Information Theory and Networks

Lecture 6: Entropy and Mutual Information

Matthew Roughan
<matthew.roughan@adelaide.edu.au>
http://www.maths.adelaide.edu.au/matthew.roughan/Lecture_notes/InformationTheory/

School of Mathematical Sciences,
University of Adelaide

October 9, 2013

Part I

Entropy and Mutual Information

> Information, defined intuitively and informally, might be something like 'uncertainty's antidote.'
> *Brian Christian,*
> *The Most Human: What Talking with Computers Teaches Us About What It Means to Be Alive*

Section 1

Entropy: properties

Simple Properties

1. Axiomatic properties hold: e.g.,
   - $H(X) \geq 0$
   - $H(\cdot)$ is a function of probabilities, not the values of X.
2. $0 \leq H(X) \leq \log |\Omega|$
   - zero iff X is deterministic
   - $\log |\Omega|$ iff X is uniform (we'll prove this in a minute)
3. For a Bernoulli RV with $p = 1/2$, we have $H(p) = 1$ bit
   1. i.e., this defines the units of information
4. $H(X|Y) \neq H(Y|X)$

Simple Properties
Axiomatic properties hold: e.g., $H(X) \geq 0$; $H(\cdot)$ is a function of probabilities, not the values of X. $0 \leq H(X) \leq \log |\Omega|$; zero iff X is deterministic; $\log |\Omega|$ iff X is uniform (we'll prove this in a minute). For a Bernoulli RV with $p = 1/2$, we have $H(p) = 1$ bit; i.e., this defines the units of information. $H(X|Y) \neq H(Y|X)$

We can do entropy in any base, but unit change:

- base 2: units are bits
- base e: units are nats

Entropy Chain Rule

Theorem (Chain Rule)

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

Proof.

$$\begin{aligned} p(x,y) &= p(x)p(y|x) \\ \log p(x,y) &= \log p(x) + \log p(y|x) \\ E[\log p(x,y)] &= E[\log p(x)] + E[\log p(y|x)]. \end{aligned}$$

by linearity of expectations, and similarly for the second form. □

Entropy Chain Rule
Theorem (Chain Rule) $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$
Proof. $p(x,y) = p(x)p(y|x)$; $\log p(x,y) = \log p(x) + \log p(y|x)$; $E[\log p(x,y)] = E[\log p(x)] + E[\log p(y|x)]$.
by linearity of expectations, and similarly for the second form. □

Entropy Chain Rule: Corollaries

Theorem (Chain Rule Corollary)

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

Don't confuse with

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

Theorem (Chain Rule Corollary)

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

But remember that $H(X|Y) \neq H(Y|X)$ in general.

Entropy Chain Rule: Corollaries
Theorem (Chain Rule Corollary)
$H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$
Don't confuse with
$H(Y,\ X|Z) = H(X|Z) + H(Y|X,\ Z)$
Theorem (Chain Rule Corollary)
$H(X) - H(X|Y) = H(Y) - H(Y|X).$
But remember that $H(X|Y) \neq H(Y|X)$ in general.

Entropy Chain Rule: General form

Theorem (Chain Rule)

Let $X_1, X_2, \ldots, X_n$ have joint PMF $p(x_1, x_2, \ldots, x_n)$, then

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1).$$

Proof.

Just use repeated applications of the two-variable chain rule, or prove directly in the same manner as the two-variable rule. □

Example:

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1).$$

Entropy Chain Rule: General form
Theorem (Chain Rule)
Let $X_1, X_2, \ldots, X_n$ have joint PMF $p(x_1, x_2, \ldots, x_n)$, then
$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1).$
Proof.
Just use repeated applications of the two-variable chain rule, or prove directly in the same manner as the two-variable rule. □
Example:
$H(X_1, X_2, X_3) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1).$

Relative Entropy Chain Rule

Theorem (Chain Rule)

$$D\big(p(x,y)\big\|q(x,y)\big) = D\big(p(x)\big\|q(x)\big) - D\big(p(y|x)\big\|q(y|x)\big)$$

Proof.

Similar to previous two-variable proof. $\square$

Relative Entropy Chain Rule
Theorem (Chain Rule)
$D(p(x,y)\|q(x,y)) = D(p(x)\|q(x)) - D(p(y|x)\|q(y|x))$
Proof.
Similar to previous two-variable proof. $\square$

Relative Entropy Properties

Theorem

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

with equality only iff $p(x) = q(x)$ for all x.

Proof.

$$-D\big(p\big\|q\big) = E\left[-\log\frac{p(X)}{q(X)}\right] \leq -\log E\left[\frac{p(X)}{q(X)}\right],$$

by Jensen's inequality, as $-\log$ is strictly convex, and so equality arises only when p/q is a constant (in this case 1 when $p = q$ for all x). Next

$$-D\big(p\big\|q\big) \leq \log E\left[\frac{q(X)}{p(X)}\right] = \log\sum_x p(x)\frac{q(x)}{p(x)} = \log\sum_x q(x) = \log 1 = 0$$

$\square$

Relative Entropy Properties
Theorem
$D(p\|q) \geq 0$
with equality only iff $p(x) = q(x)$ for all x.
Proof.
$-D(p\|q) = E\left[-\log\frac{p(X)}{q(X)}\right] \leq -\log E\left[\frac{p(X)}{q(X)}\right],$
by Jensen's inequality, as $-\log$ is strictly convex, and so equality arises only when p/q is a constant (in this case 1 when $p = q$ for all x). Next
$-D(p\|q) \leq \log E\left[\frac{q(X)}{p(X)}\right] = \log\sum_x p(x)\frac{q(x)}{p(x)} = \log\sum_x q(x) = \log 1 = 0$ $\square$

We can prove the result even more directly using Gibbs' Inequality (or the related log-sum inequality).

Corollary

Theorem

$$H(X) \leq \log |\Omega|.$$

Proof.

Take distributions $p(x)$ and compare it to the uniform distribution $u(x) = 1/|\Omega|$:

$$\begin{aligned}
D(p\|u) &= \sum_x p(x) \log \frac{p(x)}{u(x)} \\
&= -\sum_x p(x) \log u(x) + \sum_x p(x) \log p(x) \\
&= -\log u \sum_x p(x) - H(X) \\
&= \log |\Omega| - H(X)
\end{aligned}$$

And we already know that $D(p\|u) \geq 0$. □

Convexity of relative entropy

Theorem

The relative entropy $D(p\|q)$ is a convex function of (p, q), i.e., for two pairs of distributions $(p^{(1)}, q^{(1)})$ and $(p^{(2)}, q^{(2)})$.

$$\begin{aligned}
&D\Big(\lambda p^{(1)} + (1-\lambda)p^{(2)} \Big\| \lambda q^{(1)} + (1-\lambda)q^{(2)}\Big) \\
&\leq \lambda D\big(p^{(1)} \big\| q^{(1)}\big) + (1-\lambda) D\big(p^{(2)} \big\| q^{(2)}\big)
\end{aligned}$$

for all $0 \leq \lambda \leq 1$.

Proof.

The proof is just another application of Jensen's (or Gibbs') inequality, but is a bit messy, so I leave it to the reader. □

Corollary: concavity of H

Theorem

The entropy $H(X) = H(p)$ is a concave function of p, i.e.,

$$H(\lambda p^{(1)} + (1-\lambda)p^{(2)}) \geq \lambda H(p^{(1)}) + (1-\lambda)H(p^{(2)}).$$

Proof.

As before

$$H(p) = \log|\Omega| - D(p\|u),$$

so the result follows directly from the convexity of D. □

Intuitively this means that if we mixed two random variables, i.e., we take a Bernoulli trial with probability λ, and use it to select either X_1 or X_2, the resulting uncertainty is larger than the weighted mixture of the two uncertainties (as you would expect, I hope)

Corollary: concavity of H
Theorem
The entropy $H(X) = H(p)$ is a concave function of p, i.e.,
$H(\lambda p^{(1)} + (1-\lambda)p^{(2)}) \geq \lambda H(p^{(1)}) + (1-\lambda)H(p^{(2)})$.
Proof.
As before
$H(p) = \log|\Omega| - D(p\|u)$,
so the result follows directly from the convexity of D.
Intuitively this means that if we mixed two random variables, i.e., we take a Bernoulli trial with probability λ, and use it to select either X_1 or X_2, the resulting uncertainty is larger than the weighted mixture of the two uncertainties (as you would expect, I hope)

Conditioning reduces entropy

As we might expect, conditioning on Y (i.e., saying we know Y) reduces the uncertainty about X, unless they are independent.

Theorem

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

with equality only when X and Y are independent.

Conditioning reduces entropy
As we might expect, conditioning on Y (i.e., saying we know Y) reduces the uncertainty about X, unless they are independent.
Theorem
$H(X|Y) \leq H(X)$,
with equality only when X and Y are independent.

Conditioning reduces entropy

Proof. Given $p(x,y)$ define $q(x,y) = p_X(x)p_Y(y)$, where $p_X(x)$ and $p_Y(y)$ are the marginal distributions of X and Y respectively. Now define

$$I(X;Y) = D\big(p(x,y) \| q(x,y)\big) = E\left[\log \frac{p(X|Y)}{p_X(X)}\right],$$

By definition of conditional probabilities

$$E\left[\log \frac{p(X,Y)}{p_X(X)p_Y(Y)}\right] = E\left[\log \frac{p(X|Y)}{p_X(X)}\right] = E\left[\log p(X|Y)\right] - E\left[\log p_X(X)\right],$$

So

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

but we also know that $I(X;Y)$ is defined in terms of relative entropy, and hence $I(X;Y) \geq 0$, and hence the result. □

Conditioning reduces entropy
Proof. Given $p(x,y)$ define $q(x,y) = p_X(x)p_Y(y)$, where $p_X(x)$ and $p_Y(y)$ are the marginal distributions of X and Y respectively. Now define $I(X;Y) = D\big(p(x,y) \| q(x,y)\big) = E\left[\log \frac{p(X|Y)}{p_X(X)}\right]$, By definition of conditional probabilities $E\left[\log \frac{p(X,Y)}{p_X(X)p_Y(Y)}\right] = E\left[\log \frac{p(X|Y)}{p_X(X)}\right] = E\left[\log p(X|Y)\right] - E\left[\log p_X(X)\right]$, So $I(X;Y) = -H(X|Y) + H(X)$, but we also know that $I(X;Y)$ is defined in terms of relative entropy, and hence $I(X;Y) \geq 0$, and hence the result.

The quantity $I(X;Y)$ is called the mutual information, and we will get to that in a moment. In particular, we'll use the result

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

again so keep you eye on it.

Section 2

Mutual information

Section 2
Mutual information

Motivation

- We created an "information" metric before, based on a single probability, but found that entropy was a more useful idea.
- Now lets return to trying to say something useful about information
- The mutual information is a measure of the information that we learn about one random variable from another.

Motivation

- We created an "information" metric before, based on a single probability, but found that entropy was a more useful idea.
- Now lets return to trying to say something useful about information
- The mutual information is a measure of the information that we learn about one random variable from another.

Mutual Information

Define: mutual information

$$
\begin{aligned}
I(X;Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p_X(x) p_Y(y)} \\
&= D\big(p(x,y) \| q(x,y)\big) \\
&= E\left[\log \frac{p(X|Y)}{p(X)}\right],
\end{aligned}
$$

Mutual Information

Define: mutual information

$$
\begin{aligned}
I(X;Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p_X(x) p_Y(y)} \\
&= D\big(p(x,y) \| q(x,y)\big) \\
&= E\left[\log \frac{p(X|Y)}{p(X)}\right],
\end{aligned}
$$

Relationship between entropy and mutual information

We already showed that

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

- So the mutual information is the reduction in uncertainty in X given knowledge of Y.
- By symmetry
$$I(X;Y) = H(Y) - H(Y|X).$$
- Also the "self-information"
$$I(X;X) = H(X) - H(X|X) = H(X).$$

which is the idea we started with, that information and uncertainty about a random variable are really the same.

Relationship between entropy and mutual information

We already showed that
$$I(X;Y) = H(X) - H(X|Y).$$
- So the mutual information is the reduction in uncertainty in X given knowledge of Y.
- By symmetry
$$I(X;Y) = H(Y) - H(Y|X).$$
- Also the "self-information"
$$I(X;X) = H(X) - H(X|X) = H(X).$$
which is the idea we started with, that information and uncertainty about a random variable are really the same.

Mutual Information Properties

- Mutual Information is non-negative, and is zero, iff X and Y are independent (see proof of previous theorem)
- Mutual Information has a conditional form (see [CT91, p.22] for details.)
- Mutual Information has a chain rule (see [CT91, p.22] for details.)

Mutual Information Properties
- Mutual Information is non-negative, and is zero, iff X and Y are independent (see proof of previous theorem)
- Mutual Information has a conditional form (see [CT91, p.22] for details.)
- Mutual Information has a chain rule (see [CT91, p.22] for details.)

Assignment

There are lots of practice problems in [CT91, Chapter 1], which is available in electronic form in our Library. I recommend you have a go, but I won't mark these.

The assignment is to calculate the entropy of Morse code symbols, given standard frequencies of English letters.
Hints:

- Remember Morse code really has four symbols:
  - dot
  - dash
  - letter-break
  - word-break
- Model the frequencies of word-breaks as well as just letters.
  - you may need to make your own measurements of text – lots is available, e.g., at http://www.gutenberg.org/

Assignment
There are lots of practice problems in [CT91, Chapter 1], which is available in electronic form in our Library. I recommend you have a go, but I won't mark these.
The assignment is to calculate the entropy of Morse code symbols, given standard frequencies of English letters.
Hints:
- Remember Morse code really has four symbols:
  - dot
  - dash
  - letter-break
  - word-break
- Model the frequencies of word-breaks as well as just letters.
  - you may need to make your own measurements of text – lots is available, e.g., at http://www.gutenberg.org/

Reading in text can be done in Matlab, but you may find some easier approaches. For instance, in R `http://www.r-bloggers.com/text-mining-the-complete-works-of-william-shakespeare/`
I personally prefer to use Perl for tasks like this, and I can pretty much guarantee that a Perl implementation will be faster to run, and faster to write (if you learn a bit about Perl), but its up to you.

Further reading I

Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, John Wiley and Sons, 1991.