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.
- The marginal entropies for X and Y are
  $H(X) = 2$bs,
  $H(Y) = 1.75$bs.
- The joint entropy of X and Y is
  $H(X, Y) = 3.375$bs.

Information Theory: Entropy

1. What is the conditional entropy $H(Y|X)$?
2. What is the conditional entropy $H(X|Y)$?
3. What is the mutual information $I(X;Y)$ between the two random variables?

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

Relationship between conditional, joint and marginal entropy.

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

Re-arranging these equations

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

Information Theory: Entropy

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

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

Information Theory: Entropy

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

  $= 3.375 - 1.75 = 1.625$ b

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

  $= 3.375 - 2.0 = 1.375$ b

Information Theory: Entropy

The formula for computing mutual information $I(X;Y)$ is

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

- $H(X) = 2\text{b}$
- $H(Y) = 1.75\text{b}$
- $H(X,Y) = 3.375\text{b}$

Information Theory: Entropy

The formula for computing mutual information $I(X;Y)$ is

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

- $H(X)$ = 2b
- $H(Y)$ = 1.75b
- $H(X,Y)$ = 3.375b

$$I(X;Y) = 3 + 1.75 - 3.375 = 0.375\text{b}$$

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.

Self Information

Self-information This is defined by the following mathematical formula:$I(A) = logbP(A)$
The self-information of an event measures the amount of one's surprise evoked by the event. The negative logarithm $logbP(A)$ can also be written as

$$log_b \frac{1}{P(A)}$$

Note that log(1) = 0, and that $|log(P(A))|$ increases as P(A) decreases from 1 to 0. This supports our intuition from daily experience. For example, a low-probability event tends to cause more "surprise".

Example

For a simple example, we will take a short phrase and derive our probabilities from a frequency count of letters within that phrase. The resulting encoding should be good for compressing this phrase, but of course will be inappropriate for other phrases with a different letter distribution.
”All you base are belong to us”

Entropy

- Entropy is the uncertainty of a single random variable.
- We can define ***conditional entropy*** $H(X|Y)$, which is the entropy of a random variable conditional on the knowledge of another random variable.
- The reduction in uncertainty due to another random variable is called the ***mutual information***.

What is Information?

- Once we agree to define the information of an event ain terms of P(a), the properties (2) and (3) will be satisfied if the information in ais defined as

$$I(a) = -logP(a)$$

- Remark : The base of the logarithm depends on the unit of information to be used.

Ambiguity occurs if there is any path to some symbol whose label is a prefix of the label of a path to some other symbol. In the Huffman tree, every symbol is a ***leaf***. Thus it is impossible for the label of a path to a leaf to be a prefix of any other path label, and so the mapping defined by Huffman coding has an inverse and decoding is possible.

Example

For a simple example, we will take a short phrase and derive our probabilities from a frequency count of letters within that phrase. The resulting encoding should be good for compressing this phrase, but of course will be inappropriate for other phrases with a different letter distribution.
"All you base are belong to us"