Formal Modeling in Cognitive Science

Lecture 26: Entropy Rate; Mutual Information

Frank Keller

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

March 6, 2006

Entropy Rate

Entropy rate takes the length of the message into account:

Definition: Entropy Rate

The entropy rate of a sequence of random variables $X_1, X_2, \dots, X_n$ is defined as:

$$\begin{aligned} H_{\text{rate}} &= \frac{1}{n} H(X_1, X_2, \dots, X_n) \\ &= -\frac{1}{n} \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \cdots \sum_{x_n \in X_n} f(x_1, x_2, \dots, x_n) \log f(x_1, x_2, \dots, x_n) \end{aligned}$$

Note that we have to extend our notion of joint distribution $f(x, y)$ and joint entropy $H(X, Y)$ to arbitrarily many random variables.

Entropy Rate

- Entropy depends on the length of the message; longer messages have higher entropy (all else being equal);
- entropy rate takes this into account, it normalizes by n, the length of the message;
- intuitively, entropy rate is the entropy per character or per word in a message.

Example: simplified Polynesian

In the previous example, we computed the joint entropy of a consonant and a vowel. The per character entropy is:

$$H_{\text{rate}} = \frac{1}{n} H(X_1, X_2, \dots, X_n) = \frac{1}{2} H(C, V) = 1.218 \text{ bits}$$

Shannon's Experiments

Guessing game: an experimental subject is given a sample of English text and is asked to guess the next letter (Shannon 1951).

Assumption: subject will guess the most probably letter first, then the second most probable letter, etc.

This way we get a probability distribution over the number of guesses required to get the correct letter:

No. of guesses	1	2	3	4	5	> 5
Probability	0.79	0.08	0.03	0.02	0.02	0.06

Shannon's Experiments

Then we can then use this distribution to compute the entropy rate of English. The results show:

- $H_{\text{rate}}(\text{English})$ is between 0.6 and 1.3 bits per character if estimated by humans;
- if humans gamble on the outcome, $H_{\text{rate}}(\text{English})$ is between 1.25 and 1.35 bpc;
- if we estimated it from a 500M word corpus, then $H_{\text{rate}}(\text{English})$ 1.75 bpc.

Modern estimates use word-guessing, not letter-guessing.

Mutual Information

Definition: Mutual Information

If X and Y are discrete random variables and $f(x, y)$ is the value of their joint probability distribution at (x, y), and $f(x)$ and $f(y)$ are the marginal distributions of X and Y, respectively, then:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} f(x, y) \log \frac{f(x, y)}{f(x) f(y)}$$

is the mutual information (MI) of X and Y.

Intuitively, mutual information is the reduction in uncertainty of X due to the knowledge of Y.

Mutual Information

We can also express mutual information in terms of entropy:

Theorem: Mutual Information

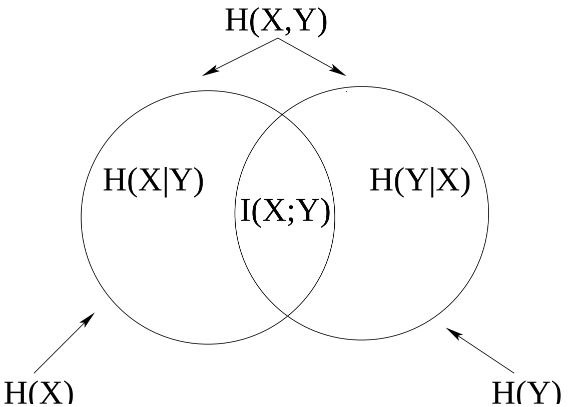
If X and Y are discrete random variables with joint entropy $H(X, Y)$ and the marginal entropy of X is $H(X)$, then:

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X, Y) \end{aligned}$$

This follows from the definition of conditional entropy in terms of joint entropy.

Mutual Information

The relationship between mutual information and entropy can be visualized using a Venn diagram:



Mutual Information

Properties of mutual information:

- Intuitively, $I(X;Y)$ is the amount of information X and Y contain about each other;
- $I(X;Y) \geq 0$ and $I(X;Y) = I(Y;X)$;
- $I(X;Y)$ is a measure of the *dependence* between X and Y:
  - $I(X;Y) = 0$ if and only if X and Y are independent;
  - $I(X;Y)$ grows not only with the dependence of X and Y, but also with $H(X)$ and $H(Y)$;
- $I(X;X) = H(X)$; entropy as "self-information" of X.

Mutual Information

Example: simplified Polynesian

Back to simplified Polynesian, with the following joint probability distribution:

$f(x,y)$	p	t	k	$f(y)$
a	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{1}{16}$	$\frac{1}{2}$
i	$\frac{1}{16}$	$\frac{3}{16}$	0	$\frac{1}{4}$
u	0	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
$f(x)$	$\frac{1}{8}$	$\frac{3}{4}$	$\frac{1}{8}$	

Let's compute the mutual information of a consonant and a vowel:

$$I(V;C) = H(V) - H(V|C)$$

Mutual Information

Example: simplified Polynesian

First compute the entropy of a vowel:

$$\begin{aligned} H(V) &= -\sum_{y \in V} f(y) \log f(y) \\ &= -(\frac{1}{2}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{4} + \frac{1}{4}\log\frac{1}{4}) \\ &= 1.5 \text{ bits} \end{aligned}$$

We have already computed $H(V|C) = 1.375$ bits (last lecture), so we can now compute:

$$I(V;C) = H(V) - H(V|C) = 0.125 \text{ bits}$$

Pointwise Mutual Information

- Mutual information is defined over *random variables.*
- Pointwise mutual information is defined over *values of random variables;*
- Example: MI over vowels and consonants; pointwise MI over the letters a and p;
- Intuitively, pointwise MI is the amount of information provided by the occurrence of event y about the occurrence of event x.

Pointwise Mutual Information

Definition: Pointwise Mutual Information

If X and Y are discrete random variables with the joint distribution $f(x, y)$ and the marginal distributions $f(x)$ and $f(y)$, then:

$$I(x; y) = \log \frac{f(x, y)}{f(x)f(y)}$$

is the pointwise mutual information at (x, y).

Pointwise Mutual Information

Example: simplified Polynesian

Compute the pointwise mutual information of a and p and of i and p:

$$I(a; p) = \log \frac{f(a, p)}{f(a)f(p)} = \log \frac{\frac{1}{16}}{\frac{1}{8} \cdot \frac{1}{2}} = 0$$

$$I(i; p) = \log \frac{f(i, p)}{f(i)f(p)} = \log \frac{\frac{1}{16}}{\frac{1}{8} \cdot \frac{1}{4}} = 1$$

Summary

- Entropy rate is the per-word or per-character entropy;
- the entropy rate of English can be estimated using experiments with humans or approximated using a large corpus;
- mutual information $I(X; Y)$ is the reduction in uncertainty of X due to the knowledge of Y;
- graphically, it's the intersection of two entropies;
- if X and Y are independent, then $I(X; Y) = 0$;
- pointwise mutual information: same for points of a distributions, instead of for the whole distribution.

References

Shannon, Claude E. 1951. Prediction and entropy of printed English. *Bell Systems Technical Journal* 30:50–64.