Mathematical Foundations

FSNLP, chapter 2

Christopher Manning and Hinrich Schütze

© 1999–2002

Entropy

- Entropy is the average uncertainty of a random variable:

  Entropy $H(\mathrm{p}) = H(X) = -\sum_{x \in X} \mathrm{p}(x) \log_2 \mathrm{p}(x)$

- Suppose you are reporting the result of rolling an 8-sided die. Then the entropy is:

  $$H(X) = -\sum_{i=1}^{8} \mathrm{p}(i) \log \mathrm{p}(i) = -\sum_{i=1}^{8} \frac{1}{8} \log \frac{1}{8} = -\log \frac{1}{8} = 3 \text{ bits}$$

- Entropy measures are expectations:

  $$H(X) = E\left(\log \frac{1}{\mathrm{p}(X)}\right)$$

Simplified Polynesian

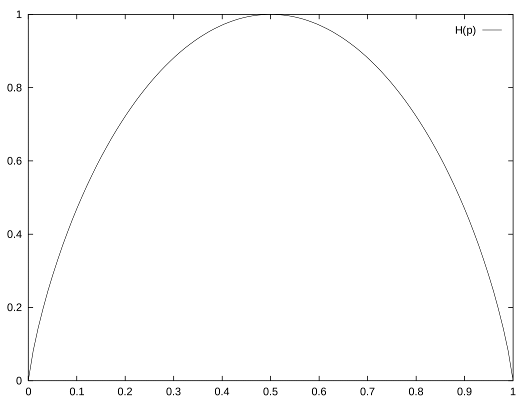
- Simplified Polynesian appears to be just a random sequence of letters, with the letter frequencies as shown:
p	t	k	a	i	u
1/8	1/4	1/8	1/4	1/8	1/8
- Then the per-letter entropy is:

  $$\begin{aligned} H(P) &= -\sum_{i \in \{p,t,k,a,i,u\}} P(i) \log P(i) \\ &= -[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4}] = 2\frac{1}{2} \text{ bits} \end{aligned}$$

  We can design a code that on average takes $2\frac{1}{2}$ bits a letter:

p	t	k	a	i	u
100	00	101	01	110	111

The entropy of a weighted coin



Joint/conditional entropy

- Joint entropy:

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

- Conditional entropy:

  $$\begin{aligned} H(Y|X) &= \sum_{x \in X} \mathrm{p}(x) H(Y|X = x) \\ &= \sum_{x \in X} \mathrm{p}(x) \left[-\sum_{y \in \mathcal{Y}} \mathrm{p}(y|x) \log \mathrm{p}(y|x) \right] \\ &= -\sum_{x \in X} \sum_{y \in \mathcal{Y}} \mathrm{p}(x,y) \log \mathrm{p}(y|x) \end{aligned}$$

The relation of mutual information I to entropy H

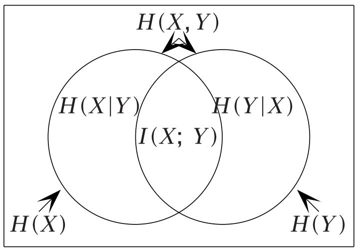
- By the chain rule for entropy,

  $$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

  Therefore,

  $$H(X) - H(X|Y) = H(Y) - H(Y|X) = I(X;Y)$$

- This is called the *mutual information* between X and Y



Mutual information

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

- Since $H(X|X) = 0$, note that:

$$H(X) = H(X) - H(X|X) = I(X;\, X)$$

- Hence, entropy is also called self-information
- MI of dependent variables depends on their entropy

Mutual information in classification problems

- A large mutual information means that knowing a lot about one variable tells you a lot about the other
- It is thus a useful metric for various problems in building supervised classification systems:
  - The standard decision tree induction algorithm uses MI to decide the attribute on which to split a node:
    - information gain= $I(C;\, A)$
  - Brown et al. (1991a) used MI to find the most informative single feature for WSD, and to partition its values to correspond to senses
  - MI is often used for feature selection in text categorization (e.g., with Bernoulli Naive Bayes models)

KL divergence or relative entropy

Two pmfs $\mathrm{p}(x)$ and $\mathrm{q}(x)$:

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

Say $0 \log \frac{0}{\mathrm{q}} = 0$, otherwise $\mathrm{p} \log \frac{\mathrm{p}}{0} = \infty$.

$$D(\mathrm{p} \,\|\, \mathrm{q}) = E_{\mathrm{p}} \left(\log \frac{\mathrm{p}(X)}{\mathrm{q}(X)} \right)$$

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

KL divergence

- Measure of how different two probability distributions are
- The average number of bits that are wasted by encoding events from a distribution p with a code based on a not-quite-right distribution q.
- $D(\mathrm{p} \,\|\, \mathrm{q}) \geq 0$; $D(\mathrm{p} \,\|\, \mathrm{q}) = 0$ iff $\mathrm{p} = \mathrm{q}$
- Not a metric: not commutative, doesn't satisfy triangle equality

Cross entropy

- Entropy = uncertainty
- Lower entropy = efficient codes = knowing the structure of the language = measure of model quality
- Entropy = measure of surprise
- How surprised we are when w follows h is pointwise entropy:

$$H(w|h) = -\log_2 \mathrm{p}(w|h)$$

$\mathrm{p}(w|h) = 1$? $\mathrm{p}(w|h) = 0$?

- Total surprise:

$$\begin{aligned} H_{\text{total}} &= -\sum_{j=1}^{n} \log_2 \mathrm{m}(w_j | w_1, w_2, \ldots, w_{j-1}) \\ &= -\log_2 \mathrm{m}(w_1, w_2, \ldots, w_n) \end{aligned}$$

Formalizing through cross-entropy

- Our model of language is $\mathrm{q}(x)$. How good a model is it?
- Idea: use $D(\mathrm{p} \,\|\, \mathrm{q})$, where p is the correct model.
- Problem: we don't know p.
- But we know roughly what it is like from a corpus
- Cross entropy:

$$\begin{aligned} H(X, \mathrm{q}) &= H(X) + D(\mathrm{p} \,\|\, \mathrm{q}) \\ &= -\sum_x \mathrm{p}(x) \log \mathrm{q}(x) \\ &= E_{\mathrm{p}}(\log \frac{1}{\mathrm{q}(x)}) \end{aligned}$$

Cross entropy

- Cross entropy of a language $L = (X_i) \sim \mathrm{p}(x)$ according to a model m:
$$H(L, \mathrm{m}) = -\lim_{n\to\infty} \frac{1}{n} \sum_{x_{1n}} \mathrm{p}(x_{1n}) \log \mathrm{m}(x_{1n})$$
- If the language is 'nice':
$$H(L, \mathrm{m}) = -\lim_{n\to\infty} \frac{1}{n} \log \mathrm{m}(x_{1n})$$
I.e., it's just our average surprise for large n:
$$H(L, \mathrm{m}) \approx -\frac{1}{n} \log \mathrm{m}(x_{1n})$$
- Since $H(L)$ is fixed if unknown, minimizing cross-entropy is equivalent to minimizing $D(\mathrm{p} \,\|\, \mathrm{m})$
- Assuming: independent test data, $L = (X_i)$ is stationary [does't change over time], ergodic [doesn't get stuck]

Entropy of English text

Character n-grams over a 27 letter alphabet

Model		Cross entropy (bits)	
"zeroth order"	uniform	4.75	$(\log 27)$
"first order"	unigram	4.03	
"second order"	bigram	3.32	
"third order"	trigram	3.1	
	fourgram	2.8	
Shannon's experiment		1.3 (1.34)	

(Cover and Thomas 1991: 140)

Perplexity

$$\begin{aligned} \mathrm{perplexity}(x_{1n}, \mathrm{m}) &= 2^{H(x_{1n}, \mathrm{m})} \\ &= \mathrm{m}(x_{1n})^{-\frac{1}{n}} \end{aligned}$$

(Note that $H(\cdot)$ has been overloaded between total entropy and entropy rate; perplexity is always per decision.)