

Information theory

- the maths of telegraphy
- provides powerful conceptual tools
 - entropy
 - conditional entropy
 - mutual information
 - divergence
- allows objective comparison of systems
- a basis for training and improvement of systems.

Telegraphy

Shannon developed Information Theory to answer questions about idealized (and real) telegraph systems.

- This effort produced some very useful abstractions of e.g. messages and the channels across which they pass.
- How good can (lossless) data compression be? It turns out that there is an upper bound on this.
- How fast can data be transmitted (perfectly)? It turns out that there is a channel-specific upper bound on this, and an optimal code which will approach the upper bound. Finding the code might still be hard.

Formulae

- Entropy

$$H(X) = -E[\log P(X)]$$

- Joint entropy

$$H(X, Y) = -E_{X,Y}[\log P(X, Y)]$$

- Conditional entropy

$$H(X|Y) = -E[\log P(X|Y)]$$

- Average mutual information

$$I(X; Y) = -E \left[\log \frac{P(X, Y)}{P(X)P(Y)} \right]$$

- Divergence

$$D(P||Q) = E_P \left[\frac{P(X)}{Q(X)} \right]$$

Entropy: How hard is a decision?

- How much choice is there?
 - Measure relative to some number of balanced binary decisions.
 - The natural measure of information is therefore *bits*.
 - Natural to use logarithms, these are usually base 2.
- Tossing a fair octahedral die gives

$$\log_2(8) = 3$$

bits of information. Before, we had a choice of 8, now the possibilities are reduced to 1. If we want to tell a friend about the outcome, we must give her at least 3 bits of information. In this case it is easy to do this, we just pass the binary code for the number. If the die were biased, a good code might be harder to find.

Designing an information measure: I

What must an information measure be like? Call it f for the moment. First consider choices from a uniform distribution of size M . If $M = NL$ we want $f(M) = f(N) + f(L)$ because you could either choose directly (top line) or hierarchically by first choosing the group, then the member,

1	2	3	4	5	6	7	8	9	10	11	12
Group 1			Group 2			Group 3			Group 4		
1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3	4.1	4.2	4.3

We want these two methods to cost the same amount.

Designing an information measure: II

- As M gets bigger, we want $f(M)$ to increase monotonically.
- Hierarchical decomposition. Generalised version of last slide, for non-uniform distributions. Split the choices into two groups governed by $P(\text{Group1}) = p, P(\text{Group2}) = 1 - p$. Want

$$H(\text{Group1} \cup \text{Group2}) = H(p, 1 - p) + pH(\text{Group1}) + (1 - p)H(\text{Group2})$$

- Want $H(p, 1 - p)$ to be a continuous function of p .

From all of these it follows (see Jelinek: *Statistical Methods for Speech Recognition* ch 7) that

$$H(p_0, p_1, \dots, p_{m-1}) = \sum_{i=0}^{m-1} p_i \log(p_i)$$

is the only choice for an information measure.

Simplified Polynesian

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

$$\begin{aligned}H(P) &= -E[\log(P)] \\&= -\left[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4}\right] \\&= -\left[4 \times -\frac{3}{8} + 2 \times -\frac{2}{4}\right] \\&= 5/2\end{aligned}$$

Entropy: Fair coin

$$H(P(v_1) \dots P(v_n)) = \sum_i^n -P(v_i) \log_2 P(v_i)$$

$$H\left(\frac{1}{2}, \frac{1}{2}\right) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1 \text{ bit}$$

Entropy: Biased coin

$$H(0.98, 0.02) = -0.98 \log_2 0.98 - 0.02 \log_2 0.02 = 0.14 \text{bits}$$

Data-Intensive Grocery Selection

Colour	Size	Skin	Eat it?
Red	Big	Shiny	No
Red	Small	Shiny	Yes
Red	Small	Rough	Yes
Green	Big	Rough	Yes
Green	Small	Rough	No
Brown	Small	Rough	Yes
Brown	Big	Rough	No
Brown	Small	Shiny	No

Goals

- Nobody gets poisoned
- We ask as few questions as possible

In General

$$I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$

Positive and negative examples

$$I\left(\frac{4}{4+4}, \frac{4}{4+4}\right) = -\frac{4}{4+4} \log_2 \frac{4}{4+4} - \frac{4}{4+4} \log_2 \frac{4}{4+4} = 1 \textit{bit}$$

Attribute tests

Choose on the basis of information content for each attribute.

$$\begin{array}{ll} p_{red} = 2 & n_{red} = 1 \\ p_{green} = 1 & n_{green} = 1 \\ p_{brown} = 1 & n_{brown} = 2 \end{array} \quad (1)$$

If we test i th attribute

$$I\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right) = -\frac{p_i}{p_i + n_i} \log_2 \frac{p_i}{p_i + n_i} - \frac{n_i}{p_i + n_i} \log_2 \frac{n_i}{p_i + n_i}$$

for *Colour*, this would be:

$$I\left(\frac{1}{1+2}, \frac{2}{1+2}\right) = -\frac{1}{1+2} \log_2 \frac{1}{1+2} - \frac{2}{1+2} \log_2 \frac{2}{1+2}$$

0.91 (call this *Remainder(Colour)*).

Choose best attribute

$$\begin{aligned} \textit{Remainder}(\textit{Colour}) &= \frac{3}{8}I\left(\frac{2}{3}, \frac{1}{3}\right) + \frac{2}{8}I\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{3}{8}I\left(\frac{1}{3}, \frac{2}{3}\right) \\ &= 0.91 \end{aligned}$$

which leaves less information to find than

$$\begin{aligned} \textit{Remainder}(\textit{Size}) &= \frac{3}{8}I\left(\frac{1}{3}, \frac{2}{3}\right) + \frac{5}{8}I\left(\frac{3}{5}, \frac{2}{5}\right) \\ &= 0.9475 \end{aligned}$$

or

$$\begin{aligned} \textit{Remainder}(\textit{Skin}) &= \frac{3}{8}I\left(\frac{1}{3}, \frac{2}{3}\right) + \frac{5}{8}I\left(\frac{3}{5}, \frac{2}{5}\right) \\ &= 0.9475 \end{aligned}$$

Joint and conditional entropy

$$H(X, Y) = -E_{X, Y}[\log P(X, Y)]$$

this is really just entropy with a shift of perspective, in which the pair X, Y is the new random variable. Because $P(X, Y) = P(X)P(Y|X)$

$$\begin{aligned} H(X, Y) &= -E_{X, Y}[\log P(X)P(Y|X)] \\ &= -E_{X, Y}[\log(P(X) + \log(P(Y|X))] \\ &= H(X) - E_{X, Y}[\log(P(Y|X))] \end{aligned}$$

so we dignify the last term with the name *conditional entropy* and the notation $H(Y|X)$. [Calculation from p65].

Mutual information

We just derived a chain rule for entropy:

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

from which it follows that

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

and we call this quantity $I(X; Y)$. If Y is informative about X then $H(X|Y) < H(X)$. In the limit Y determines X and $H(X|Y) = 0$.

Properties of Mutual information

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(X) - (H(X, Y) - H(Y)) \\ &= H(X) + H(Y) - H(X, Y) \\ &= -E_x[\log P(x)] - E_y[\log(P(y))] + E_{x,y}[\log P(x, y)] \\ &= E \left[\log \frac{P(x, y)}{P(x)P(y)} \right] \end{aligned}$$

The Noisy channel model

Another take on entropy

- Imagine watching ticker tape.
- How much information will I gain when I see s_i ?
- How predictable is s_i from its context?

- enumerate the possible next symbol – one of $s_1 \dots s_n$.
- From context $s_1 \dots s_{i-1}$ we have estimates of the probabilities $p(s_k | s_1 \dots s_{i-1})$
- Each such outcome will gain $-\log p(s_k | s_1 \dots s_{i-1})$ bits of information.
- Weighted sum of outcomes:

$$-\sum_{k=1}^n p(s_k | s_1 \dots s_{i-1}) \log p(s_k | s_1 \dots s_{i-1})$$

- Each time we see a symbol we are more or less surprised.
- If you can reliably predict the next symbol from context, you will not be surprised.
- the information gain will be low.

Cross entropy

- You may have a less good language model $p_M(w_k|s_1 \dots s_{i-1})$

$$H(P_M) = - \sum_{K=1}^n p(w_k|s_1 \dots s_{i-1}) \log p_M(w_k|s_1 \dots s_{i-1})$$

- In this quantity is called cross-entropy. It is always more than the entropy with the correct model.

Cross entropy as a measure of language models

- A good language model provides reliable predictions.
- Tends to *minimize* entropy.
- It also minimizes perplexity which is just $2^{\text{Crossentropy}}$
- If we can construct a series of models with decreasing cross entropy, we approach, and maybe reach, the correct model.

Measuring the entropy of English (Shannon)

- Uniform model, 27 chars equiprobable, so $\log_2(27) = 4.76$ bits per char
- Taking account of letter frequency 4.03
- Taking account of bigrams 2.8
- Human experiment 1.34
- The true entropy of English is lower than any of these.

Relative entropy: or divergence

- Compare two distributions $P(x)$ and $Q(x)$.
- Like $H(p) - H(q)$ which would be

$$\sum_{x \in X} P(x) \log(P(x)) - \sum_{x \in X} Q(x) \log(Q(x))$$

- But both expectations are calculated wrt. P

$$D(P||Q) = \sum_{x \in X} P(x) \log(P(x)) - \sum_{x \in X} P(x) \log(Q(x))$$

- As if P were the truth and Q an approximation. Always positive, because it is the difference between true and cross entropy.

- Divergence is also:

$$D(P||Q) = -E \left[\frac{P(x)}{Q(x)} \right]$$

- so mutual information:

$$I(X;Y) = -E \left[\log \frac{p(x,y)}{p(x)p(y)} \right]$$

is the divergence between the true distribution and the one you get if you assume independence of X and Y .