Richard W. Hamming

Learning to Learn

The Art of Doing Science and Engineering

Session 13: Information Theory

Information

Shannon identifies information with surprise

For example: Telling someone it is smoggy in Los Angeles is not much of a surprise, and therefore not much new information

Surprise defined

p is the probability of the event

I (*p*) is information gained from that event

$$I(p) = -\log_2 p = \log_2\left(\frac{1}{p}\right)$$

Information learned from independent events is additive

$$I(p_1 p_2) = I(p_1) + I(p_2)$$

Definition Confounding

Information Theory has not "defined" information

It actually measures "surprise"

Shannon's definition may suffice for machines, but it does not represent what we normally think of as information

Should have been called "Communication Theory" and not "Information Theory"

Definition Confounding

Realize how much the definition distorts the common view of information

Illustrates a point to examine whenever new definitions presented

- How far does the proposed definition agree with the original concepts you had, and how far does it differ?

Information Entropy: *H*(*P*)

The average amount of information in the system

$$H(P) = \sum_{i=1}^{q} p_i I(p_i) = \sum_{i=1}^{q} p_i \log\left(\frac{1}{p_i}\right)$$

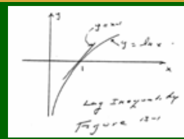
Not the same as physical entropy even though mathematical form is similar

Gibbs Inequality
(mathematical interlude)

$$\sum p_i \log\left(\frac{q_i}{p_i}\right) \le 0$$

Here q & p are independent probability distributions

Note the form is nearly identical to *H*(*P*)



From fig 13.1: $\log x \le x - 1 \quad (0 \le x < \infty)$

$$\sum p_i \left\{\frac{q_i}{p_i} - 1\right\} = \sum q_i - \sum p_i = 1 - 1 = 0$$

so $H(P) = \log q$ where $q_i = 1/q$ and $q = 1-p$

Kraft Inequality: *K*
(mathematical interlude continues)

Given a uniquely decodable code $K = \sum \frac{1}{2^{l_i}} \le 1$

Where l_i is length of code segment *i*

Define the pseudoprobabilities $Q_i = \frac{2^{-l_i}}{K}$ where $\sum [Q_i] = 1$

It then follows from Gibbs $\sum_{i=1}^{q} p_i \log\left(\frac{1}{K p_i 2^{l_i}}\right) \le 0$

Substituting Q_i for q_i

And finally the "*Noiseless Coding Theorem of Shannon*"

$$H(P) \le \log K + \sum p_i l_i \le L = \text{ average code length}$$

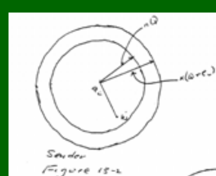
Channel Capacity defined

The maximum amount of information that can be sent through the channel reliably

With *n* bits sent, expect *nQ* errors

Given a large enough *n*

- you can force the probability of falling outside the *nQ* boundary as small as you please



Channel Capacity

Based upon random encoding with no error correction

With *M* messages of length *n* there are 2^{Mn} code books

- Leaves the possibility for destructive overlap

Proved the possibility of overlap is very small by averaging over all 2^{Mn} code books for the average error

Using sufficiently large *n*'s will reduce the probability of error while simultaneously maximizing the flow of information through the channel

Thus if the average error is suitably small, then at least one code will be suitable: "Shannon's noisy coding theorem"

What does it mean?

"Sufficiently large *n*" necessary to ensure information flow is approaching channel capacity may be so large as to be too slow

Error-correcting codes avoid the slowness at the cost of some channel capacity

- Use computable functions, rather than lengthy random code books
- When many errors are corrected, the performance compared to channel capacity is quite good

In Practice

If your system provides error correction, use it!

Solar-system exploration satellites

- Extreme total-power limitations of system about 5W, so restricted transmission power and distance/background noise induce errors.
- Aggressive error-correcting codes enabled more effective use of available bandwidth as errors were self correcting

Hamming codes may not guarantee use near optimal channel capacity, but does guarantee error-correction reliability to a specified level

Shannon coding only states a "probably low error" given a long enough series of bits, but will push those bits near channel capacity

In Practice

Information theory does not *tell* you how to design, but gives point the way towards efficient designs

Remember, information theory applies to data communications and is not necessarily relevant to *human communication*

Final Points

Reuse of established terms as definitions in a new area *should* fit our previous beliefs, but often do not and have some degree of distortion and non-applicability to the way we thought things were.

Definitions don't actually define things, just suggest how those things should be handled.

Final Points

All definitions should be inspected, not only when proposed but later when they apply to the conclusions drawn.

- Were they framed to get the desired result?
- Are they applicable under differing conditions?

Beware: initial definitions often determine what you find or see, rather than describe what is actually there.

- Are they creating results which are circular tautologies vice actual results?