Statistics for Computing

MA4413 Lecture 12A

Kevin O'Brien

kevin.obrien@ul.ie

Dept. of Mathematics & Statistics,
University *of* Limerick

Autumn 2011

Codes

Recall from last weeks lectures, this table below where a source of size 4 has been encoded in binary codes with symbol 0 and 1.

X	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6
x_1	00	00	0	0	0	1
x_2	01	01	1	10	01	01
x_3	00	10	00	110	011	001
x_4	11	11	11	111	0111	0001

Code Classifications

- **Code 1** This code is fixed length, but not distinct. Two symbols have the same binary representation. Due to this flaw it is no longer considered.
- **Code 2** This code is fixed length and distinct.
- **Code 3** This code is not uniquely decodable. Again due to this flaw, we will no longer consider it.

Prefix-free codes

- A prefix-free code is one in which no codeword is a prefix in another.
- Note that every prefix-free code is decipherable, but the converse is not true.
- In code 4, none of the codewords appear as prefixes for other codewords.
- For code 5, each code word are prefixes for the subsequent codeword.
- Both code 4 and 5 are uniquely decodable.
- Code 6 is prefix free and uniquely decodable.

Word Length

- These codes use code lengths between 1 and 4.
- For code 2; $n_1 = n_2 = n_3 = n_4 = 2$.
- For code 6; $n_1 = 1$, $n_2 = 2$, $n_3 = 3$,$n_4 = 4$.

Word Length

- Suppose that the symbols $\{x_1, x_2, x_3, x_4\}$ appear with the following probabilities $\{0.4, 0.3, 0.2, 0.1\}$
- The average code word length E(L) per source symbol is given by

$$E(L) = \sum_{i=1}^{m} P(x_i) n_i$$

- For each code compute $E(L)$.

Word Length

- **Code 1 and 2** Codes are fixed length $E(L) = 2$
- **Code 3** Recall: Code is flawed

$$E(L) = (0.4 \times 1) + (0.3 \times 1) + (0.2 \times 2) + (0.1 \times 2) = 1.3$$

- **Code 4**

$$E(L) = (0.4 \times 1) + (0.3 \times 2) + (0.2 \times 3) + (0.1 \times 3) = 1.9$$

- **Code 5 and 6**

$$E(L) = (0.4 \times 1) + (0.3 \times 2) + (0.2 \times 3) + (0.1 \times 4) = 2$$

Word Length : Interpretation

- Code 4 would require 190 binary digits to transmit 100 symbols. The transmission would be uniquely decodable.
- Code 3 would require 130 binary digits to transmit 100 symbols. The transmission would be not be uniquely decodable, and the intended message would be unclear.
- For the other codes, each would require 200 digits.
- Code 4 is seemingly the best choice.

Entropy and Efficiency

Given that the entropy of the input source is $H(X) = 1.85$b, compute the efficiency η for each code.

- **Code 1 and 2**
  $\eta = H(X)/E(L) = [1.85/2] \times 100\% = 92.5\%$
- **Code 3**
  Recall that this code is flawed
  $\eta = H(X)/E(L) = [1.84/1.3] \times 100\% = 142\%$
- **Code 4**
  $\eta = H(X)/E(L) = [1.85/1.9] \times 100\% = 97.2\%$
- **Code 5 and 6**
  $\eta = H(X)/E(L) = [1.85/2] \times 100\% = 92.5\%$

Instantaneous Codes

- Recall from previous lecture
- A uniquely decodable code is called an instantaneous code if the end of any code word is recognizable without examining subsequent code symbols.
- The instantaneous codes have the property that no code word is a prefix of another code word.
- Codes 2,4 and 6 are prefix-free codes, hence they are instantaneous codes.

Kraft inequality

- ***Kraft's inequality*** gives a necessary and sufficient condition for the existence of a uniquely decodable code for a given set of codeword lengths (more so variable length codes)
- More specifically, Kraft's inequality limits the lengths of codewords in a prefix code, and can be thought of in terms of a constrained budget to be spent on codewords, with shorter codewords being more expensive.

Kraft inequality

- Let X be a DMS with alphabet ($x_i = \{1, 2, \ldots, m\}$).
- Assume that the length of the assigned binary code word corresponding to x_i is n_i.
- Kraft inequality is given as

$$K = \sum_{i=1}^{m} 2^{-n_i} \leq 1$$

Kraft inequality

- If Kraft's inequality holds with strict inequality (i.e. $K < 1$), the code has some redundancy.
- If Kraft's inequality holds with strict equality (i.e. $K = 1$), the code in question is a complete code.
- The closer K is to 1, the more efficient the code is.
- If Kraft's inequality does not hold, the code is not uniquely decodable.

Kraft inequality

- Note that the Kraft inequality assures us of the existence of an instantaneously decodable code with code word lengths that satisfy the inequality.
- However it does not show us how to obtain these code words, nor does it say that any code that satisfies the inequality is automatically uniquely decodable

Kraft inequality

Compute the value for K in each case, and determine whether Kraft's Inequality is observed.

- **code 1 and 2** K = 1 so Kraft's inequality is obeyed.

$$K = \sum_{i=1}^{m} 2^{-n_i} = (2^{-2} \times 4) = 1$$

- **code 3** K = 1.5 so Kraft's inequality is not obeyed.

$$K = \sum_{i=1}^{m} 2^{-n_i} = (2^{-1} + 2^{-1} + 2^{-2} + 2^{-2}) = 1.5$$

- **code 4** K = 1 so Kraft's inequality is obeyed.

$$K = \sum_{i=1}^{m} 2^{-n_i} = (2^{-1} + 2^{-2} + 2^{-3} + 2^{-3}) = 1$$

- **code 5 and 6** K = 0.9375 so Kraft's inequality is obeyed.

$$K = \sum_{i=1}^{m} 2^{-n_i} = (2^{-1} + 2^{-2} + 2^{-3} + 2^{-4}) = 0.9375$$

Kraft inequality: Fixed length codes.

- Consider a 6 symbol alphabet: $\{x_1, x_2, x_3, x_4, x_5, x_6\}$
- If a fixed length code is to be used, how many digits are required for each symbol.
- Each codeword must be distinct.
- Answer: We must have 3 digits in every codeword.
- Compute K for this alphabet, and use Kraft's Inequality to appraise this code.
- $K = 6 \times 2^{-3} = 0.750$. This code is uniquely decodable, but not very efficient.