Information Theory
Entropy

www.Stats-Lab.com

Twitter: @StatsLabDublin

Information Theory: Entropy

- The input source to a noisy communication channel is a random variable X over the four symbols $\{a, b, c, d\}$.
- The output from this channel is a random variable Y over these same four symbols.

Information Theory: Entropy

The joint distribution of these two random variables is as follows:

	x=a	x=b	x=c	x=d
y=a	1/8	1/16	1/16	1/4
y=b	1/16	1/8	1/16	0
y=c	1/32	1/32	1/16	0
y=d	1/32	1/32	1/16	0

Information Theory: Entropy

1. Write down the marginal distribution for X and compute the marginal entropy $H(X)$.
2. Write down the marginal distribution for Y and compute the marginal entropy $H(Y)$.

Information Theory: Entropy

The marginal distribution of these two random variables is as follows:

	x=a	x=b	x=c	x=d	P(Y)
y=a	1/8	1/16	1/16	1/4	
y=b	1/16	1/8	1/16	0	
y=c	1/32	1/32	1/16	0	
y=d	1/32	1/32	1/16	0	
P(X)					

Information Theory: Entropy

The marginal distribution of these two random variables is as follows:

	x=a	x=b	x=c	x=d	P(Y)
y=a	1/8	1/16	1/16	1/4	0.50
y=b	1/16	1/8	1/16	0	0.25
y=c	1/32	1/32	1/16	0	0.125
y=d	1/32	1/32	1/16	0	0.125
P(X)	0.25	0.25	0.25	0.25	

Information Theory: Entropy

- H(X), the entropy of X, is computed as

$$H(X) = -\sum P(x_i)\log_2 P(x_i)$$

- Computing the logarithms

x_i	a	b	c	d
$P(x_i)$	0.25	0.25	0.25	0.25
$\log(P(x_i))$				

Information Theory: Entropy

$$H(X) = -\sum P(x_i) \log_2 P(x_i)$$

x_i	a	b	c	d
$P(x_i)$	0.25	0.25	0.25	0.25
$\log\left(P(x_i)\right)$	-2	-2	-2	-2
$P(x_i) \times \log\left(P(x_i)\right)$				

Information Theory: Entropy

- H(Y), the entropy of Y, is computed as

$$H(Y) = -\sum P(y_j) \log_2 P(y_j)$$

- Computing the logarithms

y_j	a	b	c	d
$P(y_j)$	0.50	0.25	0.125	0.125
$\log(P(y_j))$				

Information Theory: Entropy

$$H(Y) = -\sum P(y_j) \log_2 P(y_j)$$

y_j	a	b	c	d
$P(y_j)$	0.50	0.25	0.125	0.125
$\log(P(y_j))$	-1	-2	-3	-3
$P(y_j) \times \log(P(y_j))$				

Information Theory: Entropy

- H(X), the entropy of X, is

$$H(X) = 2\text{b}.$$

- H(Y), the entropy of Y, is

$$H(Y) = 1.75\text{b}.$$

Information Theory: Entropy

Entropies: Example (f)

- To compute the joint entropy $H(X,Y)$, we will use $H(X,Y) = -\sum P(x_i,y_y)\log_2 P(x_i,y_j)$
- This means we should compute the entropy component for each cell of the table, and sum up all the resultant terms.
- To save time, we will aggregate similar results,
  - there are 4 cells where the probability is $1/32$,
  - 6 cells with probability $1/16$,
  - 2 cells with probability $1/8$
  - and 1 cell with probability $1/4$.
- Solving

$$H(X,Y) = [4 \times -\frac{1}{32}\log_2\frac{1}{32}] + [6 \times -\frac{1}{16}\log_2\frac{1}{16}] + \ldots + [1 \times -\frac{1}{4}\log_2\frac{1}{4}]$$

Entropies: Example (g)

- Simplifying

$$H(X,Y) = [4 \times -\frac{1}{32}\log_2\frac{1}{32}] + [6 \times -\frac{1}{16}\log_2\frac{1}{16}] + \ldots + [1 \times -\frac{1}{4}\log_2\frac{1}{4}]$$

- Simplifying

$$H(X,Y) = [-\frac{4}{32} \times -5] + [-\frac{6}{16} \times -4] + [-\frac{2}{8} \times -3] + [-\frac{1}{4} \times -2]$$

- $H(X,Y) = 27/8$ b.

Entropies: Example (h)

From last lecture, two useful relationships among the types of entropies are

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

Re-arranging these formulae

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

Entropies: Example (i)

Re-arranging these formulae

- $H(X|Y) = H(X,Y) - H(Y) = 27/8 - 14/8 = 13/8$ b.
- $H(Y|X) = H(X,Y) - H(X) = 27/8 - 16/8 = 11/8$ b.
- Remark $1.75 = 14/8$ and $2 = 16/8$.
- Also: we will derive $H(Y|X$ and $H(X|Y)$ from first principles in a tutorial.

Entropies: Example (j)

There are three alternative ways to obtain the answer:

- $I(X;Y) = H(Y) - H(Y|X) = 7/4 - 11/8 = 3/8$ b.
- $I(X;Y) = H(X) - H(X|Y) = 2 - 13/8 = 3/8$ b.
- $I(X;Y) = H(X) + H(Y) - H(X,Y) =$
  $2 + 7/4 - 27/8 = (16 + 14 - 27)/8 = 3/8$b.

Kraft inequality

- Let X be a DMS with alphabet ($x_i = \{1, 2, ..., m\}$). Assume that the length of the assigned binary code word corresponding to x, is n.
- A necessary and sufficient condition for the existence of an instantaneous binary code is

$$K = \sum_{i=1}^{m} 2^{-n_i} \leq 1$$

  which is known as the **Kraft inequality**.
- Note that the Kraft inequality assures us of the existence of an instantaneously decodable code with code word lengths that satisfy the inequality. But it does not show us how to obtain these code words, nor does it say that any code that satisfies the inequality is automatically uniquely decodable

Data compression(1)

Data compression is the science (and art) of representing information in a compact form. Having been the domain of a relatively small group of engineers and scientists, it is now ubiquitous.

It has been one of the critical enabling technologies for the on-going digital multimedia revolution for decades. Without compression techniques, none of the ever-growing Internet, digital TV, mobile communication or increasing video communication would have been practical developments.

Data compression(1)

Data compression is an active research area in computer science. By ”compressing data”, we actually mean deriving techniques or, more specifically, designing efficient algorithms to:

- represent data in a less redundant fashion
- remove the redundancy in data
- implement coding, including both encoding and decoding.

Data compression(2)

The key approaches of data compression can be summarized as modelling + coding. Modelling is a process of constructing a knowledge system for performing compression. Coding includes the design of the code and product of the compact data form.