Entropy and Perplexity

Raf Alvarado
UVA DS 5001

Concepts of entropy, information, and perplexity

Review

Source : Message :: Grammar : Discourse and Text

Language models

Probabilistic models of the **generative grammar** of the **source**

Probabilistic is just one kind

Also rules-based, discriminant (NNs), etc.

In other words ...

Source is a stochastic process comprising **random variables**

Parameters inferred from the message (as opposed to speculation about the internal mechanism)

Motivation

What is the best way to **encode a message** so that it can be interpreted accurately by its receiver?

Literally: How can we get our messages across?

Concretely: How can we ensure that **the message you read is the one that I wrote?**

I.e. How can we ensure that $\mathbf{M1} == \mathbf{M2}$? Or is *inferable* ...

`HELP WATSON` → `HELP WATSON!` **GOOD**

→ `HELL WATSON!` **BAD**

Note that this is not the same as how can we make sure we are *understood*.

That requires more . . .

Approach

One approach to the problem is to provide a **CODEC** that is **robust to message corruption**

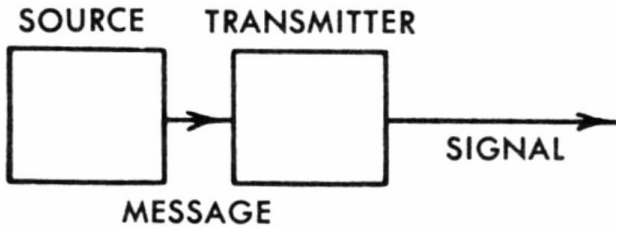
CODEC = Coder / Decoder (not just for video compression)

A CODEC is a set of **rules** for **encoding** and **decoding** a message

A mapping of one message onto another, A → 65 (ASCII)

A CODEC is **shared** by **transmitter** and **receiver**

Like Morse code in telegraphy



How can we design a CODEC

that will produce readable messages
when parts of the message are missing or incorrect?

Intuition

We can build on some observable properties of communication

First, languages are **redundant**

```
Y prbl cn ndrstnd ths sntnce.

Y-- pr-b--l- c-n -nd-rst-nd th-s s-nt-nce.
```

Shannon estimated that English is about 50% redundant

i.e. half the letters are can be removed and still be intelligible

Second, there appears to be an inverse relationship between **symbol length** and **symbol frequency**

See Morse Code . . .

A •–	J •–––	S •••	1 •———–
B –•••	K –•–	T –	2 ••–––
C –•–•	L •–••	U ••–	3 •••––
D –••	M ––	V •••–	4 ••••–
E •	N –•	W •––	5 •••••
F ••–•	O –––	X –••–	6 –••••
G ––•	P •––•	Y –•––	7 ––•••
H ••••	Q ––•–	Z ––••	8 –––••
I ••	R •–•	0 –––––	9 ––––•

More frequent letters are shorter because that **saves time and energy** when encoding ... But it also reveals something about the nature of language.

Insight

Redundancy and variable length suggest

That there are **properties of probability distributions**

That can be exploited to create an CODEC

If you make a message redundant enough

and you know how much corruption to expect in your channel

then you can make a message that survives transmission

Consider the following three simple language models for a symbol set $\mathbf{A} = (A_1, A_2, A_3, A_4)$

You can see how they differ in
how much uncertainty they possess

HIGH	
A_1	.25
A_2	.25
A_3	.25
A_4	.25

MEDIUM	
A_1	.25
A_2	.50
A_3	.125
A_4	.125

LOW	
A_1	0.0
A_2	0.0
A_3	1.0
A_4	0.0

also no information bc... well yk it's 0

this doesn't provide a lot of information (a broken record bc their saying the same thing over and over again)

Shannon's Entropy Formula

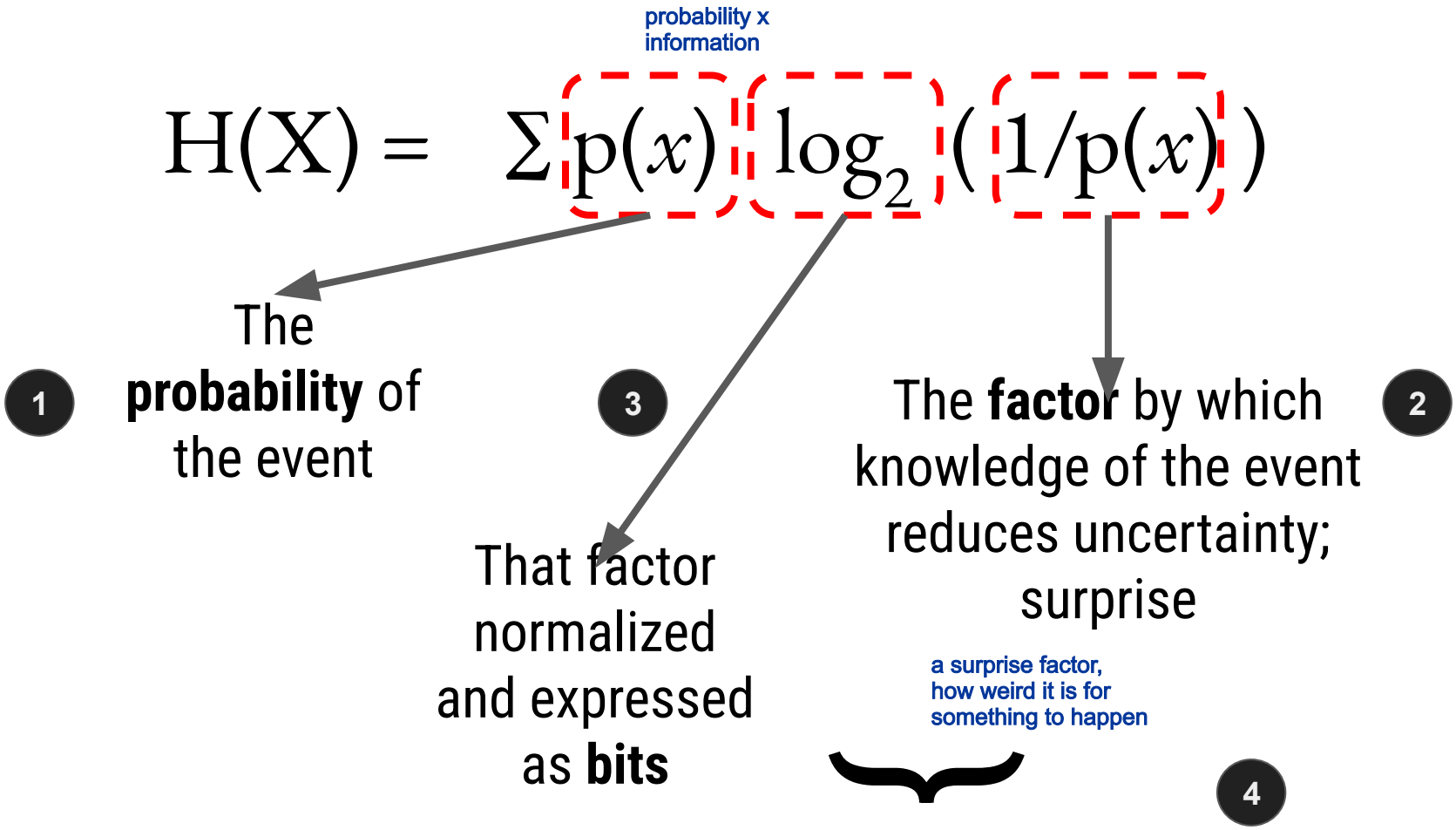
$$H(X) = -\Sigma\, p(x) \log_2(p(x))$$

$$H(X) = \Sigma\, p(x) \log_2(1/p(x))$$

We use **log base 2** because we want to our value to be expressed as **bits**

Think of bits as a **unit of measure for the property of information**

Shannon's Entropy Formula



Shannon's Entropy Formula

the expectation for surprise

$$H(X) = -\Sigma\, p(x) \log_2 p(x)$$

$$H(X) = \Sigma\, p(x) \log_2 (1/p(x))$$

$$i(x) = \log_2 (1/p(x))$$

information associated with x (this is what 'information' the information theory is talking about)

$$H(X) = \Sigma\, p\, i = \mathbf{p}\, \mathbf{i}$$

$$H(X) = \mathbf{E}[i(x)]$$

H	Entropy
X	A random variable; e.g. a language source
x	A member of X; e.g. a symbol
$p(x)$	the probability of x
$i(x)$	the information of x

Entropy and Redundancy

Entropy as a metric is **sensitive to symbol set size**

The **maximum entropy** of a random variable is just the log of the the number of possibilities (events)

Die $-(\frac{1}{2} * \log_2(\frac{1}{2}) * 2) = \log_2(2) = 1$

Coin $-(\frac{1}{6} * \log_2(\frac{1}{6}) * 6) = \log_2(6) = 2.59$

Alphabet (26 chars) $= \log_2(26) = 4.70$

Redundancy is a normalizing measure:

$$R(X) = 1 - H(X) / H_{MAX}(X)$$

Lower redundancy means more entropy

Ways to think of Entropy

As **property of the probability distribution** of a random variable (RV)

As degree of **uncertainty** (in the outcome of an RV)

As degree of **equiprobability**, i.e. maximum entropy = equiprobability

As average **minimum message length** of re-encoded messages

As **average number of decisions** to produce an outcome

As **size of search space** of the possible values

As **surprisal** -- how surprised we are to get a result

Entropy and Information

Entropy is often called "Shannon entropy"

Entropy and Information are often regarded as **synonyms**, but they are **not**

> **Entropy is the expectation of information**

We use the terms in **opposite** ways

> E.g. in a decision tree, "information **gain**" is **loss** of entropy (increase in certainty)
>
> They are opposite sides of the same coin (Stone)

Also -- the word "information" is **overloaded**

> Has many related but distinct meanings, e.g. entropy, form, news

Conjectures about Information

Information is **physical** but not **material**

- It can be measured — has magnitude — but has neither mass nor energy
  - "Information is information, not matter or energy. No materialism which does not admit this can survive at the present day." (Wiener 1948: 132)
- Inescapable hylomorphism!

Information is **uniquely a property of texts . . .**

- Entropy applies to other phenomena to the extent that we model these as text -- e.g. genes, stock market prices, etc.

"Information"

Ambiguous term . . .

- May mean "**news**"
- Sometimes used to mean **raw data** (as in "YouTube generates X amount of information per minute")
- Shannon meant it to refer to **optimally compressed** message length

Measured in **bits** — a bit is a unit of selection

- Proportional to message length

As normalized surprise

- Units of surprise are on the same order as message length

Importance to Text Analytics

Aside from a solution to the problems of communication, the concept of entropy provides **a way to think about text** detached from situation

> Entropy is a way to talk about the **significance** of text **without** talking about **meaning**, or semantics, which relies on reference

Natural Language Processing relies heavily on the concept of entropy and Shannon's theorems

> In fact, the core concept of NLP, the Language Model, begins with Shannon

Entropy is used to evaluate the **performance** of language models and to compute distance (divergence) between vectors ...

Cross Entropy

You can compare the entropy of two distributions (p, q) from the same source to get their difference

This is called **cross entropy**

$$H(p, q) = -\sum_{x} p(x) \log q(x)$$

p is the **real** language model and q the **estimated** model

As a performance measure, it measures the **entropy gap** between two distributions

AKA Cross Entropy **Loss**

Cross Entropy

Since we don't have access the real model, we substitute the maximum entropy for the real distribution (i.e. the case where all events are equally probable)

$$H(p, q) = -\sum_{x} p(x) \log q(x)$$

$$H(p, q) \approx -\frac{1}{N} \log_2 q(W)$$

cross entropy

Perplexity

Now, perplexity is just a way of expressing this cross entropy in a way that is considered more interpretable

$$H(W) = -\frac{1}{N}\log_2 P(W) = -\frac{1}{N}\log_2 P(w_1, w_2, \ldots, w_N)$$

$$PP(W) = 2^{H(W)} = 2^{-\frac{1}{N}\log_2 P(w_1, w_2, \ldots, w_N)}$$

So, perplexity is just the amount of information in a sentence (message) $w_1 \ldots w_N$ normalized by its length N (since surprise/entropy increases with N)

Perplexity

Measures on average, how many likely words we must choose from for the next word predicted by a language model

Smaller values mean more certainty (less surprise)

- And that our model is better

Perplexity

Just the amount of surprisal in a sentence (message) $w_1 \ldots w_N$ normalized by its length N, since surprise/entropy increases with N

$$\begin{aligned} \mathrm{PP}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}} \end{aligned}$$

In log form (i.e. log likelihood ℓ), we get this:

$$PP(W) = 2^{-\frac{\ell(W)}{N}}$$

$$i_2(W) = -\ell(W)$$

$$PP(W) = 2^{\frac{i_2(W)}{N}}$$

Cross Entropy, Information, and Perplexity

$$H(p,q) = -\sum_x p(x) log_2(q(x))$$

$$H(p,q) = \sum_x p(x) log_2(\frac{1}{q(x)})$$

$$i(x) = log_2(\frac{1}{q(x)})$$

$$H(p,q) = \sum_x p(x) i(x)$$

$$H(p,q) = \vec{p} \cdot \vec{i}$$

$$p_u = \frac{1}{N}$$

$$H_{cross} = H(p_u, q)$$

$$H_{cross} = \sum_x \frac{1}{N} i(x)$$

$$H_{cross} = \frac{1}{N} \sum_x i(x)$$

$$H_{cross} = \frac{\sum_x i(x)}{N}$$

$$PP(p_u, q) = 2^{H_{cross}}$$

Perplexity

Here's a simple way to think of Perplexity PP:

$$PP = 2^{\text{mean}(i)}$$

Probability	$p(x)$	Relative frequency of x in the long run. See Kolmogorov's axioms (1933).
Surprisal	$s = 1/p$	1 = The Banal ∞ = The Miraculous
Information	$i = \log_2(1/p)$ $i = -\log_2(p)$ $i = \log_2(s)$	Normalized surprise, expressed as bits. Related to message length.
Self Entropy	$h = p \log_2(1/p)$ $h = p \log_2(s)$ $h = p\, i$	Contribution of event to expectation of information.
Entropy	$H = E[i]$ $H = SUM(h)$	Expectation of information in a distribution. Also like "regularized" probability

Maximum Entropy	$H_{max} = \log_2(N)$	Assume distribution is equiprobable, so $p = 1/N$
Redundancy	$R = 1 - H/H_{max}$	Normalized H since H is unbounded. Low redundancy means low predictability.
Divergence	$D_{KL}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$ $JSD(P \parallel Q) = \frac{1}{2} D(P \parallel M) + \frac{1}{2} D(Q \parallel M)$	Non-symmetric distance measure between probabilistic vectors
Cross-entropy	$H(p, q) = -\sum_x p(x) \log q(x)$ $H(p, q) = H(p) + D_{KL}(p \Vert q)$	Compares two distributions over the same sample space, e.g. data and model.
Perplexity	$PP(W) = 2^{H(W)}$	A measure of how well a language model performs at predicting sentences.