Uncertainty and Entropy

Information sources: uncertainty and entropy

- Using simple combinatorics, we know that the state of n bits corresponds to one of 2^n possible outcomes. Around 1921 Hartley observed that the information content of a system is the logarithm of the number of possible outcomes.

Example

A decimal digit represents one of 10 outcomes, so the information content of a decimal digit is $\log_2 10 = 3.3219$ bits

- Shannon noted that the information content depends on the probability of the events, not just on the number of outcomes.
- For example, a biased coin toss has less uncertainty than a fair coin toss and therefore less information is supplied (uncertainty reduced) by the toss of a biased coin.
- Uncertainty is lack of knowledge about an outcome, while information is what is obtained when the outcome is known.

Information source

Definition

An information source is a probability distribution (or density) on a set of outcomes.

In other words, an information source is a collection of random variables. Examples of information sources:

- Discrete: pdfs for sample spaces $\{H, T\}$ $\{1, 2, 3, 4, 5, 6\}$.
- Continuous: Gaussian random variable on $\mathbf{R}$, Exponential random variable on $\mathbf{R}+$, Gaussian vectors in $\mathbf{R}^n$.

Entropy definition and examples

- We are more surprised when an unlikely outcome occurs than when a likely outcome occurs. A useful measure of the surprise of an event with probability p is

$$\log\left(\frac{1}{p}\right) = -\log(p)$$

- Shannon defined the information content of a source to be the average surprise. He called this information content the entropy and used the symbol H, taken from statistical mechanics.

Definition

The entropy (uncertainty, self information) of a discrete random variable X is

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)} = -\sum_x p(x) \log p(x)$$

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)} = E_{p(x)}\left[\log \frac{1}{p(x)}\right]$$

- Note that the entropy of a random variable depends only on the probability of the outcomes and not on the values of the random variable. Thus the entropy is scale invariant and shift invariant.
- By convention, $0 \log 0 = -0 \log\left(\frac{1}{0}\right) = 0$, since $\lim_{x \to 0} x \log x = 0$. Thus, contributions to the entropy sum of a small probabilities is negligible. (But the contribution due to many small probabilities may be significant.)

- The units for entropy depend on the base of the logarithm used in the definition. When base 2 is used, information is measured in *bits* (binary digits).
- Base 10 corresponds to (decimal) *digits*.
- When natural logarithms (base e) are used, which is convenient for differential entropy of continuous random variables, the units are called *nats*.

Facts:

- $H(X) \geq 0$, since $\log \frac{1}{p(x)} \geq 0$ when $p(x) \leq 1$
- $H(X)$ is finite when the range of X is finite
- If X has uniform pdf on a range on n values, then

$$H(X) = \sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{1/n} = \log n$$

Less obvious fact:

- The maximum entropy for a random variable with n values is $\log n$, which is achieved only by the uniform distribution[1].

[1]

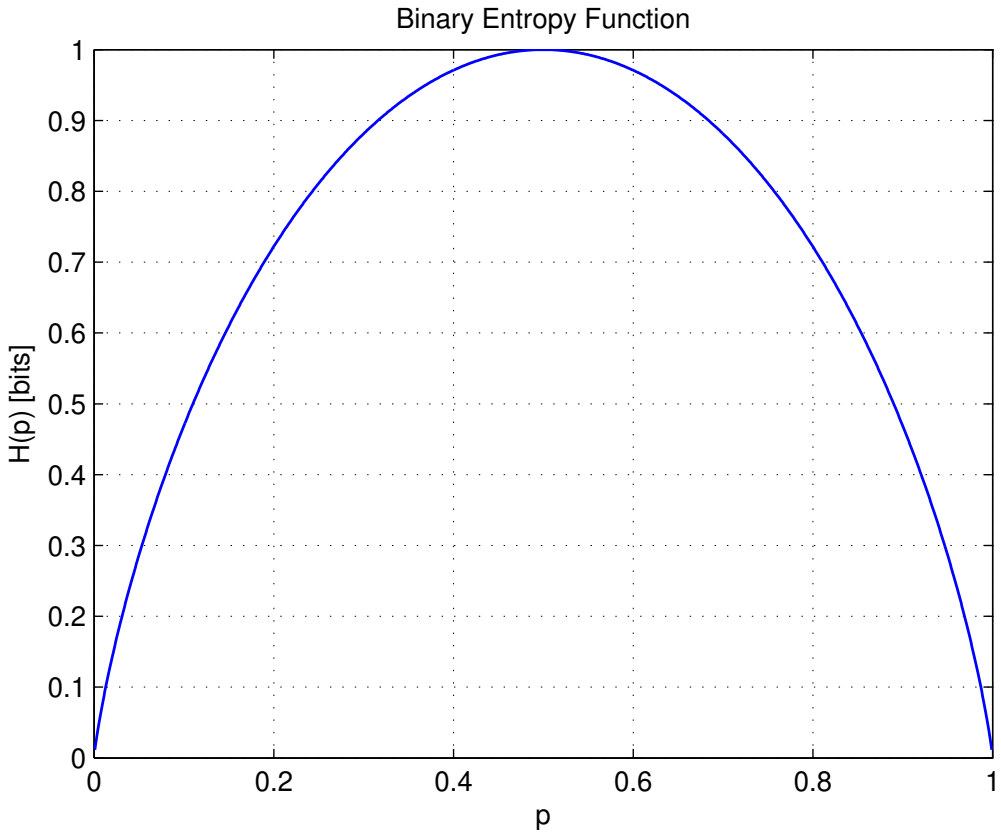
- proof at the end of this section

i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
	$\sum_i p_i \log_2 \frac{1}{p_i}$		4.1

Shannon information content of the 27 possible outcomes when a random character is picked from an English document

Binary entropy function

- Entropy of biased coin is $H(p) = p \log {}^{1}/_{p} + (1-p) \log {}^{1}/_{(1-p)}$.
- The function $H_2(p) = H(\{p, 1-p\})$ is called the binary entropy function. $H_2(0) = H_2(1) = 0$.
- The maximum value of $H_2(p)$ is $H_2\left(\frac{1}{2}\right) = 1$ bit.



Examples of entropy

- A “fair" coin toss (equiprobable, uniform) has entropy (uncertainty, information) 1 bit.
- One roll of a fair (unbiased) die: $H = \log_2 6 = 2.58496$ bits.
- The entropy of the sum of two dice is 3.2744 bits, which is less than $\log 11 = 3.4594$ bits because the outcomes are not equally probable.

Joint entropy

- In order to study information sources that are more complicated than simple random variables, such as random processes, we must introduce the notions of joint and conditional entropies of several random variables.
- Then we can define mutual information, which is the key quantity used in the Channel Coding Theorem.

Definition

The joint entropy of random variables X and Y with joint pdf $p(x, y)$ is

$$H(X, Y) = \sum_{x,y} p(x, y) \log \frac{1}{p(x, y)} = E_{p(x,y)} \left[\log \frac{1}{p(x, y)} \right]$$

- The entropy of two fair coin tosses is 2 bits.

Example

Let X and Y be dependent random variables with joint pdf shown in the table below.

X \ Y	0	1
0	$1/4$	$1/4$
1	$1/2$	0

$H(X,Y) = \frac{1}{4}\log\frac{1}{1/4} + \frac{1}{4}\log\frac{1}{1/4} + \frac{1}{2}\log\frac{1}{1/2} = \frac{2}{4}\log 4 + \frac{1}{2}\log 2 = \frac{2}{4}\cdot 2 + \frac{1}{2}\cdot 1 = \frac{3}{2}$bits

Fact

Suppose X and Y are independent $p(x,y) = p(x)\,p(y)$, sometimes denoted $X \perp Y$. Then:

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

Proof.

$$H(X,Y) = \sum_{x,y} p(x,y) \log \frac{1}{p(x,y)} = \sum_{x,y} p(x)\,p(y) \log \frac{1}{p(x)p(y)} =$$
$$\sum_{x,y} p(x)\,p(y) \log \frac{1}{p(x)} + \sum_{x,y} p(x)\,p(y) \log \frac{1}{p(y)} =$$
$$\sum_{y} p(y) \sum_{x} p(x) \log \frac{1}{p(x)} + \sum_{x} p(x) \sum_{y} p(y) \log \frac{1}{p(y)} =$$
$$\sum_{y} p(y)\,H(X) + \sum_{x} p(x)\,H(Y) = H(X) + H(Y)$$

□

This result can be obtained more concisely as follows:
$H(X,Y) = E[-\log p(x,y)] = E[-\log p(x)\,p(y)] =$
$E[-(\log p(x) + \log p(y))] = E[-\log p(x)] + E[-\log p(y)] =$
$H(X) + H(Y)$

Conditional entropy

Definition

$H(Y|X) = E_{p(x)}[H(Y|X=x)]$,
where $H(Y|X=x) = \sum_y p(y|x) \log \frac{1}{p(y|x)}$

hence:

$$H(Y|X) = \sum_x p(x) \sum_y p(y|x) \log \frac{1}{p(y|x)} =$$

$$\sum_{x,y} p(x)\, p(y|x) \log \frac{1}{p(y|x)} = \sum_{x,y} p(x,y) \log \frac{1}{p(y|x)} =$$

$$E_{p(x,y)}\left[\log \frac{1}{p(y|x)}\right]$$

Example

For the random variables of the previous example,

$p(x, y)$

X \ Y	0	1
0	$1/4$	$1/4$
1	$1/2$	0

$p(y|x)$

X \ Y	0	1
0	$1/2$	$1/2$
1	1	0

we calculate entropies in bits:

$H(Y|X=0) = 1,$

$H(Y|X=1) = 0 \Rightarrow H(Y|X) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0 = \frac{1}{2}$

Conditioning reduces entropy

note that:
$H(Y) = H_2\left(\frac{1}{4}\right) = 0.8113 > \frac{1}{2} = H(Y|X)$.

Fact

Conditioning reduces entropy:

$$H(Y|X) \leq H(Y)$$

When conditioning does not reduce entropy?

Consider $p(x,y) = p(x)\,p(y)$, i.e. $X \perp Y$

$$H(Y|X) = \sum_{x,y} p(x,y) \log \frac{1}{p(y|x)} = \sum_{x,y} p(x,y) \log \frac{p(x)}{p(y,x)} =$$

$$\sum_{x,y} p(x)\,p(y) \log \frac{p(x)}{p(x)p(y)} = \sum_{x} p(x) \sum_{y} p(y) \log \frac{1}{p(y)} = H(Y)$$

$$H(Y|X) = H(Y) \textit{ iff } X \perp Y$$

Chain rule

Theorem

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

Proof.

This follows from the chain rule for joint probability. Concise proof: □

$$H(X, Y) = E_{p(x,y)}\left[\log \frac{1}{p(x,y)}\right] = E_{p(x,y)}\left[\log \frac{1}{p(x)p(y|x)}\right] =$$

$$E_{p(x,y)}\left[\log \frac{1}{p(x)}\right] + E_{p(x,y)}\left[\log \frac{1}{p(y|x)}\right] = H(X) + H(Y|X)$$

Mutual information

- The definition of conditional entropy was not designed simply to guarantee that the chain rule $H(X, Y) = H(X) + H(Y|X)$ would hold.
- Conditional entropy is a physically meaningful quantity. On average, Y can be represented with $H(Y|X)$ bits once X is known.

Fact: $H(Y|X) \leq H(Y)$ with equality if and only if X and Y are independent.
The difference $H(Y) - H(Y|X)$ can be interpreted as how much information X gives about Y on average.

Definition

Mutual information:

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

We can manipulate this expression:

$$I(X;Y) = E_{p(y)}\left[\log \frac{1}{p(y)}\right] - E_{p(x,y)}\left[\log \frac{1}{p(y|x)}\right] =$$
$$E_{p(x,y)}\left[\log \frac{p(y|x)}{p(y)}\right] = E\left[\log \frac{p(x)p(y|x)}{p(x)p(y)}\right] = E\left[\log \frac{p(x,y)}{p(x)p(y)}\right]$$

hence the Mutual information is also given by:

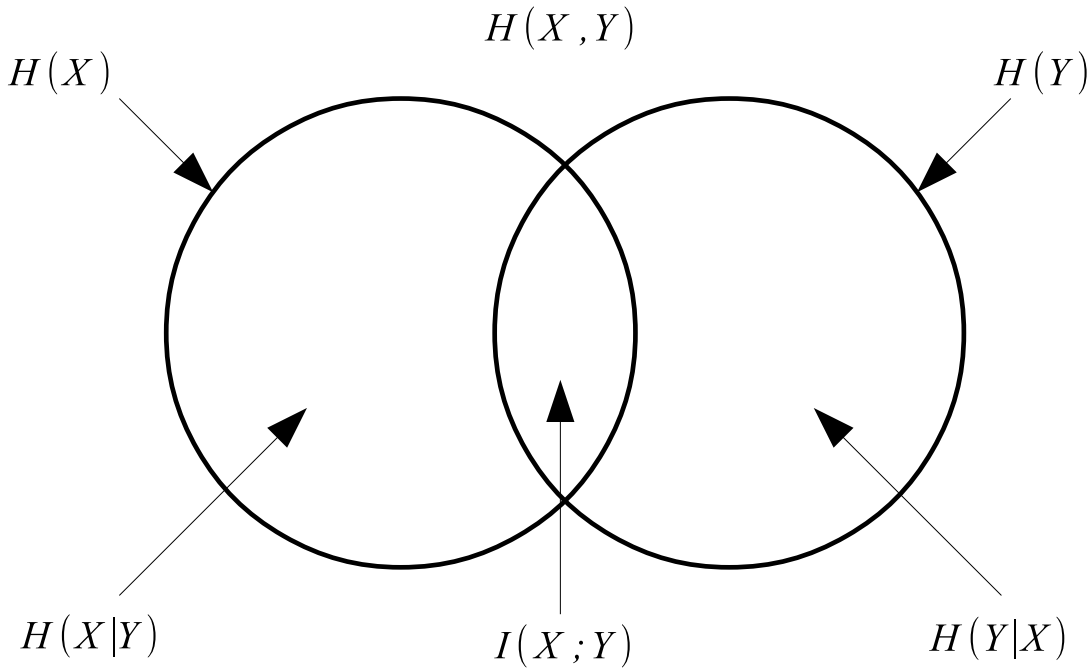
$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

This definition is symmetric in X and Y .
Other formulas:

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

Venn diagram

We can consider $I(X;Y)$ to be the information common to X and Y . This leads to the following Venn diagram and relations between entropy and mutual information:



$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$H(X) + H(Y) = I(X;Y) + H(X,Y)$$

Relative entropy

- Mutual information is a special case of *relative entropy*, also called *Kullback-Leibler* distance or *cross entropy*.

Definition

Let $p(x)$ and $q(x)$ be pdfs on a sample space. The relative entropy is

$$D\left(p(x)\,\|q(x)\right) = \sum_x p\left(x\right) \log \frac{p\left(x\right)}{q\left(x\right)} = E_{p(x)}\left[\log \frac{p\left(x\right)}{q\left(x\right)}\right]$$

- Conventions used in $D\left(p\,\|q\right)$ formula: $\log\left(\frac{0}{0}\right) = 0$, $0 \log\left(\frac{0}{q}\right) = 0$, $p \log\left(\frac{p}{0}\right) = +\infty$. It will be seen that $D\left(p\,\|q\right)$ is nonnegative and is a nonsymmetric distance measure.

Information Inequality

Our main application of Jensen's inequality ($E[f(X)] \geq f(E[X])$ for f convex) is the Information Inequality

Theorem

For any probability distribution functions $p(x)$ and $q(x)$

$$D(p \| q) \geq 0$$

with equality if and only if $p(x) \equiv q(x)$

If $p(x) > 0$ but $q(x) = 0$ then $D(p\|q) = +\infty$. And since terms in $D(p\|q)$ corresponding to $p(x) = 0$ are 0, we may assume that $p(x) > 0$ and $q(x) > 0$ for every x.
The trick is to express $D(p\|q)$ in terms of the convex function $-\log$:

Proof.

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \sum_x p(x) \left[-\log \frac{q(x)}{p(x)} \right] \geq$$

$$-\log \left(\sum_x p(x) \frac{q(x)}{p(x)} \right) = -\log \left(\sum_x q(x) \right) = -\log 1 = 0$$

□

Applications of the information inequality

Fact

$I(X;Y) \geq 0$; *and* $I(X;Y) = 0$ *iff* $X \perp Y$

Proof.

□

Using the elegant definition of mutual information:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} = D\left(p(x,y) \,\|\, p(x)\,p(y)\right) \geq 0$$

with equality if and only if $p(x,y) = p(x)\,p(y)$, that is, X and Y are independent.

Applications of the information inequality

Fact

Uniform distribution is Maximum Entropy

Proof.

□

Let $u(x) = \frac{1}{n}$ be the uniform distribution on the range $\mathcal{X}$ of X.

$0 \leq D\left(p(x) \| u(x)\right) = \sum_x p\left(x\right) \log \frac{p(x)}{1/n} =$

$\sum_x p\left(x\right) \log p(x) + \sum_x p\left(x\right) \log n = \log n - H(x)$

i.e.

$H(x) \leq \log n$, and $H(x) = \log n$ only for X being uniform.

Acknowledgments and further reading

- Additional material for this section can be found in the Cover textbook sections 2.1 – 2.6.