Formal Modeling in Cognitive Science 1 (2005–2006)

School of Informatics, University of Edinburgh
Lecturers: Mark van Rossum, Frank Keller

Tutorial 10: Codes; KL Divergence; Noisy Channel Model

Week 11 (20–24 March, 2006)

1. **Properties of Codes**

   Given are a random variable X and the codes C_1, C_2, and C_3 as follows:

x	a	b	c	d
$f(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
$C_1(x)$	0	1	10	11
$C_2(x)$	0	10	110	111
$C_3(x)$	0	00	000	0000

   (a) Describe each of the codes using the properties non-singular, uniquely decodable, and instantaneous.

   (b) Compute the expected code length $L(C)$ for each of the codes.

   (c) Which of the codes satisfies the Kraft inequality?

2. **Shannon and Huffman Coding; Kullback-Leibler Divergence**

   Suppose you have a corpus of size 25, which has 5 word types, each with the following frequencies:

John	Think	Said	Mary	Bill
5	7	3	8	2

   (a) Assume a random variable X that assigns each word a probability based on its corpus frequency. Compute the entropy of X.

   (b) Devise an optimal binary code for X using Shannon coding.

   (c) Devise an optimal instantaneous binary code for X using Huffman coding.

   (d) Compare the expected code lengths of the two codes.

   (e) Assign the Huffman code a distribution $g(x)$ based on its code lengths and compare this distribution to the original distribution $f(x)$ using the Kullback-Leibler divergence.

3. **Noisy Channel Model**

   (a) Assume a binary symmetric channel with the probability of error $p = 0.15$. The probability distribution over the input is given by $f(0) = 0.9$ and $f(1) = 0.1$. Assume we observe the output 1. What is the probability that it was generated by the input 1?

   (b) Word segmentation is the task of finding the word boundaries in a given string of letters. For example, it would involve turning the string "statisticalandphysicalmodelingcanbecombined" into the string "statistical and physical modeling can be combined". How can the noisy channel model be applied to this task?