Statistische Methoden

Information Theory, Part II

Matthew Crocker

Computerlinguistik
Universität des Saarlandes

Recall ...

- Entropy: $$H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$
- In general:
  - Entropy measures: uncertainty, sometimes called *self-information*
  - Entropy is a lower bound for the average number of bits required
  - Entropy measures the quality of our models

- Joint Entropy: the amount of information necessary to specify the value of two discrete random variables:

$$H(p(x,y)) = H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x,y)$$

- Conditional Entropy: the amount of information needed to communicate Y, given that message X has been communicated:

$$H(p(y \mid x)) = H(Y \mid X) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(y \mid x)$$

Polynesian revisited

- Assume the following (slightly different) per-letter frequencies:

p	t	k	a	i	u
1/16	3/8	1/16	1/4	1/8	1/8

  - H(X)= 2 x 1/16 $\log_2$ 16+2 x 1/8 $\log_2$ 8 + 1/4 $\log_2$ 4 +3/8 $\log_2$ 8/3
    = 1/2+3/8+1/2+3/8 $\log_2$ 8/3 = 1.9 per letter

  - Suppose we discover that, in Simplified Polynesian,all words consist of Consonant-Vowel (CV) sequences. (note: margin probs are per syllable, not per letter - thus, twice their per-letter probs)

	p	t	k	
a	1/16	3/8	1/16	1/2
i	1/16	3/16	0	1/4
u	0	3/16	1/16	1/4
	1/8	3/4	1/8	

- We can calculate H(C,V) directly from the table, i.e. treat each possible pair (syllable) as an event:
  - H(C,V)=1/4 $\log_2$ 16 + 6/16 $\log_2$ 16/3 + 3/8 $\log_2$ 8/3 = 2.436 per syllable

Chain rule for joint entropy

- Chain rule for entropy:

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(x,y) \\
&= -E_{p(x,y)}(\log_2 p(x,y)) \\
&= -E_{p(x,y)}(\log_2 p(x)p(y\mid x)) \\
&= -E_{p(x,y)}(\log_2 p(x) + \log_2(p(y\mid x)) \\
&= -E_{p(x)}(\log_2 p(x)) - E_{p(x,y)}(\log_2 p(y\mid x)) \\
&= H(X) + H(Y\mid X)
\end{aligned}
$$

- In general

$$H(X_1,\ldots,X_n) = H(X_1) + H(X_2\mid X_1) + \ldots + H(X_n\mid X_1,\ldots,X_{n-1})$$

Polynesian continued

	p	t	k	
a	1/16	3/8	1/16	1/2
i	1/16	3/16	0	1/4
u	0	3/16	1/16	1/4
	1/8	3/4	1/8	

- H(C,V)=H(C)+H(V|C)

$$H(C) = 2 \times \frac{1}{8}\log_2 8 + \frac{3}{4}\log_2 \frac{4}{3}$$
$$= \frac{3}{4} + \frac{3}{4}(2 - \log_2 3) = \frac{9}{4} - \frac{3}{4}\log_2 3 \approx 1.061$$

$$H(V \mid C) = \sum_{c=p,t,k} p(C=c)H(V \mid C=c)$$
$$= \frac{1}{8}H(V \mid p) + \frac{1}{8}H(V \mid k) + \frac{3}{4}H(V \mid t)$$
$$= \frac{1}{8}H(\frac{1}{2},\frac{1}{2},0) + \frac{1}{8}H(\frac{1}{2},0,\frac{1}{2}) + \frac{3}{4}H(\frac{1}{2},\frac{1}{4},\frac{1}{4})$$
$$= 2 \times \frac{1}{8} \times 1 + \frac{3}{4}(\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2)$$
$$= \frac{1}{4} + \frac{3}{8} + \frac{3}{8} + \frac{3}{8} = \frac{11}{8} \approx 1.375$$

$$H(C,V) = H(C) + H(V \mid C)$$
$$= 1.061 + 1.375 = 2.436$$

Entropy rate

- Since information in a message depends on message length, we often normalize to the per-letter/per-word entropy rate:

$$H_{rate} = \frac{1}{n} H(X_1,...,X_n) = \frac{1}{n} H(X_{1n}) = -\frac{1}{n} \sum_{X_{1n}} p(x_{1n}) \log_2 p(x_{1n})$$

- Entropy rate for language:
  - "Language" is a stochastic process generating a sequence of tokens, $L=(X_i)$ e.g., all the words you hear, utter, appear in Die Zeit, etc...
  - We define the entropy of the language as the entropy rate for that process:

$$H_{rate}(L) = \lim_{n \to \infty} \frac{1}{n} H(X_1,..., X_n)$$

  - Recall: $H(X_1,..., X_n) = H(X_1) + H(X_2 \mid X_1) + ... + H(X_n \mid X_1,..., X_{n-1})$
  - Or, "the entropy rate of language is the limit of the entropy rate of a sample of the language, as the sample gets longer and longer" (M&S)

So, what is information?

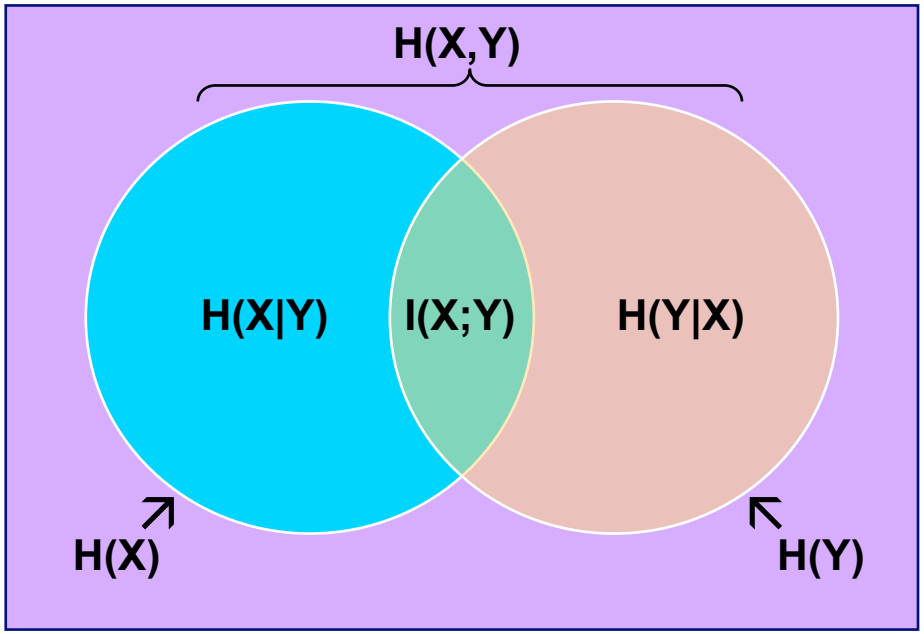
It's a change in what you *don't know*.

It's a change in the entropy.

$$I(x;y) = H(x) - H(x \mid y)$$

Mutual Information

- Recall: chain rule for entropy
  - H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)
- Therefore:
  - H(X) - H(X|Y) = H(Y) - H(Y|X) = I(X;Y)
- Mutual Information: The reduction in uncertainty for one variable due to knowing about another.



Mutual Information, continued

- Calculating Mutual Information:

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= H(X) + H(Y) - H(X,Y) \\
&= \sum_x p(x) \log_2 \frac{1}{p(x)} + \sum_y p(y) \log_2 \frac{1}{p(y)} + \sum_{x,y} p(x,y) \log_2 p(x,y) \\
&= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}
\end{aligned}
$$

- Mutual Information:
  - Symmetric, non-negative measure of common information
  - Measures the distance of a joint distribution from independence
  - *I(X;Y)* = 0 when *X,Y* are independent
  - MI grows as a function of *both* dependence and entropy

Simplified Polynesian

	p	t	k	
a	1/16	3/8	1/16	1/2
i	1/16	3/16	0	1/4
u	0	3/16	1/16	1/4
	1/8	3/4	1/8	

- Recall the following per-syllable distribution

$$H(V \mid C) = \sum_{c = p,t,k} p(C = c) H(V \mid C = c)$$

$$= \frac{1}{8} H(V \mid p) + \frac{1}{8} H(V \mid k) + \frac{3}{4} H(V \mid t)$$

$$= \frac{1}{8} H(\frac{1}{2}, \frac{1}{2}, 0) + \frac{1}{8} H(\frac{1}{2}, 0, \frac{1}{2}) + \frac{3}{4} H(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$$

$$= 2 \times \frac{1}{8} \times 1 + \frac{3}{4} (\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2)$$

$$= \frac{1}{4} + \frac{3}{8} + \frac{3}{8} + \frac{3}{8} = \frac{11}{8} \approx 1.375$$

$$I(V;C) = H(V) - H(V \mid C)$$

$$H(V) = 2 \times \frac{1}{4} \log_2 4 + \frac{1}{2} \log_2 2 = \frac{3}{2}$$

$$H(V \mid C) = \frac{11}{8}$$

$$I(V;C) = \frac{12}{8} - \frac{11}{8} = \frac{1}{8}$$

$$I(V;C) = \sum_{x,y} p(v,c) \log_2 \frac{p(v,c)}{p(v)p(c)}$$

$$= \frac{1}{16} \log_2 \frac{\frac{1}{16}}{\frac{1}{16}} + \frac{3}{8} \log_2 \frac{\frac{3}{8}}{\frac{3}{8}} + \frac{1}{16} \log_2 \frac{\frac{1}{16}}{\frac{1}{16}} + \frac{1}{16} \log_2 \frac{\frac{1}{16}}{\frac{1}{32}} + \frac{3}{16} \log_2 \frac{\frac{3}{16}}{\frac{3}{16}} + \frac{3}{16} \log_2 \frac{\frac{3}{16}}{\frac{3}{16}} + \frac{1}{16} \log_2 \frac{\frac{1}{16}}{\frac{1}{32}}$$

$$= \frac{1}{16} + \frac{1}{16} = \frac{1}{8}$$

Mutual Information

- Recall, Mutual Information: a measure of the reduction in uncertainty for one random variable due to knowing about another:

$$I(X;Y) = H(X) - H(X \mid Y)$$

$$= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

  - Sometimes called *average mutual information*
- Pointwise Mutual Information of two individual elements as a measure of association:

$$I(x',y') = \log_2 \frac{p(x',y')}{p(x')p(y')}$$

$$= \log_2 \frac{p(x' \mid y')}{p(x')} \qquad = \log_2 \frac{p(y' \mid x')}{p(y')}$$

  - *"The amount of information provided by the occurrence of event y´ about the occurrence of event x´."*

Computing Mutual Information

- We can compute the probabilities using ML estimation:

$$I(x', y') = \log_2 \frac{p(x', y')}{p(x')p(y')}$$

$$= \log_2 \frac{\frac{c(w^1 w^2)}{N}}{\frac{c(w^1)}{N} \times \frac{c(w^2)}{N}} \qquad = \log_2 \frac{N \times c(w^1 w^2)}{c(w^1)c(w^2)}$$

- Simple example:

	W_1=new	W_1≠new	
W_2=companies	8	4667	4675
W_2≠companies	15820	14287181	14303001
	15828	14291848	14307676

$$I(new, companies) = \log_2 \frac{14307676 \times 8}{15828 \times 4675} \approx .63$$

More on Mutual Information

- MI provides a similar ranking as the *t* test:

Word 1	Word 2	C(w1)	C(w2)	C(w1 w2)	I(w1,w2)
Ayatollah	Khomeini	42	20	20	18.38
Agatha	Christie	30	117	20	16.31
cassette	recorder	77	59	20	15.94
unsalted	butter	24	320	20	15.19
over	many	13484	10570	20	1.01
like	people	14093	14776	20	0.46
time	last	15019	1569	20	0.29

- Consider translation:
  - Canadian Hansards: “House of Commons” and “Chambre de communes”
  - What is a good translation of “house”
  - MI fails to capture the fact that *house* usually occurs <u>without</u> *communes*
  - X^2 makes the right ranking

	chambre	¬chambre	MI	X^2
house	31950	12004		
¬house	4793	848330	4.1	553610
	communes	¬communes		
house	4974	38980		
¬house	441	852682	4.2	88405

Entropy and Language Models

- So far, we have used entropy to find the most efficient code for transmitting messages.

- Recall Simplified Polynesian: 6 letters
  - Naive model 3 bits assumes uniform distribution
  - Zero-order model 2.5 bits uses per-letter frequencies
  - Syllable model 1.22 bits uses syllabic constraints/structure
  - : : :
  - Language model ? Bits uses language structure

- The more structure and regularities a model captures, the lower our uncertainty, or entropy, will be.

- We can use entropy as a measure of the quality of our models

Relative Entropy

- For two PMFs, p(x) and q(x), for an event space X, we can compute relative entropy as follows:

$$D(p \,\|\, q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)}$$

  - Also known as: Kullback-Leibler(KL) divergence

- KL-divergence compares the entropy of the two distributions

- Intuitively, the KL-divergence between *p* and *q* is the average number of bits that are wasted (or the additional bits required) by encoding events from a distribution *p* with a code based on distribution *q*.
  - Non-symmetric

Relative Entropy

Theorem: Properties of the Kullback-Leibler Divergence

1. $D(f||g) \geq 0$;
2. $D(f||g) = 0$ iff $f(x) = g(x)$ for all $x \in X$;
3. $D(f||g) \neq D(g||f)$;
4. $I(X;Y) = D(f(x,y)||f(x)f(y))$.

- Recall that Mutual Information measures the distance of a joint distribution from independence, thus Mutual Information and Relative Entropy are related in the following way:

$$I(X;Y) = D(p(x,y) \,\|\, p(x)p(y))$$

$$= \sum_{x,y \in X,Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

Evaluating language models with entropy

- Often, we want to construct a probabilistic model of some linguistic phenomena.
  - Represent events (e.g. letters, words, or sentences that 'occur') by *X*
  - Assume some true probability distribution for *X*: *p(x)*
  - In building a model, *m*, of *p*, we want to minimise *D(p||m)*
- Cross entropy: $H(X,q) = H(X) + D(p \,\|\, m)$

$$= -\sum_x p(x)\log_2 p(x) + \sum_x p(x)\log_2 \frac{p(x)}{m(x)}$$

$$= \sum_x p(x)\log_2 \frac{p(x)}{m(x)} - p(x)\log_2 p(x)$$

$$= \sum_x p(x)\log_2 p(x) + p(x)\log_2 \frac{1}{m(x)} - p(x)\log_2 p(x)$$

$$= \sum_x p(x)\log_2 \frac{1}{m(x)}$$

Entropy of English

- Per-letter entropy of English:
  - ASCII = 8 log 256
  - Uniform = 4.76 log 27
  - Unigram = 4.03 first order
  - Bigram = 2.8 second order
  - Gzip = 2.5
  - Trigram = 1.76 (Brown *et al.* 1992)
  - Human = 1.25 (Shannon)
    1.34 (Cover & Thomas)
- A notational variant of cross entropy is perplexity:

$$\text{perplexity}(X_{1n}, m) = 2^{H(X_{1n}, m)}$$

$$= m(X_{1n})^{-\frac{1}{n}}$$

- ... for when bigger is better: “perplexity of *k* means you are as suprised as if you had to guess between *k* equiprobable choices”

Summary: Information Theory

- Entropy:
  - Measures the average uncertainty present for a single random variable
  - More ‘knowledge’ means lower uncertainty, entropy
  - We represent this as the number of bits required, on average, to transmit an event.
- Entropy and Language Modelling:
  - Models with lower entropy can be considered better since they presumably encode more knowledge about the structure and relationships of the modelled language.
- Relative Entropy: D(p||q), the distance between two pmfs
- Cross entropy $H(X,m)=H(X)+D(p||m)$
  - Task: find the model, *m*, which minimises cross entropy!