Introduction to Information Theory

Eric Filiol

ESIEA - Laval
Laboratoire de cryptologie et de virologie opérationnelles
$(C+V)^O$
filiol@esiea.fr

2013 - 2014

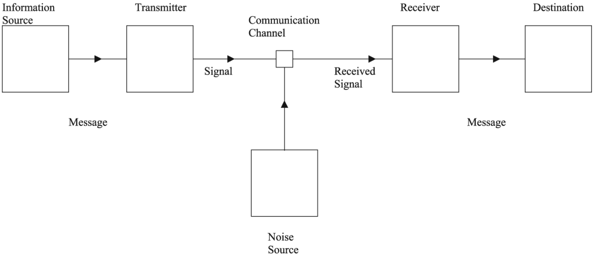
Introduction : What is Information

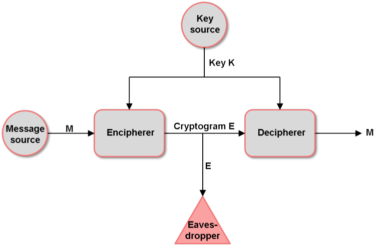
- Consider the following propositions
  - A A race between two equally matched horses is less uncertain that a race between six evenly matched horses.
  - B The outcome of a spin on a roulette wheel is more uncertain than the throw of a die.
  - C The throw of a fair die is more uncertain than the throw of a biased die in which the probabilities are $\frac{1}{10}$ of getting each of the numbers 1 to 5 and probability $\frac{1}{2}$ of getting a 6.
- What about their validity ? Could you formalize easily those propositions ? What is uncertainty ?
- Uncertainty means/implies also the effort one should make to guess information (the case of cryptology).

Introduction : What is Information (2)

- Information theory deals with mathematical problems regarding the representation, the storage, the transformation, the transmission of information.
- Modern life is overwhelmed by all types of information.
- It is necessary to define what is information and how to define a measure of information.

Information Theory (1)





Information Theory (2)

- Information must be represented (essential difference with ideas or concepts).
  - Information is intrinsic to the existence of a physical medium (paper, air, hard disk, copper wire, optical fiber...).
- The concept of *unpredictability* is essential : why transmit an information which is obvious. Information is by essence unpredictable. So any measure of information will be that of its unpredictability degree/level.
- However very often a part only of received information is new (e.g. cell telephone number or Social insurance code). So information is never totally unexpected.

Information Theory (3)

- We have to use *signals* and *encoding*. We then have to choose *signs* or *symbols* to build *messages*.
- Key issues :
  - Cost of information representation, transmission, storage...
  - Required properties : unicity (to avoid ambiguity and equivocation), transinformation, universality...
- Without loss of generality we will consider only discrete signals/symbols.
- Two kind of information : *useful information* and *parasite information* or *noise*. The difference between the two is relative and subjective.
- Theory developped by C. E. Shannon in 1948-1949. Initial works by Harry Nyqyst and Ralph Hartley (1920).

Agenda

Agenda

1. Introduction : What is Information and Information Theory ?
2. Uncertainty
3. Entropy and Its Properties
4. Conditional Entropy
5. Measure of Information
6. Conclusion
7. Bibliography

Introduction : Statistical Description

- Suppose that X and Y are two distinct random variables such that

$$P(X=0)=p \quad P(X=1)=1-p \text{ while } P(Y=100)=p \quad P(Y=200)=1-p$$

- Any definition of uncertainty should give X and Y the same uncertainty. In other words, it should be a function on the probability p only.
- This definition and the relevant properties should extend to variables taking more than 2 values.

Definition of Uncertainty

The uncertainty of a random variable X, which takes the values x_i with probabilities $p_i, (1 \leq i \leq n)$ is to be a function *only* of the probabilities $p_1, \ldots, p_n$.

Let us denote this function $H(p_1, \ldots, p_n)$.

Postulates for $H(p_1, \ldots, p_n)$

A1 $H(p_1, \ldots, p_n)$ is maximum whenever $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$

A2 For any permutation $\pi \in S(n)$ we have $H(p_1, \ldots, p_n) = H(p_{\pi(1)}, \ldots, p_{\pi(n)})$

A3 $H(p_1, \ldots, p_n) \geq 0$ and $H(p_1, \ldots, p_n) = 0$ whenever $\exists i \in [1, \ldots, n]$ such that $p_i = 1$

A4 $H(p_1, \ldots, p_n, 0) = H(p_1, \ldots, p_n)$

A5 $H(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) \leq H(\frac{1}{n+1}, \frac{1}{n+1}, \ldots, \frac{1}{n+1})$.

A6 $H(p_1, \ldots, p_{n-1})$ should be a continuous and strictly concave function.

A7 $\forall (n, m) \in \mathbb{N}^2, H(\frac{1}{nm}, \frac{1}{nm}, \ldots, \frac{1}{nm}) = H(\frac{1}{n}, \ldots, \frac{1}{n}) + H(\frac{1}{m}, \ldots, \frac{1}{m})$

A8 Let $p = p_1 + \ldots p_m$ and $q = q_1 + \ldots q_n$ with each p_i and q_j being non negative and p, q being positive while $p + q = 1$, we must have

$$H(p_1, \ldots, p_m, q_1, \ldots, q_n) = H(p, q) + p.H(\frac{p_1}{p}, \ldots, \frac{p_m}{p}) + q.H(\frac{q_1}{q}, \ldots, \frac{q_n}{q})$$

Entropy Theorem

Theorem

Let $H(p_1, \ldots, p_n)$ be a function defined for any $n \in \mathbb{N}$ and $\forall (p_1, \ldots, p_n) \in \mathbb{R}^n$ with $p_i \in [0, \ldots, 1] \subset \mathbb{R}$ such that $\sum_{i=1}^{n} p_i = 1$.
If h is to satisfy the axioms [A1]-[A8], then

$$H(p_1, \ldots, p_n) = -\lambda \sum_{i=1}^{n} p_i \log(p_i)$$

with λ any positive constant and where the sum is for those i for which $p_i > 0$.

- The system of axioms [A1]-[A8] has been proposed by (Shannon, 1948). It is not minimal (Aczél & Daróczy, 1975).
- Proof left as exercice.

Random Variable Entropy

- For X an random variable that takes a finite set of values with probabilities $p_1, \ldots, p_n$, We define *Entropy* or *Uncertainty* of X as

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

  under the same condition as in the Entropy Theorem.
- Axiom [A1] implies that $H(\frac{1}{2}, \frac{1}{2}) = 1$. This expresses that the information unit is the *bit* (standing for Binary unIT).
- Exercices.

Agenda

Properties of Entropy

- Let X be a random vector which takes only a finite number of values $u_1, u_2, \ldots, u_n$. We define its entropy by

$$H(X) = -\sum_{i=1}^{n} p(u_i) \log_2(p(u_i))$$

  - For $n = 2$ with $X = (U, V)$ and $p_{ij} = P(U = u_i, V = b_j)$ then we write
$$H(X) = H(U, V) = -\sum_{i,j} p_{ij} \log_2(p_{ij})$$

- More generally, if $X_1, X_2, \ldots, X_n$ is a collection of random variables each taking only a finite number of values, we can consider the random vector $X = (X_1, X_2, \ldots, X_n)$ which takes also a finite number of values and define the *joint entropy* by

$$H(X) = H(X_1, X_2, \ldots, X_n) = -\sum p(x_1, x_2, \ldots, x_n) \log_2(p(x_1, x_2, \ldots, x_n))$$

where $p(x_1, x_2, \ldots, x_n) = p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$.

A First Inequality on Entropy

Theorem

For any $n \in \mathbb{N}$

$$H(p_1, p_2, \dots, p_n) \leq \log_2(n)$$

with equality if and only if $p_i = \frac{1}{n} \quad \forall i \in [1, \dots, n] \subset \mathbb{N}$.

- Proof left as an exercise (hint : $\log_e$ is a concave function).

Key Lemma

If $(p_i : 1 \leq i \leq n)$ is a given probability distribution, then the minimum of

$$G(q_1, \dots, q_n) = -\sum_{i=1}^{n} q_i \log_2(q_i)$$

over all probability distributions $(q_1, \dots, q_n)$, is achieved when $q_k = p_k, \ (1 \leq k \leq n)$.

A Second Inequality on Entropy

The previous Lemma is useful to prove the following key Theorem.

Theorem

If X and Y are any two random variables taking only a finitely many values, then

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

with equality holding if and only if X and Y are independent.

- This can be extended to more than two random variables e.g. $X_1, X_2, \ldots, X_n$ the equality holding when the variables are mutually independent.
- We extend this result two any pair of random vectors (U, V) and we have

$$H(U,V) \leq H(U) + H(V)$$

Agenda

Conditional Entropy

Suppose that X Y are random variables on a probability space Ω, taking many finitely values, and A is an event in Ω.

- We define the *conditional entropy* of X given A by

$$H(X|A) = -\sum_{i=1}^{n} P(X = x_i|A)\log_2(P(X = x_i|A))$$

- We define in the same way the conditional entropy of X given Y (or the *equivocation* of Y about X) by

$$H(X|Y) = -\sum_{y_j}^{n} H(X|Y = y_j).P(Y = y_j)$$

where $H(X|Y = y_j) = -\sum_{x_i} P(X = x_i|Y = y_j)\log_2(P(X = x_i|Y = y_j))$

- $H(X|Y)$ is the uncertainty on X given a particular value of Y, averaged over the range of values that Y can take. In other words, it the remaining uncertainty about X after Y has been observed.

Properties of Conditional Entropy

- Trivial property :
$$H(X|X) = 0$$
- If X and Y are independent then we have :
$$H(X|Y) = H(X)$$
- $H(X|Y)$ is the uncertainty on X given a particular value of Y, averaged over the range of values that Y can take (exercice).
- This notion extends easily to random vectors.
$$H(U|V) = -\sum_{i=1}^{n} H(U|V = v_i).P(V = v_i)$$
- $H(U|V)$ measures the uncertainty about U contained in V and we can prove that (proof left as an exercise)
$$H(U|V) = 0 \text{ if and only if } U = g(V) \text{ for some } G$$

Properties of Conditional Entropy (2)

Theorem : Chain rule

For any two pair of random variables X and Y that take only a finitely many values, and for U and V two random vectors each taking only a finite set of values then

$$H(X,Y) = H(Y) + H(X|Y) \text{ and } H(U,V) = H(V) + H(U|V)$$

- This result expresses mathematically the idea that conditional entropy of X given Y correctly measures the remaining uncertainty (proof left as an exercise).
- We then can give the following corollary (proof left as an exercise).

Corollary

For any pair of X and Y (random variables or random vectors)

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

with equality if and only if X and Y are independent.

Properties of Conditional Entropy (3)

Corollary

For any three random variables X, Y and Z that take only a finitely many values, then

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

- The proof is similar to that of the Chain rule theorem.

Corollary

Let $X_1, X_2, \ldots, X_n$ random variables drawn according to $p(x_x, x_2, \ldots, x_n)$. Then

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

- Proof by using the previous corollary.

Agenda

Introduction

- We would like to have a measure of information while until now we have just defined a measure of uncertainty.
- First attempt by (Hartley, 1928).
- Suppose E_1 and E_2 two events on probability space Ω, or respective probability p_1 and p_2. Any "natural" measure of information I should satisfy
$$I(p_1, p_2) = I(p_1) + I(p_2)$$
- I must be a continuous, positive function, so for any event E we choose
$$I(E) = -\log_2(P(E))$$
- Let us extend this concept to random variables and random vectors to define the useful concept of *transinformation* or *mutual information*.

Transinformation or Mutual Information

Let X and Y two random variables. We want to express the amount of information that Y reveals about X. we denote this $I(X;Y)$ or $I(X|Y)$.

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

We then have

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

$$I(X;Y) = 0 \text{ if and only if } X \text{ and } Y \text{ are independent}$$

$$I(X;Y) = I(Y;X)$$

Agenda

Conclusion

- Uncertainty and information are essentially the same quantities.
- The removal of uncertainty can be considered as giving information.
- Both are measured with the mathematical concept of entropy.
- The use of base 2 for the logarithm defines the unit of entropy as the *bit*.
- Go now to the computer room to practice with exercices.

Agenda

Essential Bibliography

A few papers are available on the Moodle repository for this lecture.

- Aczél, J. & Daróczy, Z. (1975). *On Measures of Information and Their Characterizations*. Academic press.
- Hartley, R.V.L. (1928). Transmission of Information, *Bell System Technical Journal*, Volume 7, Number 3, pp. 535–563.
- Nyquist, H. (1924). Certain Factors Affecting Telegraph Speed. *Bell System Technical Journal*, 3, 324–346.
- Nyquist, H. (1928). Certain Topics in Telegraph Transmission Theory, *Trans. AIEE*, vol. 47, pp. 617–644.
- Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656.
- Welsh, D. (1988). *Codes and Cryptography*, Oxford Science Publications.