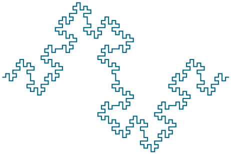
An Introduction to Information Theory

Carlton Downey



November 12, 2013

Introduction

- Today's recitation will be an introduction to *Information Theory*
- Information theory studies the quantification of Information
  - Compression
  - Transmission
  - Error Correction
  - Gambling
- Founded by Claude Shannon in 1948 by his classic paper "A Mathematical Theory of Communication"
- It is an area of mathematics which I think is particularly elegant

Outline

Outline

Motivation: Casino

- You're at a casino
- You can bet on coins, dice, or roulette
  - Coins = 2 possible outcomes. Pays 2:1
  - Dice = 6 possible outcomes. Pays 6:1
  - roulette = 36 possible outcomes. Pays 36:1
- Suppose you can predict the outcome of a single coin toss/dice roll/roulette spin.
- Which would you choose?

Motivation: Casino

- You're at a casino
- You can bet on coins, dice, or roulette
  - Coins = 2 possible outcomes. Pays 2:1
  - Dice = 6 possible outcomes. Pays 6:1
  - roulette = 36 possible outcomes. Pays 36:1
- Suppose you can predict the outcome of a single coin toss/dice roll/roulette spin.
- Which would you choose?
- Roulette. But why? these are all fair games

Motivation: Casino

- You're at a casino
- You can bet on coins, dice, or roulette
  - Coins = 2 possible outcomes. Pays 2:1
  - Dice = 6 possible outcomes. Pays 6:1
  - roulette = 36 possible outcomes. Pays 36:1
- Suppose you can predict the outcome of a single coin toss/dice roll/roulette spin.
- Which would you choose?
- Roulette. But why? these are all fair games
- Answer: Roulette provides us with the most *Information*

Motivation: Coin Toss

- Consider two coins:
  - Fair coin C_F with $P(H) = 0.5, P(T) = 0.5$
  - Bent coin C_B with $P(H) = 0.99, P(T) = 0.01$
- Suppose we flip both coins, and they both land heads
- Intuitively we are more "surprised" or "Informed" by first outcome.
- We know C_B is almost certain to land heads, so the knowledge that it lands heads provides us with very little information.

Motivation: Compression

- Suppose we observe a sequence of events:
  - Coin tosses
  - Words in a language
  - notes in a song
  - etc.
- We want to record the sequence of events in the smallest possible space.
- In other words we want the shortest representation which preserves all information.
- Another way to think about this: How much information does the sequence of events actually contain?

Motivation: Compression

To be concrete, consider the problem of recording coin tosses in unary.

$$T, T, T, T, H$$

Approach 1:

H	T
0	00

$$00, 00, 00, 00, 0$$

We used 9 characters

Motivation: Compression

To be concrete, consider the problem of recording coin tosses in unary.

$$T, T, T, T, H$$

Approach 2:

H	T
00	0

$$0, 0, 0, 0, 00$$

We used 6 characters

Motivation: Compression

- Frequently occuring events should have short encodings
- We see this in english with words such as “a”, “the”, “and”, etc.
- We want to maximise the information-per-character
- seeing common events provides little information
- seeing uncommon events provides a lot of information

Outline

Information

- Let X be a random variable with distribution $p(X)$.
- We want to quantify the information provided by each possible outcome.
- Specifically we want a function which maps the probability of an event $p(x)$ to the information $I(x)$
- Our metric $I(x)$ should have the following properties:
  - $I(x_i) \geq 0 \quad \forall i.$
  - $I(x_1) > I(x_2)$ if $p(x_1) < p(x_2)$
  - $I(x_1, x_2) = I(x_1) + I(x_2)$

Information

$$I(x) = f(p(x))$$

- We want $f()$ such that $I(x_1, x_2) = I(x_1) + I(x_2)$
- We know $p(x_1, x_2) = p(x_1)p(x_2)$
- The only function with this property is log(): $\log(ab) = \log(a) + \log(b)$
- Hence we define:

$$I(X) = \log(\frac{1}{p(x)})$$

Information: Coin

Fair Coin:

h	t
0.5	0.5

$$
\begin{aligned}
I(h) &= \log(\frac{1}{0.5}) = \log(2) = 1 \\
I(t) &= \log(\frac{1}{0.5}) = \log(2) = 1
\end{aligned}
$$

INFORMATION: COIN

Bent Coin:

h	t
0.25	0.75

$$I(h) = \log(\frac{1}{0.25}) = \log(4) = 2$$
$$I(t) = \log(\frac{1}{0.75}) = \log(1.33) = 0.42$$

Information: Coin

Really Bent Coin:

h	t
0.01	0.99

$$I(h) = \log(\frac{1}{0.01}) = \log(100) = 6.65$$
$$I(t) = \log(\frac{1}{0.99}) = \log(1.01) = 0.01$$

Information: Two Events

Question: How much information do we get from observing two events?

$$
\begin{aligned}
I(x_1, x_2) &= \log(\frac{1}{p(x_1, x_2)}) \\
&= \log(\frac{1}{p(x_1)p(x_2)}) \\
&= \log(\frac{1}{p(x_1)}\frac{1}{p(x_2)}) \\
&= \log(\frac{1}{p(x_1)}) + \log(\frac{1}{p(x_2)}) \\
&= I(x_1) + I(x_2)
\end{aligned}
$$

Answer: Information sums!

Information is Additive

- I(k fair coin tosses) = $\log \frac{1}{1/2^k} = k$ bits
- So:
  - Random word from a 100,000 word vocabulary:
    I(word) = $\log(100,000)$ = 16.61 bits
  - A 1000 word document from same source:
    I(documents) = 16,610 bits
  - A 480 pixel, 16-greyscale video picture:
    I(picture) = $307,200 \times \log(16)$ = 1,228,800 bits
- A picture is worth (a lot more than) 1000 words!
- In reality this is a gross overestimate

Information: Two Coins

Bent Coin:

x	h	t
p(x)	0.25	0.75
I(x)	2	0.42

$$
\begin{aligned}
I(hh) &= I(h) + I(h) = 4 \\
I(ht) &= I(h) + I(t) = 2.42 \\
I(th) &= I(t) + I(h) = 2.42 \\
I(th) &= I(t) + I(t) = 0.84
\end{aligned}
$$

Information: Two Coins

Bent Coin Twice:

hh	ht	th	tt
0.0625	0.1875	0.1875	0.5625

$$
\begin{aligned}
I(hh) &= \log(\frac{1}{0.0625}) = \log(4) = 4 \\
I(ht) &= \log(\frac{1}{0.1875}) = \log(4) = 2.42 \\
I(th) &= \log(\frac{1}{0.1875}) = \log(4) = 2.42 \\
I(tt) &= \log(\frac{1}{0.5625}) = \log(4) = 0.84
\end{aligned}
$$

Outline

Entropy

- Suppose we have a sequence of observations of a random variable X.
- A natural question to ask is what is the average amount of information per observation.
- This quantitity is called the *Entropy* and denoted $H(X)$

$$H(X) = E[I(X)] = E[\log(\frac{1}{p(X)})]$$

ENTROPY

- Information is associated with an *event* - heads, tails, etc.
- Entropy is associated with a *distribution* over events - p(x).

Entropy: Coin

Fair Coin:

x	h	t
p(x)	0.5	0.5
I(x)	1	1

$$\begin{aligned} H(X) &= E[I(X)] \\ &= \sum_i p(x_i)I(X) \\ &= p(h)I(h) + p(t)I(t) \\ &= 0.5 \times 1 + 0.5 \times 1 \\ &= 1 \end{aligned}$$

Entropy: Coin

Bent Coin:

x	h	t
p(x)	0.25	0.75
I(x)	2	0.42

$$
\begin{aligned}
H(X) &= E[I(X)] \\
&= \sum_i p(x_i) I(X) \\
&= p(h)I(h) + p(t)I(t) \\
&= 0.25 \times 2 + 0.75 \times 0.42 \\
&= 0.85
\end{aligned}
$$

Entropy: Coin

Very Bent Coin:

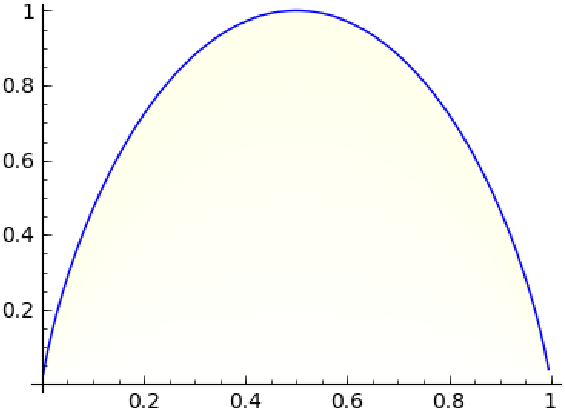
x	h	t
p(x)	0.01	0.99
I(x)	6.65	0.01

$$
\begin{aligned}
H(X) &= E[I(X)] \\
&= \sum_i p(x_i) I(X) \\
&= p(h)I(h) + p(t)I(t) \\
&= 0.01 \times 6.65 + 0.99 \times 0.01 \\
&= 0.08
\end{aligned}
$$

Entropy: All coins

Entropy: All coins

$$H(P) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$$



Alternative Explanations of Entropy

$$H(S) = \sum_i p_i \log \frac{1}{p_i}$$

- Average amount of information provided per event
- Average amount of surprise when observing a event
- Uncertainty an observer has before seeing the event
- Average number of bits needed to communicate each event

The Entropy of English

27 Characters (A-Z, space)
100,000 words (average 5.5 characters each)

- Assuming independence between successive *characters*:
  - Uniform character distribution: $\log(27) = 4.75$ bits/characters
  - True character distribution: 4.03 bits/character
- Assuming independence between successive *words*:
  - Uniform word distribution: $\frac{\log(100{,}000)}{6.5} = 2.55$ bits/character
  - True word distribution: $\frac{9.45}{6.5} = 1.45$ bits/character
- True Entropy of English is much lower

Types of Entropy

- There are 3 Types of Entropy
  - Marginal Entropy
  - Joint Entropy
  - Conditional Entropy
- We will now define these quantities, and study how they are related.

Marginal Entropy

- A single random variable X has a *Marginal Distribution*

$$p(X)$$

- This distribution has an associated *Marginal Entropy*

$$H(X) = \sum_i p(x_i) \log \frac{1}{p(x_i)}$$

- Marginal entropy is the average information provided by observing a variable X

JOINT ENTROPY

- Two or more random variables X, Y have a *Joint Distribution*

$$p(X, Y)$$

- This distribution has an associated *Joint Entropy*

$$H(X,Y) = \sum_i \sum_j p(x_i, y_j) \log \frac{1}{p(x_i, y_j)}$$

- Marginal entropy is the average *total* information provided by observing two variables X, Y

Conditional Entropy

- Two random variables X, Y also have two *Conditional Distributions*

$$p(X|Y) \quad \text{and} \quad P(Y|X)$$

- These distributions have associated *Conditional Entropys*

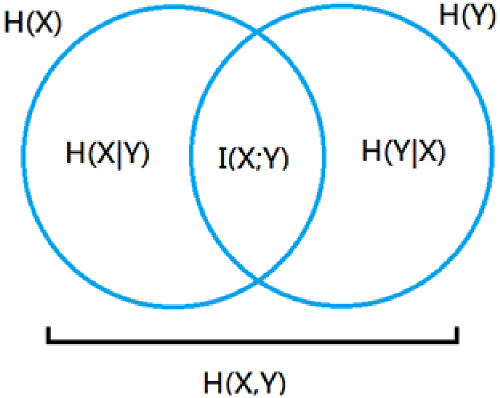
$$\begin{aligned} H(X|Y) &= \sum_j p(y_j) H(X|y_j) \\ &= \sum_j p(y_j) \sum_i p(x_i|y_j) \log \frac{1}{p(x_i|y_j)} \\ &= \sum_i \sum_j p(x_i, y_j) \log \frac{1}{p(x_i|y_j)} \end{aligned}$$

- Conditional entropy is the average *additional* information provided by observing X, given we already observed Y

Types of Entropy: Summary

- **Entropy:** Average information gained by observing a single variable
- **Joint Entropy:** Average *total* information gained by observing two or more variables
- **Conditional Entropy:** Average *additional* information gained by observing a new variable

Entropy Relationships



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

$$
\begin{aligned}
H(X,Y) &= \sum_{i,j} p(x_i, y_j) \log\left(\frac{1}{p(x_i, y_j)}\right) \\
&= \sum_{i,j} p(x_i, y_j) \log\left(\frac{1}{p(y_j|x_i)p(x_i)}\right) \\
&= \sum_{i,j} p(x_i, y_j) \left[\log\left(\frac{1}{p(x_i)}\right) + \log\left(\frac{1}{p(y_j|x_i)}\right)\right] \\
&= \sum_{i} p(x_i) \log\left(\frac{1}{p(x_i)}\right) + \sum_{i} p(x_i, y_j) \log\left(\frac{1}{p(y_j|x_i)}\right) \\
&= H(X) + H(Y|X)
\end{aligned}
$$

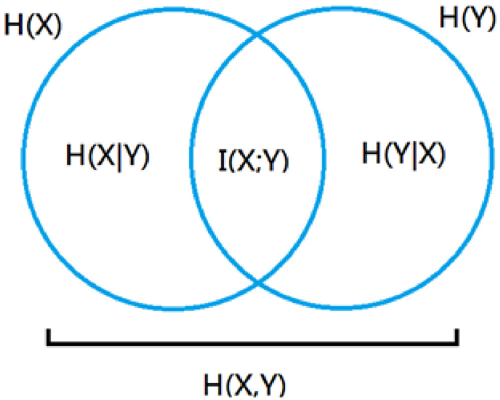
RELATIONSHIP: $H(X, Y) \leq H(X) + H(Y)$

- We know $H(X, Y) = H(X) + H(Y|X)$
- Therefore we need only show $H(Y|X) \leq H(Y)$
- This makes sense, knowing X can only decrease the addition information provided by Y.

RELATIONSHIP: $H(X, Y) \leq H(X) + H(Y)$

- We know $H(X, Y) = H(X) + H(Y|X)$
- Therefore we need only show $H(Y|X) \leq H(Y)$
- This makes sense, knowing X can only decrease the addition information provided by Y.
- Proof? Possible homework =)

Entropy Relationships



Mutual Information

- The *Mutual Information* $I(X;Y)$ is defined as:

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

- The mutual information is the amount of information shared by X and Y.
- It is a measure of how much X tells us about Y, and vice versa.
- If X and Y are independent then $I(X;Y) = 0$, because X tells us nothing about Y and vice versa.
- If $X = Y$ then $I(X;Y) = H(X) = H(Y)$. X tells us everything about Y and vice versa.

EXAMPLE

Marginal Distribution:

X	sun	rain
$P(X)$	0.6	0.4

Y	hot	cold
$P(Y)$	0.6	0.4

Conditional Distribution:

Y	hot	cold
$P(Y\|X = \text{sun})$	0.8	0.2

Y	hot	cold
$P(Y\|X = \text{rain})$	0.3	0.7

Joint Distribution:

	hot	cold
sun	0.48	0.12
rain	0.12	0.28

Example: Marginal Entropy

Marginal Distribution:

X	sun	rain
$P(X)$	0.6	0.4

$$
\begin{aligned}
H(X) &= \sum_i p(x_i)\log\left(\frac{1}{p(x_i)}\right) \\
&= 0.6\log\left(\frac{1}{0.6}\right) + 0.4\log\left(\frac{1}{0.4}\right) \\
&= 0.97
\end{aligned}
$$

Example: Joint Entropy

Joint Distribution:

	hot	cold
sun	0.48	0.12
rain	0.12	0.28

$$
\begin{aligned}
H(X) &= \sum_{i,j} p(x_i, y_i) \log(\frac{1}{p(x_i, y_i)}) \\
&= 0.48 \log(\frac{1}{0.48}) + 2\left[0.12 \log(\frac{1}{0.12})\right] + 0.28 \log(\frac{1}{0.28}) \\
&= 1.76
\end{aligned}
$$

Example: Conditional Entropy

Joint Distribution:

	hot	cold
sun	0.48	0.12
rain	0.12	0.28

Conditional Distribution:

Y	hot	cold
$P(Y\|X = \text{sun})$	0.8	0.2

Y	hot	cold
$P(Y\|X = \text{rain})$	0.3	0.7

$$
\begin{aligned}
H(Y|X) &= \sum_{i,j} p(x_i, x_j) \log\left(\frac{1}{p(y_i|x_i)}\right) \\
&= 0.48 \log\left(\frac{1}{0.8}\right) + 0.12 \log\left(\frac{1}{0.2}\right) + 0.12 \log\left(\frac{1}{0.3}\right) + 0.28 \log\left(\frac{1}{0.7}\right) \\
&= 0.79
\end{aligned}
$$

Example: Summary

- Results:
  - $H(X) = H(Y) = 0.97$
  - $H(X, Y) = 1.76$
  - $H(Y|X) = 0.79$
  - $I(X; Y) = H(Y) - H(Y|X) = 0.18$
- Note that $H(X, Y) = H(X) + H(Y|X)$ as required.
- Interpreting the Results:
  - $I(X; Y) > 0$, therefore X tells us something about Y and vice versa
  - $H(Y|X) > 0$, therefore X doesn't tell us everything about Y

Motivation Recap

- Gambling: Coins vs. Dice vs. Roulette
- Prediction: Bent Coin vs. Fair Coin
- Compression: How to best record a sequence of events

Outline

Prefix Codes

- Compression maps events to code words
- We already saw an example when we mapped coin tosses to unary numbers
- We want mapping which generates short encodings
- One good way of doing this is prefix codes

Prefix Codes

- Encoding where no code word is a prefix of any other code word.
- Example:

Event	a	b	c	d
Code Word	0	10	110	111

- Previously we reserved 0 as a separator
- If we use a prefix code we do not need a separator symbol

$$1010001101111110111 \quad = \quad bbaacdcd$$

Distribution as Prefix Codes

- Every probability distribution can be thought of as specifying an encoding via the Information $I(X)$
- Map each event x_i to a word of length $I(x_i)$

Table: Fair Coin

X	h	t
$P(X)$	0.5	0.5
$I(X)$	1	1
$code(X)$	1	0

DISTRIBUTION AS PREFIX CODES

- Every probability distribution can be thought of as specifying an encoding via the Information $I(X)$
- Map each event x_i to a word of length $I(x_i)$

Table: Fair 4-Sided Dice

X	1	2	3	4
$P(X)$	0.25	0.25	0.25	0.25
$I(X)$	2	2	2	2
$code(X)$	11	10	01	00

Distribution as Prefix Codes

- Every probability distribution can be thought of as specifying an encoding via the Information $I(X)$
- Map each event x_i to a word of length $I(x_i)$

Table: Bent 4-Sided Dice

X	1	2	3	4
$P(X)$	0.5	0.25	0.125	0.125
$I(X)$	1	2	3	3
$code(X)$	0	10	110	111

DISTRIBUTION AS PREFIX CODES

- Prefix codes built from the distribution are optimal
  - Information is contained in the smallest possible number of characters
  - Entropy is maximized
- Encoding is not always this obvious. e.g. How to encode a bent coin
- Question: If use a different (suboptimal) encoding, how many extra characters do I need

KL Divergence

KL Divergence

- The expected number of additional bits required to encode p using q, rather than p using p.

$$
\begin{aligned}
D_{KL}(p||q) &= \sum_i p(x_i)\left|code_q(x_i)\right| - \sum_i p(x_i)\left|code_p(x_i)\right| \\
&= \sum_i p(x_i) I_q(x_i) - \sum_i p(x_i) I_p(x_i) \\
&= \sum_i p(x_i)\log\left(\frac{1}{q(x_i)}\right) - \sum_i p(x_i)\log\left(\frac{1}{p(x_i)}\right)
\end{aligned}
$$

KL Divergence

- The KL Divergence is a measure of the 'Dissimilarity' of two distributions
- If p and q are similar, then $KL(p||q)$ will be small.
    - Common events in p will be common events in q
    - This means they will still have short code words
- If p and q are dissimilar, then $KL(p||q)$ will be large.
    - Common events in p may be uncommon events in q
    - This means commonly occuring events might be given long codewords

SUMMARY

Motivation

Information

Entropy
- Marginal Entropy
- Joint Entropy
- Conditional Entropy
- Mutual Information

Compressing Information
- Prefix Codes
- KL Divergence