C I & S S Lab

UNIVERSITÀ PARTHENOPE
DiST
DIPARTIMENTO DI SCIENZE E TECNOLOGIE

Intelligent Signal Processing

Introduction to Information Theory

Angelo Ciaramella

Introduction

- *Information theory is a branch of applied mathematics and electrical engineering involving the quantication of information*

- *Claude E. Shannon (1948)*
  - Finds fundamental limits on signal processing operations, such as compressing data and reliably storing and communicating data

Information Theory

- What's information?
  - Information is the reduction of uncertainty
  - Some (informal) axioms
    - if something is certain its uncertainty = 0
    - uncertainty should be maximum if all choices are equally probable
    - uncertainty (information) should add for independent sources

Information Theory

- How to measure information content?
  - Let X be a random variable whose outcome x takes values in $\{a_1, \ldots, a_L\}$ with probabilities $\{p_1, \ldots, p_L\}$
  - Shannon's information content for the outcome x = a_i

$$H(x = a_i) = \log_2\left(\frac{1}{P(x = a_i)}\right) = \log_2\left(\frac{1}{p_i}\right)$$

  - Entropy

$$H(X) = \sum_i p_i \log_2\left(\frac{1}{p_i}\right) = -\sum_i p_i \log_2(p_i)$$

sensible measure of expected (average) information content

Information Theory

- Information content
  - How many bits needed to compress your data?
  - Example
    - Observe a sequence «…00000100» with $p_1 = 0.1$ (or $p_0 = 0.9$)

$$H(x=1) = \log_2\left(\frac{1}{0.1}\right) = 3.3 bits$$

$$H(x=0) = \log_2\left(\frac{1}{0.9}\right) = 0.15 bits$$

Information Theory

- Intuition
  - The «1» has less information
    - you don't get too much surprised with a 0
  - You don’t learn too much with a 0
  - The «1» is
    - more improbable
    - more surprising
    - more informative

Information Theory

The entropy of an ensemble

$$H(X) \equiv \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)},$$

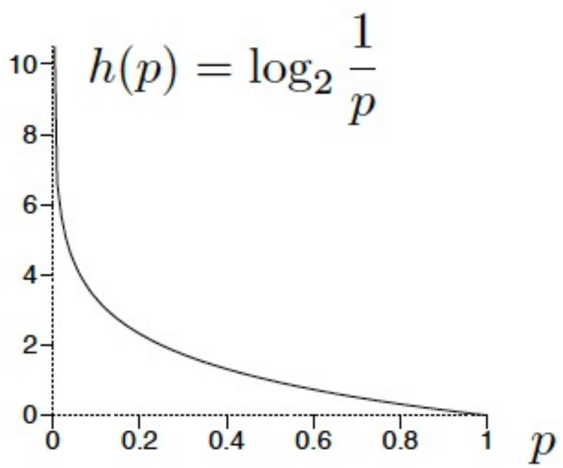
$P(x) = 0$ that $0 \times \log 1/0 \equiv 0$ $\lim_{\theta \to 0^+} \theta \log 1/\theta = 0$

i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
	$\sum_i p_i \log_2 \frac{1}{p_i}$		4.1

Table 2.9. Shannon information contents of the outcomes a–z.

Information and uncertainty

- Consider a binary random variable that can take two values with probabilities *p* and *1 – p*

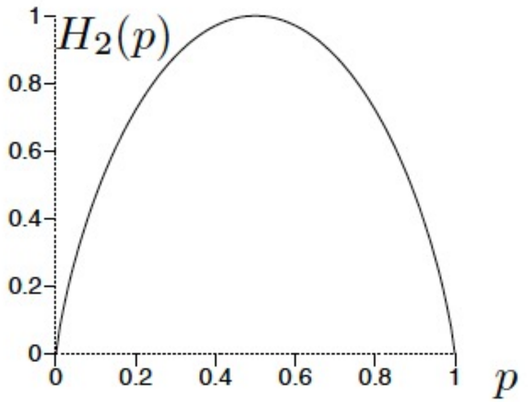


p	$h(p)$	$H_2(p)$
0.001	10.0	0.011
0.01	6.6	0.081
0.1	3.3	0.47
0.2	2.3	0.72
0.5	1.0	1.0

Shannon information content of an outcome with probability *p*, as a function of *p*. The less probable an outcome is, the greater its Shannon information content.

Information and uncertainty

- Consider a binary random variable that can take two values with probabilities *p* and *1 – p*



p	$h(p)$	$H_2(p)$
0.001	10.0	0.011
0.01	6.6	0.081
0.1	3.3	0.47
0.2	2.3	0.72
0.5	1.0	1.0

$$H_2(p) = H(p, 1-p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{(1-p)}$$

Information and uncertainty

- Improbable events are more informative, but less frequent on average

- The entropy satisfies the two first axioms
  - observation of a certain event carries no information
  - maximum information is carried by uniformly probable events

Information under independence

- Variables x and y that are independent

$$P(x, y) = P(x)P(y)$$

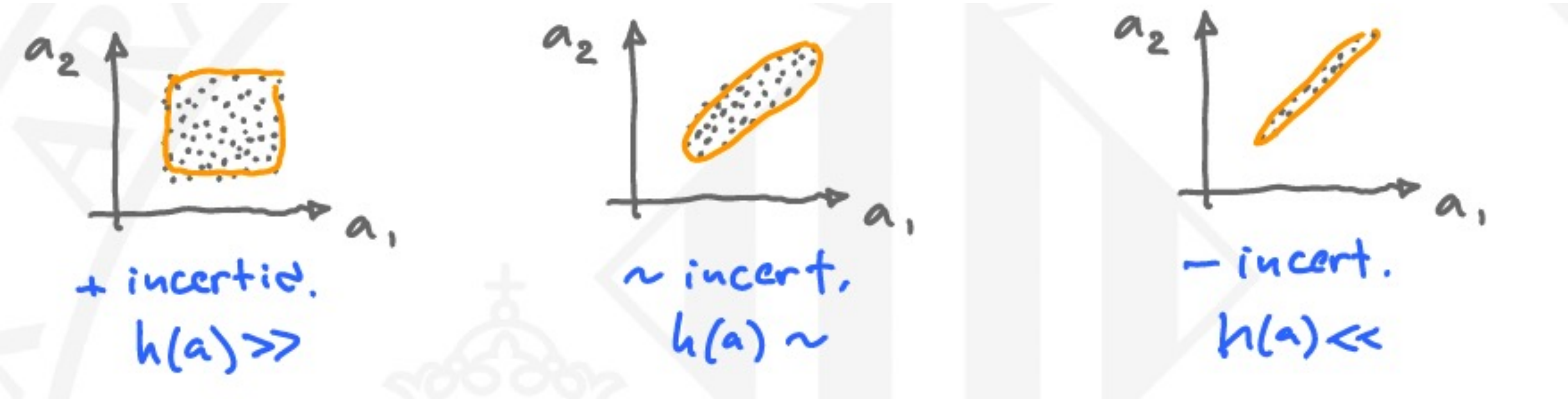
$$h(x, y) = \log \frac{1}{P(x, y)} = \log \frac{1}{P(x)P(y)} = \log \frac{1}{P(x)} + \log \frac{1}{P(y)}$$

$$h(x, y) = h(x) + h(y)$$

- Shannon's information content

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

Differential Entropy



Differential Entropy

- Vecrtor **a** with PDF *P*(**a**)

$$H(\mathbf{a}) = \int P(\mathbf{a}) \log_2\left(\frac{1}{P(\mathbf{a})}\right) d\mathbf{a} = -\int P(\mathbf{a}) \log_2(P(\mathbf{a})) d\mathbf{a}$$

entropy is related to the PDF volume

$$H(\mathbf{a}) = \frac{1}{2}\ln(2\pi e\sigma^2)$$ Unidimensional Gaussian

$$H(\mathbf{a}) = \frac{1}{\log(2)}\ln((2\pi e\sigma)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}})$$ Multidimensional Gaussian

More about Entropy

- ## Joint Entropy

$$H(X,Y) = \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log \frac{1}{P(x,y)}$$

$$H(X,Y) = H(X) + H(Y) \ \text{ iff } \ P(x,y) = P(x)P(y)$$

- ## Conditional Entropy

$$\begin{aligned} H(X \mid Y) &\equiv \sum_{y \in \mathcal{A}_Y} P(y) \left[\sum_{x \in \mathcal{A}_X} P(x \mid y) \log \frac{1}{P(x \mid y)} \right] \\ &= \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log \frac{1}{P(x \mid y)}. \end{aligned}$$

More about Entropy

- Chain rule for information content

$$\log \frac{1}{P(x,y)} = \log \frac{1}{P(x)} + \log \frac{1}{P(y\,|\,x)} \qquad h(x,y) = h(x) + h(y\,|\,x)$$

- Chain rule for entropy

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

More about Entropy

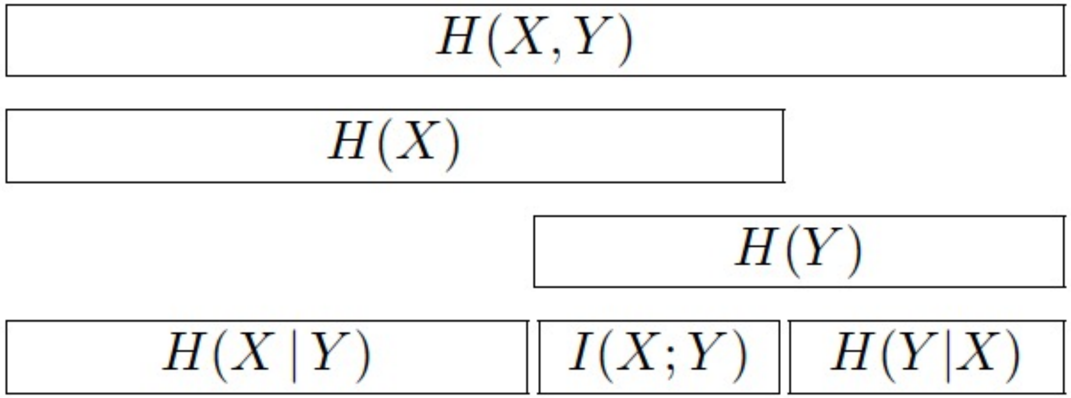
- Mutual Information

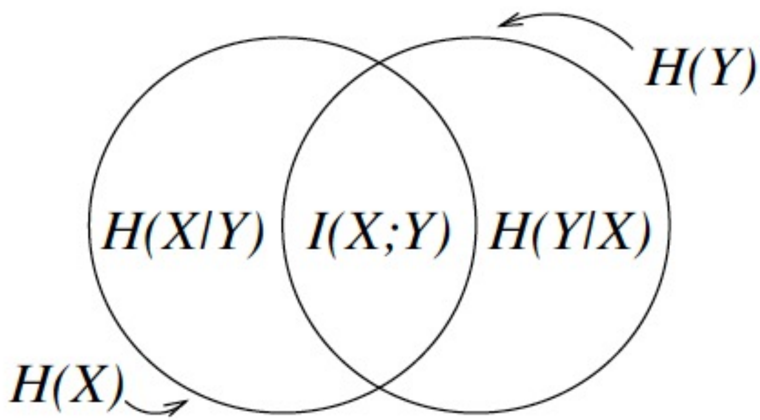
$$I(X;Y) \equiv H(X) - H(X|Y)$$

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

$$I(X;Y) \geq 0$$

More about Entropy





Kullback-Leibler Divergence

- Kind of distance

$$D_{KL}(P(\mathbf{a}), Q(\mathbf{a})) = \int P(\mathbf{a}) \log_2 \left(\frac{P(\mathbf{a})}{Q(\mathbf{a})} \right) d\mathbf{a}$$

$$D_{KL} \geq 0$$

$$D_{KL} = 0 \quad iff \; P(\mathbf{a}) = Q(\mathbf{a})$$

A distance $d(\cdot\|\cdot)$ must fulfil three conditions:

- Positiveness: $d(x\|y) \geq 0$ $d(x\|y) = 0$ iff $x = y$:)
- Triangle inequality: $d(x\|z) \geq d(x\|y) + d(y\|z)$:)
- Symmetry: $d(x\|y) = d(y\|x)$:(

Cross-Entropy

- Two distributions **p** and **q**

$$H(\mathbf{p},\mathbf{q}) = H(\mathbf{p}) + D_{KL}(\mathbf{p}||\mathbf{q})$$

$$D_{KL}(\mathbf{p}||\mathbf{q}) = H(\mathbf{p},\mathbf{q}) - H(\mathbf{p})$$

Cross-Entropy

- Two distributions **p** and **q**

$$H(\mathbf{p},\mathbf{q}) = -\sum_i \mathbf{p}\log_2(\mathbf{q}) = -\sum_i \mathbf{p}\log_2(\frac{\mathbf{pq}}{\mathbf{p}}) =$$

$$-\left[\sum_i (\mathbf{p}\log_2(\mathbf{p}) + \mathbf{p}\log_2(\frac{\mathbf{q}}{\mathbf{p}}))\right] = H(\mathbf{p}) + D_{KL}(\mathbf{p}\|\mathbf{q})$$

Consequence: For discrete **p** and **q** this means:

$$H(\mathbf{p},\mathbf{q}) = -\sum_i \mathbf{p}\log_2(\mathbf{q}) \neq H(\mathbf{q},\mathbf{p}) = -\sum_i \mathbf{q}\log_2(\mathbf{p})$$

More on MI

- Mutual Information

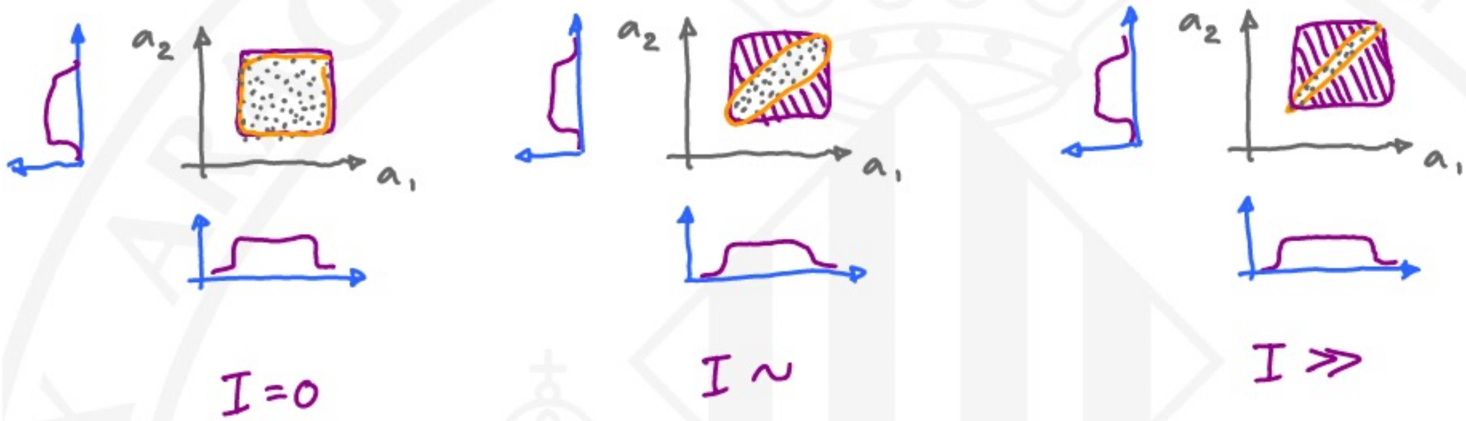
$$I(x,y) = \sum_{x}\sum_{y} p(x,y)\log\left(\frac{p(x,y)}{p_1(x)p_2(x)}\right)$$

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

More on MI

Intuition on mutual information

* Intuición: $I = \sum_x h_x - h$ ≡ diferencia entre el volumen del producto de marginales frente al volumen de la conjunta



Cuanto mayor es la relación entre las variables mayor es la diferencia entre los volumenes (entropías)

More on MI

- Mutual Information

$$I(x,y) = \sum_{x}\sum_{y} p(x,y)\log\left(\frac{p(x,y)}{p_1(x)p_2(x)}\right)$$

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

Data compression

- how many bits are needed to describe the outcome of an experiment

- One way of measuring the information content of a random variable is simply to count the number of possible outcomes

$$|A_X|$$

- binary name to each outcome, the length of each name would be

$$\log_2|A_X| \quad bits \qquad |A_X| \text{ a power 2}$$

Data compression

- Raw bit content

$$H_0(X) = \log_2 |A_X|$$

- Risk
  - δ the probability that there will be no name for an outcome x
- Compression strategy with risk δ
  - Smallest sufficient subset

$$P(x \in S_\delta) \geq 1 - \delta$$

- can be constructed by ranking the elements of A_X in order of decreasing probability and adding successive elements starting from the most probable elements until the total probability is greater than (1 - δ).

Data compression

$$\mathcal{A}_X = \{\, \mathtt{a},\ \mathtt{b},\ \mathtt{c},\ \mathtt{d},\ \mathtt{e},\ \mathtt{f},\ \mathtt{g},\ \mathtt{h}\, \},$$
$$\mathcal{P}_X = \{\, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{3}{16}, \tfrac{1}{64}, \tfrac{1}{64}, \tfrac{1}{64}, \tfrac{1}{64}\, \}.$$

The raw bit content of this ensemble is 3 bits, corresponding to 8 binary names

$$P(x \in \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}) = 15/16$$

$\delta = 0$	
x	$c(x)$
a	000
b	001
c	010
d	011
e	100
f	101
g	110
h	111

$\delta = 1/16$	
x	$c(x)$
a	00
b	01
c	10
d	11
e	−
f	−
g	−
h	−

Source Coding Theorem

- Essential bit content of X is

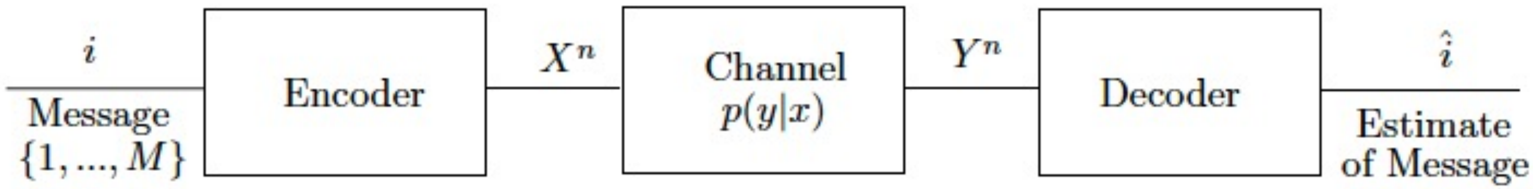
$$H_\delta(X) = \log_2 |S_\delta|$$

- Shannon's Source Coding Theorem
  - Let X be an ensemble with entropy $H(X) = H$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$

$$\left| \frac{1}{N} H_\delta(X^N) - H \right| < \epsilon$$

$$(X_1, X_2, \ldots, X_N)$$

Communication channel



Message is the index set from which a message is drawn

Discrete Memoryless Channel

- Discrete Memoryless Channel (DMC)
  - consists of two finite sets *X* and *Y* and a collection of probability mass functions p(y|x)

$$(X, p(y|x), Y)$$

- (*M*, *n*) code for the channel (*X*, p(*y*|*x*),*Y*)
  - encoding function $g: \{1: M\} \rightarrow X^n$, which is a mapping from the index set to a set of codewords or codebook
  - decoding function $f: Y^n \rightarrow \{1: M\}$, which is a deterministic rule assigning a number (index) to each received vector

Channel Coding Theorem

- Channel Coding Theorem
  - Let us dene the channel capacity as follows

$$C = \max_{p_X(x)} I(X;Y)$$

for a discrete memoryless channel a rate R is achievable if and only if R < C

Channel Coding Theorem

- Channel Coding Theorem
  - even though the channel introduce errors, the information can still be reliably sent over the channel at all rates up to channel capacity
    - the noisiness of the channel does not limit the reliability of the transmission but only its rate
  - Shannon's key idea
    - sequentially use the channel many times, so that the law of large number comes into effect
  - Shannon's outline of the proof is indeed strongly based on the concept of typical sequences and in particular on a joint typicality based decoding rule
  - Shannon proves that choosing the codes at random is asymptotically the best choice whatever the channel is
    - for finite n the knowledge of the channel may help to choose a better code