Notes on Information Theory

Cristian Di Pietrantonio, Michele Laurenti, Lorenzo Madeddu

October 12, 2016

Chapter 1

Introduction

1.1 Lesson 1

Information Theory is born from one man, Cloude Shannon (1926 - 2001); it has to do with probability, algebra, coding theory, ergodic theory and appears in daily life. In communication you want to be understood, but you also want to avoid redundancy and words tend to be lengthy; there is a conflict of interest. Why some words are short and others are long? Short words encode common abstract pictures whereas long words must carry the whole information not available in common knowledge [nessuna fonte].

Suggested readings:

- Tom Cover, Joy Thomas, Information Theory, Wiley;
- IT: cooling theorems for discrete memoryless systems, Korner;
- IT, Robert Ash.

1.2 Lesson 2

Define $\mathcal{M}$ as the finite set of all messages over an alphabet Σ. A message m is an element of $\mathcal{M}$. Hartley, who introduced this concept before Shannon, stated that the maximum amount of information is equal to $\log_2 |\mathcal{M}|$, since one can identify elements of $\mathcal{M}$ with binary strings of length $\log_2 |\mathcal{M}|$. Formally

$$\mathcal{M} \sim \{0,1\}^{\lceil \log_2 |\mathcal{M}| \rceil}$$

and

$$m \sim \Delta \in \{0,1\}^{\lceil \log_2 |\mathcal{M}| \rceil}.$$

One can think of Δ as a sequence of $\log_2 |\mathcal{M}|$ binary questions that identify an element in $\mathcal{M}$; a single unit of information is a 0 or 1 (Binary

Information Theory). In the *20 question game*, where you need to minimize the number of questions to guess what a person thinks, is given a probability distribution P over $\mathcal{M}$ s.t. $P(m) \geq 0,\ \forall m \in \mathcal{M}$ and $\sum_{m \in \mathcal{M}} P(m) = 1$. The questions can be designed to make the game as short as possible if P is known.

Must be noted that $|\underline{x}| = i \Leftrightarrow \underline{x} \in \{0,1\}^i$. The number of questions asked on average must be minimized. Let $f : \mathcal{M} \to \{0,1\}^*$, in other words a mapping from the message set to the sequences of answers to find the message then we want to minimize

$$\sum_{m \in \mathcal{M}} P(m)|f(m)|. \tag{1.1}$$

This equation is tied only to the probability distribution, not to the set or the questions. If P is uniform then 1.1 is equal to $\lceil \log_2 |\mathcal{M}| \rceil$. The function f is called **code**. The code is not always invertible.

Communication happens at a distance, otherwise it would be trivial. There is some device that makes communication possible; the channel is not perfect. The noise it is random, modelled through a probability distribution. The information source is not predictable either and will have its own probability distribution too (transmission error ?). Encoding and decoding must be optimized to make communications short to reduce “lost”, and to reliably transmit, i.e. the receiver is reasonably sure that it gets what was sent. The two aspects are in contrast, you need and optimal trade-off. Usually partners in communication are separated in space and communication takes place in time. Sometimes in computer science the opposite thing happens: communication is storing information in space to retrieve it later in time. Sometimes you want to shorten the communication time, sometimes the space communication takes. The model is the same. Entropy comes from thermodynamics, introduced by Boltzmann.

1.3 Lesson 3 - Hamming Space

A space is a set that has structure. Consider the set of all binary strings $\{0,1\}^*$; we can define a space with the distance metric. The **distance** is defined as a function $d : X \times X \to \mathbb{R}_0^+$. A metric has the following properties:

1. (a) $d(x,y) \geq 0,\ \forall (x,y) \in X \times X$
   (b) $d(x,y) = 0 \Leftrightarrow x = y$
2. $d(x,y) = d(y,x), \forall (x,y) \in X \times X$
3. $d(x,y) \leq d(x,z) + d(z,y)$

What is the Hamming metric? Consider $\underline{x}, \underline{y} \in \{0,1\}^n$. Then the Hamming distance is defined as

$$d_H(\underline{x}, \underline{y}) = |\{i \mid x_i \neq y_i\}| \tag{1.2}$$

Why is it a metric? Consider $\underline{x}$, $\underline{y}$, $\underline{z} \in \{0,1\}^n$. Then, with $D = \{i \mid x_i \neq y_i\}$ is true that

$$i \in D \Rightarrow x_i \neq y_i \Rightarrow \neg(z_i = x_i \wedge z_i = y_i) \Rightarrow d_H(x_i, y_i) \leq d_H(z_i, x_i) + d_H(z_i, y_i).$$

By the additive property we can write

$$d_H(x, y) = \sum_{i=1}^{n} d_H(x_i, y_i).$$

Summing up gives the triangle inequality.

If a string $\underline{x}$ has been changed r times to become $\underline{y}$ then $d_H(\underline{x}, \underline{y}) \leq r$. Define a **Hamming ball** of radius r around $\underline{x}$ as

$$B_H = B_H(\underline{x}, r) = \{\underline{y} \mid d_H(\underline{x}, \underline{y}) \leq r\}.$$

Obviously r needn't to be an integer (use the *floor* function). Here are some properties of Hamming balls:

- if one subtracts $B_H(\underline{x}, r)$ to $\{0,1\}^n$ then the result is also an Hamming ball;
- $\{0,1\}^n = B_H(\underline{x}, n)$, $\forall \underline{x} \in \{0,1\}^n$, but if $r < n$ then the center is unique;

What we can say about $B_H(\underline{x}, r)$? For simplicity sake, consider $B_H(\underline{0}, r)$. Then the following holds:

$$|B_H(\underline{0}, r)| = \sum_{i=0}^{\lfloor r \rfloor} \binom{n}{i} \tag{1.3}$$

That is the number of ways in which we can flip to 1 the bits. From this follows that $d_H(\underline{0}, \underline{y}) + d_H(\underline{1}, \underline{y}) = n$.

We define the Hamming weight as $w_H(\underline{x}) = d_H(\underline{0}, \underline{x})$. The following holds:

$$\underline{y} \in B_H(\underline{0}, r) \Rightarrow w_H(\underline{y}) > r \Rightarrow d_H(\underline{y}, \underline{1}) < n - r \leq n - r - 1.$$

So it is true that

$$\overline{B_H(\underline{0}, n)} = B_H(\underline{1}, r') = B_H(\underline{1}, n - r - 1).$$

In other words the Hamming space can be partitioned into two Hamming balls. If n is odd an Hamming space can be partitioned in the following way:

$$\{0,1\}^n = B_H(\underline{0}, \lfloor \frac{n}{2} \rfloor) \cup B_H(\underline{1}, \lfloor \frac{n}{2} \rfloor)$$

The Hamming space cannot be partitioned into three balls. In how many balls can $\{0,1\}^n$ be partitioned? It cannot partitioned be partititoned into s balls with $3 \leq s \leq n+1$.

1.3.1 The volume of a Hamming ball

We have seen that the volume of a generic Hamming ball is described by equation 1.3. We can use the Pascal triangle to get a feeling about the order of magnitude of the Hamming ball. Consider the nth row in the triangle; since it is symmetric, if we split the row in half the sum of the terms of the first part is equal to the sum of the terms of the second part of the row. The volume of the greatest ball is $|B_H(\underline{0}, n)| = 2^n$; instead, the volume of the ball with radius $n/2$ (in the triangle's point of split) is $2^n/2 = 2^n 2^{-1} = 2^{n-1}$. So, if $r \geq n/2$ then $2^{n-1} \leq |B_H(\underline{0}, r)| \leq 2^n$. Can we bound the volume for $r < n/2$?

First we introduce the notion of **entropy**, a function $h : [0,1] \to [0,1]$ defined as follows:

$$h(t) = t \log_2(\frac{1}{t}) + (1-t) \log_2 \left(\frac{1}{1-t} \right)$$

The plot of h looks like this:

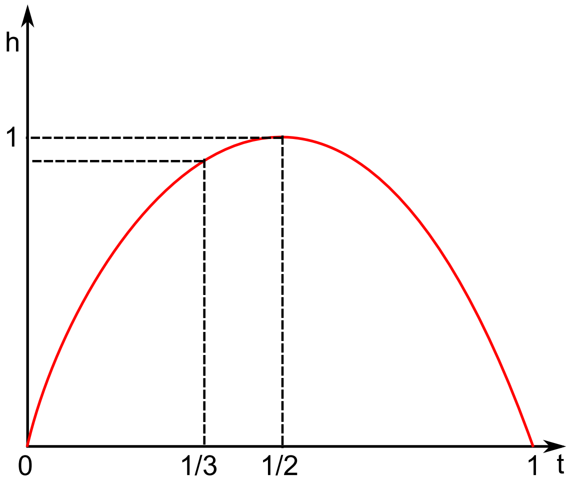

Figure 1.1: The entropy function.

This function is not defined for the values 0 and 1 but we have limits defined on these points, and they are both 0. So we artificially set $h(0) = 0$

and $h(1) = 0$. It is better to think about it as probability distribution $(t, 1-t)$ and entropy is a number attached to it. Entropy measures symmetry and with $t = 1/2$ we have maximum chaos (the future outcomes are equally likely).

1.4 Lesson 4

Theorem 1 *The upper bound on the volume of a Hamming ball, if* $r \leq n/2$*, is*

$$|B_H(\underline{0}, r)| \leq 2^{nh(\frac{r}{n})}$$

In order to prove this theorem an analogy is introduced. Suppose there is a box with a number of chickens in it. We want to count those animals without withdrawing all of them out of the box. What can be done is to take the lightest one and measure its weight w; then also the weight W of the box is measured. The number of the total chickens in the box can't be greater than w/W.

Proof. In this proof we will use a similar technique. Consider $\{0,1\}^n$. We "sparkle" a substance on the strings in the set; this substance looks like probability, but it doesn't matter. Define the weight of 1 and 0 as

$$P(1) = \frac{r}{n},\ P(0) = 1 - P(1).$$

Notice that is not an uniform distribution. Define the weight of a string as

$$P^n(\underline{x}) = \prod_{i=1}^{n} P(x_i).$$

If $A \subseteq \{0,1\}^n$ then the weight of the set A is

$$P^n(A) = \sum_{\underline{x} \in A} P^n(\underline{x}).$$

$P^n(\{0,1\}^n)$ is the total weight of the substance sparkled on the strings and it is the probability distribution of binomial. For this reason one can claim that

$$1 = P^n(\{0,1\}^n) \geq P^n(B_H(\underline{0}, r)) = \sum_{\underline{x} \in B_H(\underline{0}, r)} P^n(\underline{x})$$

at this point we "take out the lightest chicken" and write

$$\sum_{\underline{x} \in B_H(\underline{0}, r)} P^n(\underline{x}) \geq |B_H(\underline{0}, r)| \cdot \min_{\underline{x} \in B_H(\underline{0}, r)} P^n(\underline{x})$$

Which are the lightest chickens? Because we assume $r \leq n/2$ then

$$r \leq \frac{n}{2} \Rightarrow P(1) = \frac{r}{n} \leq \frac{1}{2} \Rightarrow P(1) \leq P(0).$$

It follows that the the lightest strings are the ones on the border of the ball, with r 1's. We now compute their weight.

$$\begin{aligned}\min_{\underline{x} \in B_H(\underline{0},r)} P^n(\underline{x}) = [P(1)]^r \cdot [P(0)]^{n-r} = \\ = \left(\frac{r}{n}\right)^r \left(1 - \frac{r}{n}\right)^{n-r} = \\ = \left(\frac{r}{n}\right)^{n\frac{r}{n}} \left(1 - \frac{r}{n}\right)^{n\left(1-\frac{r}{n}\right)} = \\ = [\left(\frac{r}{n}\right)^{\frac{r}{n}} \left(1 - \frac{r}{n}\right)^{\left(1-\frac{r}{n}\right)}]^n = \\ = 2^{n \log_2 [\left(\frac{r}{n}\right)^{\frac{r}{n}} \left(1-\frac{r}{n}\right)^{\left(1-\frac{r}{n}\right)}]} = \\ = 2^{n[\frac{r}{n} \log_2 \frac{r}{n} + \left(1-\frac{r}{n}\right) \log_2 \left(1-\frac{r}{n}\right)]} = 2^{-nh\left(\frac{r}{n}\right)}\end{aligned}$$

So we have

$$1 \geq |B_H(\underline{0}, r)| \frac{1}{2^{nh\left(\frac{r}{n}\right)}} \Rightarrow |B_H(\underline{0}, r)| \leq 2^{nh\left(\frac{r}{n}\right)}$$

□

Theorem 2 *The lower bound on the volume of a Hamming ball, if $r \leq n/2$, is*

$$|B_H(\underline{0}, r)| \geq \frac{1}{n+1} 2^{nh\left(\frac{r}{n}\right)}$$

Proof.

$$P(1) = \frac{r}{n},\ P(0) = 1 - P(1),\ P^n(\{0,1\}^n) = 1.$$

Consider the set of all strings of length n and partition it in the following way.

$$\mathrm{T}_q^n = \{\underline{x} \mid w_H(\underline{x}) = q\},$$

obtaining $n+1$ classes. We know that $|\mathrm{T}_q^n| = \binom{n}{q}$; we are not interested in how many strings are in that set, but what is the total weight; we want to prove

$$|T_r| \geq \frac{1}{n+1} 2^{nh\left(\frac{r}{n}\right)}.$$

There is not symmetry in the weight of the partitions so there is a weight r so that

$$\frac{P^n(\mathrm{T}_q)}{P^n(\mathrm{T}_r)} \leq 1,\ \forall q$$

In the formula above there are binomials we want to bound.

Observation 1

$$\frac{k!}{l!} \leq k^{k-l}.$$

Proof. We are going to prove this observation in two steps.

- $k \geq l$.

$$\frac{k!}{l!} = \frac{k(k-1)\cdots l(l-1)\cdots 1}{l(l-1)\cdots 1} \leq k^{k-l}.$$

- $k < l$.

$$\frac{k!}{l!} = \frac{k(k-1)\cdots 1}{l(l-1)\cdots k(k-1)\cdots 1} \leq \left(\frac{1}{k+1}\right)^{l-k} < \left(\frac{1}{k}\right)^{l-k} = k^{k-l}.$$

□

Define $p = r/n$ so that we have a distribution $P(p, 1-p)$ that picks a set and concentrate the weight (probability) on it. We observe that the probability of each string in a class depends only on the number of 1's in it. So we can write

$$\frac{P^n(\mathrm{T}_q)}{P^n(\mathrm{T}_r)} = \frac{p^q(1-p)^{n-q}|\mathrm{T}_q|}{p^r(1-p)^{n-r}|\mathrm{T}_q|} = p^{q-r}(1-p)^{r-q}\frac{\frac{n!}{q!(n-q)!}}{\frac{n!}{r!(n-r)!}} =$$

$$= p^{q-r}(1-p)^{r-q}\frac{r!}{q!}\frac{(n-r)!}{(n-q)!} \leq p^{q-r}(1-p)^{r-q}r^{r-q}(n-r)^{q-r} =$$

considering that $r = np$ it follows that

$$p^{q-r}(1-p)^{r-q}(np)^{r-q}[n(1-p)]^{q-r} = p^{q-r}(1-p)^{r-q}n^{r-q+q-r}p^{r-q}(1-p)^{q-r} =$$

$$= p^{q-r+r-q}(1-p)^{r-q+q-r} = 1.$$

So we can write

$$1 = P^n(\{0,1\}^n) = P^n\left(\bigcup_{q=0}^{n} T_q\right) = \sum_{q=0}^{n} P^n(T_q) \leq (n+1)\max_q P^n(T_q) =$$

$$= (n+1)P^n(T_r) = (n+1)|T_r|2^{-nh\frac{r}{n}}$$

From the previous result we know $|T_r| \geq \frac{1}{n+1}2^{nh(\frac{r}{n})}$ and $T_r \subseteq B_H(\underline{0}, r)$ so it follows that $|T_r| \leq |B_H(\underline{0}, r)|$. □

So this proof is important because of two reasons:

1. something related to entropy. [LISTEN THE REGISTRATION];
2. Cardinality of the Hamming ball comes up in error correction (a string that has been haltered at most r times is in a certain radius from the original string).

1.5 Lesson 5