Information Theory

Jethro Kuan

January 7, 2019

Contents

1 Introduction

1.1 How can we achieve perfect communication over an imperfect noisy communication channel?

The physical solution is to improve the characteristics of the communication channel to reduce its error probability. For example, we can use more reliable components in the communication device's circuitry.

Information theory and coding theory offer an alternative: we accept the given noisy channel as it is, and add communication systems to it to detect and correct errors introduced by the channel.

An encoder encodes the source message `s` into a transmitted message `t`, adding *redundancy* to the original message in some way. The channel adds noise to the transmitted message, yielding a received message `r`. The decoder uses the known redundancy introduced by the encoding system to infer both the original signal `s` and the added noise.

Information theory is concerned of the theoretical limitations and potentials of these systems. Coding theory is concerned with the creation of practical encoding and decoding systems.

1.2 Error-correcting codes for binary symmetric channels

1.2.1 Repetition codes

Key idea: repeat every bit of the message a prearranged number of times, and pick the bit with the majority vote.

We can describe the channel as adding a sparse noise vector n to the transmitted vector, adding in a modulo 2 arithmetic.

One can show that this algorithm is optimal by considering the maximum likelihood function of `s`.

The repetition code R_3 (repeat 3 times) has reduced the probability of error, but has also reduced the *rate* of information by a factor of 3.

1.2.2 Block codes - the (7,4) Hamming Code

Key idea: add redundancy to blocks of data instead of encoding one bit at a time.

A block code is a rule for converting a sequence of source bits `s`, of length `K`, into transmitted sequence `t` of length `N > K` bits. The Hamming code transmits `N=7` bits for every `K=4` bits.

Because the Hamming code is a linear code, it can be written compactly as a matrix:

$$\text{transmitted} = G^T \text{source}$$

where G is the generator matrix of the code.

1. Decoding for linear codes: syndrome decoding The decoding problem for linear codes can also be described in terms of matrices. We evaluate 3 parity-check bits for the received bits $r_1 r_2 r_3 r_4$, and see whether they match the three received bits $r_5 r_6 r_7$. The differences (mod 2) between these 2 triplets are called the *syndrome* of the received vector. If the syndrome is 0, then the received vector is a code word, and the most probable decoding is given by reading its first four bits.

$$G^T = \begin{bmatrix} I_4 \\ P \end{bmatrix},$$

$$H = \begin{bmatrix} -P & I_3 \end{bmatrix} = \begin{bmatrix} P & I_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

1.2.3 What performance can the best codes achieve?

We consider the `(R, p_b)` plane, where R is the rate, and p_b is the decoded bit-error probability, Claude Shannon proved that the boundary between achievable and non-achievable points meets the R axis at a non-zero value $R = C$. For any channel, there exist codes that make it possible to communicate with arbitrarily small probability of error $\$p_b$= at non-zero rates. This theorem is called the *noisy-channel coding theorem.*

The maximum rate at which communication is possible with arbitrarily small p_b is called the capacity of the channel.

$$C(f) = 1 - H_2(f) = 1 - \left[f \log_2 \frac{1}{f} + (1-f) \log_2 \frac{1}{1-f} \right]$$

2 Measuring Information Content

We view information content as the "degree of surprise" on learning the value of x, for some random variable x. This content will therefore depend on $p(x)$, and we're looking for a monotonic function $h(x)$ that expresses information content.

We would also like some desirable properties from our function $h(x)$: $h(x, y) = h(x) + h(y)$ if random variables x and y are statistically independent, since the information gained from the realization of both random variables must be additive. Since $p(x, y) = p(x)p(y)$, it's easy to see that $h(\dot{})$ must be given by the logarithm of $h(x)$. Thus, we have:

$$h(x) = -\log_2 p(x) \tag{1}$$

Then the average information a random variable transmits in the process is obtained by taking the expectation of 1 with respect to the distribution $p(x)$:

$$H[x] = -\sum_x p(x) \log_2 p(x) \tag{2}$$

3 Shannon Information Content

A set of size S can be communicated with $\log S$ bits.

4 Source Coding Theorem

We can compress N outcomes from a source X into roughly $NH(X)$ bits.

This is provable by counting the typical set.