbols (naked run/office in PM example) with a binary representation is termed *source coding*.  

• In any coding operation we replace the symbol with a codeword.

• The purpose of source coding is to reduce the number of bits required to convey the information provided by the information source: *minimize the average length of codes.*

• Conjecture: an information source of entropy $H$ needs on average only $H$ binary bits to represent each symbol.
An *instantaneous* code can be found that encodes a source of entropy $H(X)$ with an average number $L_s$ (average length) such that

$$L_s \geq H(X)$$
How does it work?

- Like many theorems of information theory, the theorem tells us nothing of how to find the code.
- However, it is useful results.
- Let us have a look how it works
Example

- Look at the activities of PM in three days with $P(O) = 0.9$
- Calculate probability
- Assign binary codewords to these grouped outcomes.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>OOO</th>
<th>OON</th>
<th>ONO</th>
<th>NOO</th>
<th>NNO</th>
<th>NON</th>
<th>ONN</th>
<th>NNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.729</td>
<td>0.081</td>
<td>0.081</td>
<td>0.081</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.001</td>
</tr>
<tr>
<td>Codeword</td>
<td>0</td>
<td>1</td>
<td>01</td>
<td>10</td>
<td>11</td>
<td>00</td>
<td>000</td>
<td>111</td>
</tr>
<tr>
<td>Codeword Length (in bits)</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1: Variable length source coding
Example

• Table 1 shows such a code, and the probability of each code word occurring.

• Entropy is

\[ H(X) = -0.729 \log_2(0.729) - 0.081 \log_2(0.081) \times 3 \]
\[ -0.009 \log_2(0.009) \times 3 - 0.001 \log_2(0.001) \]
\[ = 1.4070 \]

• The average length of coding is given by

\[ L_s = 0.729\times1 + 0.081\times1 + 2\times0.081\times2 + 2\times0.009\times2 \]
\[ + 3\times0.009 + 3\times0.001 \]
\[ = 1.2 \]
Moreover, without difficulty, we have found a code that has an average bit usage less than the source entropy.
Example

• However, there is a difficulty with the code in Table 1.

• Before a code word can be decoded, it must be parsed.

• Parsing describes that activity of breaking the message string into its component codewords.
Example

- After parsing, each codeword can be decoded into its symbol sequence.

- An instantaneously parsable code is one that can be parsed as soon as the last bit of a codeword is received.
Instantaneous code

- An instantaneous code must satisfy the prefix condition: that no codeword may be a prefix of any other code.

- For example: in the codeword, we should not use 1 11 to code two events. When we receive 11, it could be ambiguous.

- This condition is not satisfied by the code in Table 1.
Huffman coding

• The code in Table 2, however, is an instantaneously parsable code.

• It satisfies the prefix condition.
### Huffman coding

<table>
<thead>
<tr>
<th>Sequence</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.729</td>
<td>0.081</td>
<td>0.081</td>
<td>0.081</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.001</td>
</tr>
<tr>
<td>Codeword</td>
<td>1</td>
<td>011</td>
<td>010</td>
<td>001</td>
<td>00011</td>
<td>00010</td>
<td>00001</td>
<td>00000</td>
</tr>
<tr>
<td>Codeword Length (in bits)</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

### Weighted code length

\[ \sum_{i} p_i \times l_i = 0.729 \times 1 + 0.081 \times 3 \times 3 + 0.009 \times 5 \times 3 + 0.001 \times 5 \]

\[ = 1.5980 \quad \text{(remember entropy is 1.4)} \]

---

**Table 2:** OOO=A, OON=B, ONO=C, NOO=D, NNO=E, NON=F, ONN=G, NNN=H
Huffman coding

Decoding

1 1 1 0 1 1 0 1 0 1 1 0 0 0 0 0 0 0 0 0 1
Huffman coding

• The derivation of the Huffman code tree is shown in the following Figure and the tree itself is shown in the next Figure.

• In both these figures, the letter A to H have be used in replace of the sequence in Table 2 to make them easier to read.
Creating the tree:

1. Start with as many leaves as there are symbols.
2. Queue all leaf nodes into the first queue (in order).
3. While there is more than one node in the queues:
   - Remove two nodes with the lowest weight from the queues.
   - Create a new internal node, with the two just-removed nodes as children (either node can be either child) and the sum of their weights as the new weight.
   - Update the parent links in the two just-removed nodes to point to the just-created parent node.
4. Queue the new node into the second queue.
5. The remaining node is the root node; the tree has now been generated.
Huffman coding

- Prefix condition is obviously satisfied since in the tree above, each branch codes one alphabetic.
Huffman coding

- For example, the code in Table 2 uses 1.6 bits/symbol which is only 0.2 bits/symbol more bits per sequence than the theorem tells us is the best we can do.

- We might conclude that there is little point in expending the effort in finding a code less satisfying the inequality above.
Another thought

• How much have we saved in comparison with the most naïve idea?

• i.e. O=1, N=0

• $L_s = 3 \left[ P(\text{OOO}) + \ldots + P(\text{NNN}) \right] = 3$, halving it
My most favourite story (History)

• In 1951, David A Huffman and his MIT information theory classmates were given the choice of a term paper or a final exam.
• The Professor, Robert M Fano, assigned a term paper on the problem of finding the most efficient binary code.
• Huffman, unable to prove any codes were the most efficient, was about to give up when he hit upon the idea of using a frequency-sorted binary tree and quickly proved this method the most efficient.
• In doing so, the student outdid his professor, who had worked with information theory inventor Clude Shannon to develop an optimal code.
• By building the tree from the bottom up instead of the top down, Huffman avoided the major flaw of the suboptimal Shannon-Fano coding.
Coding English: Huffman Coding

Frequency for alphabetics

![Bar chart showing frequency distribution of English alphabets. The chart plots the frequency of each letter, with 'e' having the highest frequency and 'z' having the lowest.]}
Turbo coding

• Using Bayesian theorem to code and decode

• Bayesian theorem basically said we should employ priori knowledge as much as possible

• Read yourself
The most famous of all results of information theory is Shannon's channel capacity theorem.

For a given channel there exists a code that will permit the error-free transmission across the channel at a rate $R$, provided $R < C$, the channel capacity.

\[ C = B \log_2 \left( 1 + \left( \frac{S}{N} \right) \right) \, b/s \]
Channel Capacity

As we have already noted, the astonishing part of the theory is the existence of a channel capacity.

- Shannon's theorem is both tantalizing and frustrating.
Channel Capacity

- It is offers error-free transmission, but it makes no statements as to what code is required.

- In fact, all we may deduce from the proof of the theorem is that is must be a long one.

- No none has yet found a code that permits the use of a channel at its capacity.

- However, Shannon has thrown down the gauntlet, in as much as he has proved that the code

- We will not go into details here
Next week

• We had a de tour: information theory

• Next week, come back to FT

• Pressing a button will do FFT for you