CMPSCI 240: Reasoning about Uncertainty

Lecture 23: Information Theory

Andrew McGregor

University of Massachusetts

Last Compiled: April 27, 2017

Coding, Compression, and Information Theory

- Encoding messages/files as binary strings...
- Information Transmission: How to talk over a garbled phone line.
- Information Compression: How you'd design a language if you like to keep your conversations brief.

Outline

Information Theory

The concept of Information Theory was proposed by Claude Shannon in 1948 in his paper "A Mathematical Theory of Communication".

Goal: Send messages over a noisy channel, and then to have the receiver reconstruct the message with low probability of error, in spite of the channel noise.

Encoding Messages as Fixed Length Binary Strings

- Let C be a set of k messages that need to be sent.
- Consider encoding each message as a different binary string of length n.
- How large must n be such that such an encoding is possible?

$$k \leq 2^n$$

- For example, the 26 letters of the alphabet can be encoded as binary strings of length 5 since $26 \leq 2^5$:

$$a \to 00000\ ,\ b \to 00001\ ,\ c \to 00010\ ,\ d \to 00011\ ,\ \ldots z \to 11001$$

- If you tried to use binary strings of length 4, then at least two letters would have to have to same binary string.

Encoding Messages with Redundancy: Error Detecting

- Sometimes we might want to use more bits so that message is "protected" against errors that might occur in the transmission.
- E.g., suppose you have 8 possible messages corresponding to

$$000, 001, 010, 011, 100, 101, 110, 111$$

- Add one bit, a parity bit to each string such that each string now has an even number of ones.

$$0000, 0011, 0101, 0110, 1001, 1010, 1100, 1111$$

- If an odd numbers of bits get flipped, you will detect that an error has a occurred and won't be misled.
- Define the Hamming distance d between two binary strings to be the number of coordinates in which they differ. After adding parity bit, all strings have Hamming distance at least 2 from each other, e.g.,

$$d(0101, 0110) = 2$$

Change in Error Probability

- Suppose each bit gets flipped with probability $1/3$
- Before adding the parity bit, what's the probability you either get the correct message or detect that there was an error?

$$P(\text{no flips}) = (2/3)^3 = 8/27 = 0.29\ldots$$

- After adding the parity bit, what's the probability you either get the correct message or detect that there was an error?

$$\begin{aligned} P(0,\ 1,\ \text{or } 3 \text{ flips}) &= (2/3)^4 + 4 \cdot (1/3) \cdot (2/3)^3 + 4 \cdot (1/3)^3 \cdot (2/3) \\ &= 56/81 = 0.69\ldots \end{aligned}$$

Experiment Time

- All books have an ISBN (International Standard Book Number) that has either 10 digits or 13 digits.
- If the code has 10 digits $x_1 x_2 \dots x_{10}$ then x_{10} equals

  $$(11-(10x_1+9x_2+8x_3+7x_4+6x_5+5x_6+4x_7+3x_8+2x_9 \bmod 11)) \bmod 11$$

- If the code has 13 digits $x_1 x_2 \dots x_{10} x_{11} x_{12} x_{13}$ then x_{13} equals

  $$(10-(x_1+3x_2+x_3+3x_4+x_5+3x_6+x_7+3x_8+x_9+3x_{10}+x_{11}+3x_{12} \bmod 10)) \bmod 10$$

- If the numbers don't satisfy this condition, you know there's been mistake writing down the number.

Encoding Messages with Redundancy: Error Correcting

- Suppose you now have 16 possible messages corresponding to

$$0000, 0001, 0010, \ldots, 1111$$

- Now consider adding 3 bits $y_1 y_2 y_3$ to each string $x_1 x_2 x_3 x_4$ where

$$\begin{aligned} y_1 &= x_1 + x_2 + x_4 \pmod 2 \\ y_2 &= x_1 + x_3 + x_4 \pmod 2 \\ y_3 &= x_2 + x_3 + x_4 \pmod 2 \end{aligned}$$

- After encoding, all strings are Hamming distance 3 from each other.
- On receiving a string z, decode it as codeword that is closest to z.
- If only one bit is changed, it is still possible to correct the error:
  - Suppose s was the sent codeword but $d(s', z) \leq d(s, z)$ for some other codeword s'
  - Then $d(s', s) \leq d(s', z) + d(z, s) \leq 2d(s, z) = 2$ which is a contradiction since all codewords differ in at least 3 places.

Example

The resulting code has the codewords:

0000000
0001111
0010011
0011100
0100101
0101010
0110110
0111001
1000110
1001001
1010101
1011010
1100011
1101100
1110000
1111111

Change in Error Probability

- Suppose each bit gets flipped with probability $1/4$
- Before adding the extra bits, the probability you work out the correct message:

$$(3/4)^4 = 0.32$$

- After adding the extra bits, the probability you work out the correct message is at least:

$$(3/4)^7 + 7 \cdot (3/4)^6 \cdot 1/4 = 0.44$$