Information et quantité d'information

Information et quantité d'information

Définition Une information désigne un ou plusieurs événements possibles parmi un ensemble fini d'événements.

L'information permet de diminuer l'incertitude.

Exemple Considérons par exemple une source qui peut produire trois symboles a, b et c. Quand le destinataire attend un symbole, il est dans l'incertitude quant au symbole que la source va engendrer. Lorsque le symbole apparaît et qu'il arrive au destinataire, cette incertitude diminue.

Le but de la théorie de l'information est de mesurer cette incertitude avant réception.

Exemple

On recherche une lettre dans une boîte. Si on précise que la lettre se trouve dans une enveloppe **bleue**, on fournit une information qui diminuera le temps de recherche du fait que le nombre de lettres dans des enveloppes bleues est plus restreint.

Si on ajoute l'information que la lettre dans une **grande** enveloppe, on pourra abréger d'autant plus le temps de la recherche.

Définition

La quantité d'information est définie par

$$\log_2 \left(\frac{N}{n} \right)$$

- N est le nombre d'événements possibles
- n est le cardinal du sous-ensemble dénoté par l'information

qui est exprimé en *logon*.

Observation La quantité d'information est une fonction croissante.

Exemple

Dans une boîte il y a $N = 1050$ lettres dont

- $n_1 = 500$ lettres en enveloppes bleues
- $n_2 = 250$ en grandes enveloppes
- $n = 40$ en grandes enveloppes bleues

L'information :

- *la lettres est dans une enveloppe dans la boîte* vaut $\log_2\left(\frac{N}{N}\right) = \log_2(1) = 0$
- *la lettres est dans une enveloppe bleue* vaut $\log_2\left(\frac{N}{n_1}\right) = \log_2(\frac{1050}{500}) = 1,07$
- *la lettres est dans une grande enveloppe* vaut $\log_2\left(\frac{N}{n_2}\right) = \log_2(\frac{1050}{250}) = 2,07$
- *la lettres est dans une grande enveloppe bleue* vaut $\log_2\left(\frac{N}{n_3}\right) = \log_2(\frac{1050}{40}) = 3,64$

Entropie

L'entropie nous permet de mesurer la quantité moyenne d'information contenue dans un ensemble de messages et de mesurer l'incertitude.

Soit X un ensemble partitionné en n sous-ensembles X_i, $1 \leq i <\leq$ (les messages), avec

$$X = \bigcup_{i=1}^{i=n} X_i$$

Par définition, la quantité d'information liée à chaque message de X_i est

$$I(X_i) = \log_2 \left(\frac{|X|}{|X_i|} \right) = \log_2 \left(\frac{N}{n_i} \right)$$

Définition L'*entropie* de la partition est

$$H(partition) = \sum_{i=1}^{i=n} \frac{n_i}{N} \log_2 \left(\frac{N}{n_i} \right)$$

Observation Si

$$p_i = \frac{N}{n_i}$$

est la probabilité de l'apparition d'un message en X_i

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

Du fait que les X_i forment une partition de X, $\sum_{i=1}^{i=n} p_i = 1$ et l'entropie correspond à la distribution de probabilité de tous les messages possibles.

Exemple Soit une urne content 100 boules dont x blanches et $100 - x$ boules noires. On considère l'expérience qui consiste à tirer une boule.

- $I(b)$=quantité d'information liée à l'apparition d'une boule blanche
- $I(n)$=quantité d'information liée à l'apparition d'une boule noire
- H=quantite d'information moyenne par expérience = l'entropie

$$H = \frac{x}{100} \log_2 \left(\frac{100}{x} \right) + \frac{N-x}{100} \log_2 \left(\frac{100}{N-x} \right)$$

x	$N-x$	$I(b)$	$I(n)$	H
50	50	1	1	1
40	60	1,32	0,73	0,97
1	99	6,64	0,014	0,080

Codage pour un canal non bruité

Afin de transmettre un message sous la forme de signal, il faut le coder; c'est ce que l'on nomme également *codage de source*.

- écriture
- parole
- code Morse
- code ASCII
- unicode
- codes binaires i.e. sur l'alphabet $\{0, 1\}$

Soit Σ un alphabet et Σ^* l'ensemble des mots finis sur cet alphabet. Un code $C = \{c_1, ..., c_k\}$ est un sous-ensemble de Σ^*. Les éléments $c_i \in C$ sont appelés les *mots du code*.

Si

- tous les mots du code sont de même longueur, on dit que C est un code *à longueur fixe* ou code *en bloc*
- dans le cas contraire, C est un code à *longueur variable*
- si aucun mot du code n'est le préfixe d'un autre C est appelé *préfixe*

Exemple

- $C = \{0, 10, 11\}$ code préfixe de longueur variable
- $C = \{1, 01, 11\}$ code non-préfixe de longueur variable
- $C = \{000, 101\}$ code de longueur fixe

Codes optimaux

En associant des codes courts aux messages les plus fréquents et des codes longs aux messages les moins fréquents, on peut construire un code optimal : au sens où le nombre moyen de bits par symbole correspond précisément à l'entropie de l'ensemble des messages possibles.

lettre	fréquence en français	fréquence en anglais	alphabet de Morse
a	6.16	8.05	·–
b	0.40	1.62	– ···
c	5.35	3.20	– · –·
d	3.86	3.65	– ··
e	18.61	12.31	·
f	2.24	2.28	·· –·
g	1.79	1.61	– – ·
h	1.48	5.14	····
i	6.35	7.18	··
j	0.04	0.10	· – ––
k	0.13	0.52	– · –

lettre	fréquence en français	fréquence en anglais	alphabet de Morse
l	5.26	4.03	· − ··
m	1.79	2.25	−−
n	6.02	7.19	−·
o	5.12	7.94	− − −
p	2.92	2.28	· − −·
q	0.62	2.29	− − ·−
r	5.35	6.03	· − ·
s	6.96	6.59	· · ·
t	7.41	9.59	−
u	5.03	3.10	· · −
v	1.03	0.93	··
w	0.20	0.25	− ··−
y	1.39	1.88	− · −−
z	0.04	0.09	− − ··

Huffman a proposé un algorithme qui construit un code préfixe optimal

Longueur moyenne, efficacité et redondance

Soit $A = \{a_1, a_2, \ldots, a_k\}$ un alphabet de source tel que

- p_i est la probabilité d'apparition du symbole a_i
- $a_i \rightarrow c_i$ est un code où c_i est un mot de longueur ℓ_i,

Définition La *longueur moyenne* du code est défini comme:

$$L = \sum_{i=1}^{k} p_i \ell_i$$

qui correspond à la somme pondérée des longueurs de tous les mots.

La longueur moyenne coïncide avec le rapport entre le nombre de symboles binaires du message codé et le nombre de symboles de source quand le message est suffisament long pour que tous les symboles apparaissent avec une fréquence relative égale à leur probabilité.

Exemple

alphabet source	code	probabilité
a	0	$\frac{1}{2}$
b	10	$\frac{1}{4}$
c	11	$\frac{1}{4}$

- La longueur moyenne est
$L = \sum_{i=1}^{3} p_i \ell_i = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = 1.5$

- L'entropie est $H = \sum_{i=1}^{k} p_i \log_2(\frac{1}{p_i}) =$
$\frac{1}{2} \cdot \log_2(\frac{1}{2}) + \frac{1}{4} \cdot \log_2(\frac{1}{4}) + \frac{1}{4} \cdot \log_2(\frac{1}{4}) = 1.5$

Efficacité et redondance

Définition On définit l'*efficacité* comme le rapport

$$\eta = \frac{H}{L}$$

et la *redondance*

$$r = 1 - \eta$$

Exemple Soit l'alphabet de source $A = \{a, b\}$ tel que

alphabet source =symbole	probabilité	code
a	$p_a = 0,8$	0
b	$p_b = 0,2$	1

- L'entropie de la source est de

$$-0,8 \cdot \log_2(0,8) - 0,2 \cdot \log_2(0,2) = 0,72 \; logon$$

- et sa longueur moyenne est

$$0,8 \cdot 1 + 0,2 \cdot 1 = 1$$

- avec l'efficacité de $\eta = 0.72$ et la redondance de $r = 0.28$

Comment peut-on diminuer cette redondance et améliorer l'efficacité? Une idée est de coder des couples de symboles au lieu des symboles eux-mêmes.

symbole	probabilité	code
$aa = \alpha$	$p_{aa} = 0,64$	0
$ab = \beta$	$p_{ab} = 0,16$	11
$ba = \gamma$	$p_{ba} = 0,16$	100
$bb = \delta$	$p_{bb} = 0,04$	101

- L'entropie de la source est de

$$0,64 \cdot log_2 \left(\frac{1}{0,64} \right) +$$

$$2 \cdot \left(0,16 \cdot log_2 \left(\frac{1}{0,16} \right) \right) +$$

$$0,04 \cdot log_2 \left(\frac{1}{0,04} \right) = 1,45 \; logon$$

- et sa longueur moyenne est

$$1 \cdot 0,64 + 2 \cdot 0,16 + 3 \cdot (0,16 + 0,04) = 1,56$$

- avec l'efficacité de $\eta = \frac{1,45}{1,56} = 0.93$ et la redondance de $r = 0.07$

En revanche, le coût à payer est une complexification des opérations de codage et de décodage. On montre que, en faisant croître le nombre de symboles qu'on code, l'efficacité du codage peut devenir aussi proche que possible de sa limite supérieure (égale à 1). C'est précisément la signification du premier théorème de Shannon.

Arbre de Huffman

Théorèmes de Shannon 1 Théorème du codage de source: Sans perturbation, il est possible, à partir d'un alphabet quelconque, de coder les messages émis de telle sorte que le rendement soit aussi proche que souhaité de la valeur maximale, i.e. la capacité du canal.

Théorie algorithmique de l'information

ou théorie de la complexité de Kolmogorov née dans la décennie 1960, à la frontière de la logique mathématique et de l'informatique théorique.

- Kolmogorov
- Chaitin
- Solomonof

L'information d'une suite de caractères est définie comme la taille du plus petit programme permettant de l'engendrer.

La distance informationnelle entre deux suites A et B est définie comme la taille du plus court programme permettant de transformer A en B et B en A