Formal Modeling in Cognitive Science 1 (2005–2006)

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

Solutions for 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.

   Solution: C_1 is non-singluar, as it assigns each value of X a different code word. It is not uniquely decodable, as its extension is singluar. It is not instantaneous, as the codeword $C(b)$ is a prefix of $C(c)$ and $C(d)$.

   C_2 is non-singluar, as it assigns each value of X a different code word. It is also uniquely decodable, as its extension is also non-singluar. It is instantaneous, as no codeword is a prefix of any other codeword.

   C_3 is non-singluar, as it assigns each value of X a different code word. I It is not uniquely decodable, as its extension is singluar. It is not instantaneous, as several codewords are prefixes of other codewords.

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

   Solution:

$$\begin{aligned}
L(C_1) &= \textstyle\sum_{x\in X} f(x)l(x) = \frac{1}{2}\cdot 1+\frac{1}{4}\cdot 1+\frac{1}{8}\cdot 2+\frac{1}{8}\cdot 2 = 1.25 \\
L(C_2) &= \textstyle\sum_{x\in X} f(x)l(x) = \frac{1}{2}\cdot 1+\frac{1}{4}\cdot 2+\frac{1}{8}\cdot 3+\frac{1}{8}\cdot 3 = 1.75 \\
L(C_3) &= \textstyle\sum_{x\in X} f(x)l(x) = \frac{1}{2}\cdot 1+\frac{1}{4}\cdot 2+\frac{1}{8}\cdot 3+\frac{1}{8}\cdot 4 = 1.875
\end{aligned}$$

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

   Solution:

   The entropy of the distribution is:

$$H(X) = \sum_{x\in X} f(x)\log f(x) = \frac{1}{2}\log\frac{1}{2}+\frac{1}{4}\log\frac{1}{4}+\frac{1}{8}\log\frac{1}{8}+\frac{1}{8}\log\frac{1}{8} = 1.75$$

   Only C_2 assigns code words with the optimal code lengths given by the Shannon information $l(x) = -\log f(x)$. Hence the source code theorem holds for it and $H(X) \leq L(C_2) < H(X)+1$.

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.

Solution: In order to do this we need to calculate the probability $f(x)$ of each word in the corpus:

x	John	Think	Said	Mary	Bill
$f(x)$	0.2	0.28	0.12	0.32	0.08

Once you have worked out the probabilities then use the formula in (1c) to compute the entropy of $H(X)$.

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

Solution: To get an optimal code, first compute the Shannon information $-\log f(x)$ for each word. Then assign codewords of length $l(x)$ that approximate the Shannon information (we round fractional code lengths):

x	John	Think	Said	Mary	Bill
$f(x)$	0.2	0.28	0.12	0.32	0.08
$-\log f(x)$	2.32	1.84	3.06	1.64	3.6
$l(x)$	2	2	3	2	4

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

Solution: A Huffman code for this distribution is as follows:

x	John	Think	Said	Mary	Bill
$f(x)$	0.2	0.28	0.12	0.32	0.08
$C(x)$	11	00	100	01	101
$l(x)$	2	2	3	2	3

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

Solution: The Shannon code has a higher expected code length, as it assigns a longer codeword to one of the words.

(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.

Solution:

x	John	Think	Said	Mary	Bill
$g(x) = 2^{-l(x)}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$

$$D(f||g) = \sum_{x \in X} f(x) \log \frac{f(x)}{g(x)} = -0.064 + 0.046 - 0.007 + 0.114 - 0.052 = 0.044$$

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?

Solution: We can write the distribution of the input given the output (the posterior) using Bayes' theorem:

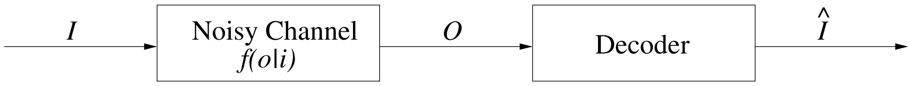
$$f(x|y) = \frac{f(y|x)f(x)}{f(y)} = \frac{f(y|x)f(x)}{\sum_z f(y|z)f(z)}$$

We know the distributions $f(y|x)$ and $f(x)$, so we can use them to compute the probability that the output 1 was generated by the input 1:

$$\begin{aligned} f(1|1) &= \frac{f(1|1)f(1)}{f(1|1)f(1)+f(1|0)f(0)} \\ &= \frac{0.85 \cdot 0.1}{0.85 \cdot 0.1 + 0.15 \cdot 0.9} = 0.39 \end{aligned}$$

(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?

Solution: We can use the noisy channel model as applied to linguistic input:



Assume that the input I is a segmented string of letters, which is passed through a noisy channel and output as an unsegmented string of letter O. We want to compute $\hat{I}$, an estimate of the original input string. The input has the distribution $f(i)$, which is a language models over segmented text. The noisy channel has the distribution $f(o|i)$, which corresponds to the conditional distribution of unsegmented strings given segmented ones.

We compute $\hat{I}$ using Bayes' theorem:

$$\hat{I} = \arg\max_i f(i|o) = \arg\max_i \frac{f(i)\,f(o|i)}{f(o)} = \arg\max_i f(i)\,f(o|i)$$

The distribution $f(o|i)$ can be estimated using data in which segmented and unsegmented strings are aligned.