Properties of Entropy and Mutual information

Who? Minqi Pan

From? Capital Normal University

When? October 18, 2011

A tiny Citation

My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann(?), he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage.'

—Claude Shannon[1]

[1] Seen on Scientific American 1971 , volume 225 , page 180

What we're gonna cover

Today

Self-information and Mutual-information

Properties of Mutual information

Properties of Entropy

Do you still remember

the Definition of Self-information.
You have already learnt this in the previous class. (?)

Do you still remember

the Definition of Self-information.
You have already learnt this in the previous class. (?)

- A measure of the information content associated with the outcome of a random variable.

Do you still remember

the Definition of Self-information.
You have already learnt this in the previous class. (?)

- A measure of the information content associated with the outcome of a random variable.
- Serves as a measure of the information content associated with the outcome of a random variable.

Do you still remember

the Definition of Self-information.
You have already learnt this in the previous class. (?)

- A measure of the information content associated with the outcome of a random variable.
- Serves as a measure of the information content associated with the outcome of a random variable.
- expressed in a unit of information – bits, nats, or hartleys – depending on the base of the logarithm used in its calculation.

Did anyone answer this?

Precisely, the Definition of Self-information is

Did anyone answer this?

Precisely, the Definition of Self-information is

$$I(\omega_n) = \log\left(\frac{1}{P(\omega_n)}\right) = -\log(P(\omega_n))$$

Propose a better name

I prefer calling it surprisal[2].

[2] seen in the 1961 book Thermostatics and Thermodynamics by Tribus.

Propose a better name

I prefer calling it surprisal[2].

- ... as it represents the "surprise" of seeing the outcome (a highly improbable outcome is very surprising).

[2]seen in the 1961 book Thermostatics and Thermodynamics by Tribus.

next thing

"Mutual-information"

next thing

"Mutual-information"

- a quantity that measures the mutual dependence of the two random variables.

next thing

"Mutual-information"

- a quantity that measures the mutual dependence of the two random variables.
- A big formula is gonna come. Behold!

next thing

"Mutual-information"

- a quantity that measures the mutual dependence of the two random variables.
- A big formula is gonna come. Behold!
- $$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\,p_2(y)}\right) dx\,dy,$$

where $p(x,y)$is now the joint probability density function ofX and$Y$, and $p1(x)$ and $p2(y)$ are the marginal probability density functions of X and Yrespectively.

next thing

"Mutual-information"

- a quantity that measures the mutual dependence of the two random variables.
- A big formula is gonna come. Behold!
- $$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\,p_2(y)}\right) dx\,dy,$$

  where $p(x,y)$is now the joint probability density function ofX and$Y$, and $p1(x)$ and $p2(y)$ are the marginal probability density functions of X and Yrespectively.
- Yipes!

Mutual-information

Don't worry, we'll only consider discrete cases.

Mutual-information

Don't worry, we'll only consider discrete cases.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\,p_2(y)}\right),$$

where $p(x, y)$ is the joint probability distribution function of $X$and Y, and $p1(x)$and $p2(y)$ are the marginal probability distribution functions of X and Y respectively.

Intuitively

truly and utterly Intuitively speaking,

Intuitively

truly and utterly Intuitively speaking,

- mutual information measures the information that X and Y share.

Intuitively

truly and utterly Intuitively speaking,

- mutual information measures the information that X and Y share.
- it measures how much knowing one of these variables reduces uncertainty about the other.

If...

if X and Y are independent.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right),$$

If...

if X and Y are independent.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\,p_2(y)}\right),$$

$$p(x,y) = p(x)p(y)$$

If...

if X and Y are independent.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right),$$

$$p(x,y) = p(x)p(y)$$

$$\log \left(\frac{p(x,y)}{p(x)\, p(y)} \right) = \log 1 = 0.$$

If...

if X and Y are independent.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\,p_2(y)}\right),$$

$$p(x,y) = p(x)p(y)$$

$$\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) = \log 1 = 0.$$

their mutual information is zero.

If...(other extreme)

if X and Y are identical

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right),$$

If...(other extreme)

if X and Y are identical

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\, p_2(y)}\right),$$

$$I(X;Y) = \sum_{y \in Y} (p(y,y) = p_1(y)) \log\left(\frac{p(y,y) = p_1(y)}{p_1(y)\, p_1(y)}\right),$$

$$I(X;Y) = \sum_{y \in Y} p_1(y) \log\left(\frac{1}{p_1(y)}\right) = H(X) = H(Y)$$

If...(other extreme)

if X and Y are identical

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right),$$

-

$$I(X;Y) = \sum_{y \in Y} (p(y,y) = p_1(y)) \log \left(\frac{p(y,y) = p_1(y)}{p_1(y)\, p_1(y)} \right),$$

$$I(X;Y) = \sum_{y \in Y} p_1(y) \log \left(\frac{1}{p_1(y)} \right) = H(X) = H(Y)$$

- then all information conveyed by X is shared with Y. Knowing X determines the value of Y and vice versa.

If...(other extreme)

if X and Y are identical

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right),$$

- $$I(X;Y) = \sum_{y \in Y} (p(y,y) = p_1(y)) \log \left(\frac{p(y,y) = p_1(y)}{p_1(y)\, p_1(y)} \right),$$

$$I(X;Y) = \sum_{y \in Y} p_1(y) \log \left(\frac{1}{p_1(y)} \right) = H(X) = H(Y)$$

- then all information conveyed by X is shared with Y. Knowing X determines the value of Y and vice versa.
- Ah, do you still remember $H()$?

$H()$

You have learnt, presumably in previous classes, that ...

$H()$

You have learnt, presumably in previous classes, that ...

- let 'message' be a specific realization of the random variable.

$H()$

You have learnt, presumably in previous classes, that ...

- let 'message' be a specific realization of the random variable.
- Shannon's entropy quantifies the expected value of the information contained in a message

$H()$

You have learnt, presumably in previous classes, that ...

- let 'message' be a specific realization of the random variable.
- Shannon's entropy quantifies the expected value of the information contained in a message
- Shannon's entropy is a measure of the uncertainty associated with a random variable

$H()$

You have learnt, presumably in previous classes, that ...

- let 'message' be a specific realization of the random variable.
- Shannon's entropy quantifies the expected value of the information contained in a message
- Shannon's entropy is a measure of the uncertainty associated with a random variable
- Shannon's entropy usually comes in units bits.

Definitions

Just another review...

Definitions

Just another review...

$$H(X) = \mathrm{E}(I(X)).$$

E is the expected value, and I is the information content of X.

Definitions

Just another review...

- $$H(X) = \mathrm{E}(I(X)).$$

  E is the expected value, and I is the information content of X.
- If p denotes the probability mass function of X then the entropy can explicitly be written as

$$H(X) = \sum_{i=1}^{n} p(x_i)\, I(x_i)$$

$$= \sum_{i=1}^{n} p(x_i) \log_b \frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

$H()$

I prefer calling it the expected surprisal!!

In a word

the expected surprisal is to...

In a word

the expected surprisal is to...

- measure disorder, measure unpredictability.

In a word

the expected surprisal is to...

- measure disorder, measure unpredictability.
- measure the average information content one is missing when one does not know the value of the random variable.

In a word

the expected surprisal is to...

- measure disorder, measure unpredictability.
- measure the average information content one is missing when one does not know the value of the random variable.
- What’s your understanding?

In a word

the expected surprisal is to...

- measure disorder, measure unpredictability.
- measure the average information content one is missing when one does not know the value of the random variable.
- What's your understanding?
- But it's only a number anyway. An artificial Mathematical Quantity.

Connecting the Dots 1

$$I(X;X) = H(X)$$

where I(X;X) is the mutual information of X with itself.

Connecting the Dots 1

$$I(X;X) = H(X)$$

where I(X;X) is the mutual information of X with itself.

Right?

Connecting the Dots 1

$$I(X;X) = H(X)$$

where I(X;X) is the mutual information of X with itself.

- Right?
- We have just shown that!

Connecting the Dots 2

In fact, (Property 3)

Connecting the Dots 2

In fact, (Property 3)

- $I(X;Y) \leq H(X)$

Connecting the Dots 2

In fact, (Property 3)

- $I(X;Y) \leq H(X)$
- $I(X;Y) \leq H(Y)$

Connecting the Dots 2

In fact, (Property 3)

- $I(X;Y) \leq H(X)$
- $I(X;Y) \leq H(Y)$
- $I(X;Y) = H(X) \Rightarrow$ implication is possible from one event to another (Channel Fully Operational!)

Connecting the Dots 2

In fact, (Property 3)

- $I(X;Y) \leq H(X)$
- $I(X;Y) \leq H(Y)$
- $I(X;Y) = H(X) \Rightarrow$ implication is possible from one event to another (Channel Fully Operational!)
- $I(X;Y) = 0 \Rightarrow$ implication is impossible from one event to another (Channel Break!)

Connecting the Dots 3

Other Trivial Properties

Connecting the Dots 3

Other Trivial Properties

- $I(X;Y) \neq 0$

Connecting the Dots 3

Other Trivial Properties

- $I(X;Y) \neq 0$
- $I(X;Y) = I(Y;X)$

Connecting the Dots 4

Nontrivial Property 4

Connecting the Dots 4

Nontrivial Property 4

- convex function

You check the book, Page 32, for proofs
I won't prove it here.

Now back to properties of Entropy

Recall

$$H(X) = \sum_{i=1}^{n} p(x_i)\, I(x_i)$$

$$= \sum_{i=1}^{n} p(x_i) \log_b \frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

First I'll talk about additivity,

Now back to properties of Entropy

Recall

$$H(X) = \sum_{i=1}^{n} p(x_i)\, I(x_i)$$

$$= \sum_{i=1}^{n} p(x_i) \log_b \frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

First I'll talk about additivity,

Why do they use logarithm to construct that formula?

Now back to properties of Entropy

Recall

$$H(X) = \sum_{i=1}^{n} p(x_i)\, I(x_i)$$

$$= \sum_{i=1}^{n} p(x_i) \log_b \frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

First I'll talk about additivity,

- Why do they use logarithm to construct that formula?
- The logarithm is used so as to provide the additivity characteristic for independent uncertainty.

Properties of Entropy

Examples about additivity,

Properties of Entropy

Examples about additivity,

- When throwing a fair dice, the probability of ’four’ is 1/6. When it is proclaimed that ’four’ has been thrown, the amount of self-information is I(’four’) = log2 (1/(1/6)) = log2 (6) = 2.585 bits.

Properties of Entropy

Examples about additivity,

- When throwing a fair dice, the probability of ’four’ is 1/6. When it is proclaimed that ’four’ has been thrown, the amount of self-information is I(’four’) = log2 (1/(1/6)) = log2 (6) = 2.585 bits.
- When, independently, two dice are thrown, the amount of information associated with throw 1 = ’two’ and throw 2 = ’four’ equals I(’throw 1 is two and throw 2 is four’) = log2 (1/P(throw 1 = ’two’ and throw 2 = ’four’)) = log2 (1/(1/36)) = log2 (36) = 5.170 bits.
  This outcome equals the sum of the individual amounts of self-information associated with throw 1 = ’two’ and throw 2 = ’four’; namely 2.585 + 2.585 = 5.170 bits.

Properties of Entropy

Prop 5. certainties

Properties of Entropy

Prop 5. certainties

I'll explain by examples

Properties of Entropy

Prop 5. certainties

- I'll explain by examples
- A series of tosses of a two-headed coin will have zero entropy,

Properties of Entropy

Prop 5. certainties

- I'll explain by examples
- A series of tosses of a two-headed coin will have zero entropy,
- since the outcomes are entirely predictable.

Properties of Entropy

Basic Properties

Properties of Entropy

Basic Properties

- Prop 1. Entropy is always non-negative.

Properties of Entropy

Basic Properties

- Prop 1. Entropy is always non-negative.
- Prop 2. Symmetry
  The measure should be unchanged if the outcomes

$$x_i$$

are re-ordered.

$$H_n\left(p_1, p_2, \ldots\right) = H_n\left(p_2, p_1, \ldots\right) = ...$$

Properties of Entropy

The Quest of Maximum

Properties of Entropy

The Quest of Maximum

- The measure should be maximal if all the outcomes are equally likely

Properties of Entropy

The Quest of Maximum

- The measure should be maximal if all the outcomes are equally likely
- (uncertainty is highest when all possible events are equiprobable).

Properties of Entropy

The Quest of Maximum

- The measure should be maximal if all the outcomes are equally likely
- (uncertainty is highest when all possible events are equiprobable).
- $$H_n(p_1, \ldots, p_n) \le H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right).$$

Properties of Entropy

The Quest of Maximum

- The measure should be maximal if all the outcomes are equally likely
- (uncertainty is highest when all possible events are equiprobable).
- $$H_n(p_1, \ldots, p_n) \leq H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right).$$
- For equiprobable events the entropy should increase with the number of outcomes.

Properties of Entropy

The Quest of Maximum

- The measure should be maximal if all the outcomes are equally likely
- (uncertainty is highest when all possible events are equiprobable).
- $$H_n(p_1, \ldots, p_n) \leq H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right).$$
- For equiprobable events the entropy should increase with the number of outcomes.
- $$H_n\underbrace{\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)}_{n} < H_{n+1}\underbrace{\left(\frac{1}{n+1}, \ldots, \frac{1}{n+1}\right)}_{n+1}.$$

Properties of Entropy

Adding 0's

Properties of Entropy

Adding 0's

- Adding or removing an event with probability zero does not contribute to the entropy:

Properties of Entropy

Adding 0's

- Adding or removing an event with probability zero does not contribute to the entropy:

$$H_{n+1}(p_1, \ldots, p_n, 0) = H_n(p_1, \ldots, p_n).$$

Properties of Entropy

Jensen inequality

Properties of Entropy

Jensen inequality

- It can be confirmed using the Jensen inequality that

Properties of Entropy

Jensen inequality

- It can be confirmed using the Jensen inequality that

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

$$\leq \log_b\left[\mathrm{E}\left(\frac{1}{p(X)}\right)\right] = \log_b(n).$$

Properties of Entropy

Conditional Entropy is defined by

Properties of Entropy

Conditional Entropy is defined by

$$H(X|Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)}$$

Properties of Entropy

Conditional Entropy is defined by

- $$H(X|Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)}$$

- where p(xi,yj) is the probability that$X = x_i$ and $Y = y_j$.

Properties of Entropy

Conditional Entropy is defined by

- $$H(X|Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)}$$
- where p(xi,yj) is the probability that$X = x_i$ and $Y = y_j$.
- This quantity should be understood as the amount of randomness in the random variable X given that you know the value of Y. For example, the entropy associated with a six-sided die is H(die), but if you were told that it had in fact landed on 1, 2, or 3, then its entropy would be equal to H(die: the die landed on 1, 2, or 3).

Properties of Entropy

Conditionally

Properties of Entropy

Conditionally

- If X and Y are two independent experiments, then knowing the value of Y doesn't influence our knowledge of the value of X (since the two don't influence each other by independence):

Properties of Entropy

Conditionally

- If X and Y are two independent experiments, then knowing the value of Y doesn't influence our knowledge of the value of X (since the two don't influence each other by independence):
-

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

Properties of Entropy

Conditionally inequality

Properties of Entropy

Conditionally inequality

- Oh knowing more reduces the expected value of surprisal.

Properties of Entropy

Conditionally inequality

- Oh knowing more reduces the expected value of surprisal.
- $$H(X|Y) \leq H(X).$$

Example on Page 28

Calculations that wraps up.

That's all

Many thanks goes to Shannon for his brilliance.

That’s all

Many thanks goes to Shannon for his brilliance.

The other thanks goes to you:)

That's all

Many thanks goes to Shannon for his brilliance.

- The other thanks goes to you:)
- Ciao!