Linguistics 696f
Hammond
Spring '03

Information Theory

A. Overview

(1)
a. Entropy
b. Huffman encoding
c. N-grams
d. Cross-entropy
e. Software

B. Entropy

(2) For comparison purposes, assume that all information is encoded with binary numbers: *bits*.

(3) $\{a, b, c, d\} = \{00, 01, 10, 11\}$

(4) Entropy
The number of digits needed in a binary encoding to represent all possible items.

C. Huffman encoding

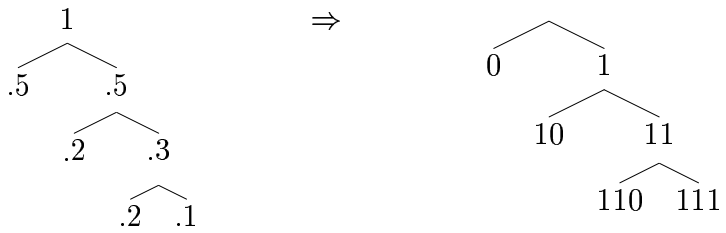
(5) Why not represent $\{a, b, c, d\}$ as $\{0, 1, 10, 11\}$?

(6) How is 1011 parsed?: $1-0-1-1$ or $10-1-1$ or $1-0-11$ or $10-11$

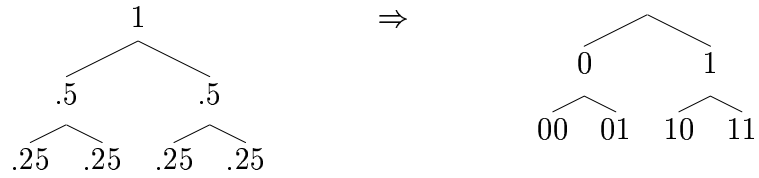
(7) There are other ways of creating a parsable binary encoding for four items, e.g. $\{1, 00, 010, 011\}$.

(8) A Huffman tree:
a. Make a binary tree from the two lowest-frequency elements.
b. Their parent node is assigned the sum of their probabilities.
c. Keep doing this until all elements are in the tree.
d. Assign 0 and 1 to each pair of nodes.
e. Read the code off the tree.

(9)



(10)



D. Formulas

(11) Entropy: $H(\mathrm{p}) = -\sum \mathrm{p}(x) \log \mathrm{p}(x)$

(12) Per-word entropy: $H_{\mathrm{rate}} = -\frac{1}{n} \sum \mathrm{p}(x_{1n}) \log \mathrm{p}(x_{1n})$

(13) Shannon-McMillan-Breiman theorem: $H(L) = \lim_{n\to\infty} \frac{1}{n} \log \mathrm{p}(w_1 w_2 \ldots w_n)$

(14) Perplexity: $2^{H(L)}$

E. Entropy & n-grams

(15) Unigram entropy of four texts:

Text	Entropy	Perplexity
a b a b a b a b a b	1	2
a b c d e a b c d e	2.32	5
a b a b a a a a a a	0.72	1.65
a a a a a b b b b b	1	2

(16) Bigram entropy of the first and fourth texts:

Text	Entropy	Perplexity
a b a b a b a b a b	0.128	1.093
a a a a a b b b b b	0.489	1.404

F. Cross entropy

(17) Cross entropy:
$H(\mathrm{p}, \mathrm{q}) = -\sum \mathrm{p}(x) \log \mathrm{q}(x)$

(18) By the Shannon-Breiman-McMillan theorem:
$H(\mathrm{p}, \mathrm{q}) = \lim_{n\to\infty} \frac{1}{n} \log \mathrm{q}(w_1 w_2 \ldots w_n)$

(19) Cross-entropy using the four texts in (15):

Text	1	2	3	4
1 (a b a b a b a b a b)	1	∞	1	1
2 (a b c d e a b c d e)	2.32	2.32	2.32	2.32
3 (a b a b a a a a a a)	1.32	∞	0.72	1.32
4 (a a a a a b b b b b)	1	∞	1	1

(20) $H(\mathrm{p}) \leq H(\mathrm{p}, \mathrm{q})$

G. Software

(21) All software is invoked on the command-line as follows: `perl program-name` If arguments are required, a suitable error message is displayed.

(22) a. `entropy.pl`: computes the per-word entropy of a text using unigrams.
b. `entropy2.pl`: calculates the per-word entropy of a text using bigrams

References

Charniak, Eugene. 1993. *Statistical Language Learning*. Cambridge: MIT Press.

Hopcroft, J.E., & J.D. Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation*. Reading: Addison-Wesley.