La formule de Shannon

Gérard LAVAU, Lycée Carnot à Dijon

Dans le n° 96 de *Feuille de Vigne*, Michel Lafond nous exposait l'algorithme de compression de Huffman, et nous précisait que cet algorithme permettait de compresser d'environ 15% un texte "ordinaire". Nous nous proposons d'expliquer en quoi la formule de Shannon permet de préciser ce point.

Nous donnons d'abord la formule de Shannon, l'appliquons sur quelques exemples, puis tenterons d'en donner une justification empirique que nous espérons assez convaincante.

Considérons une suite de lettres ou de symboles constituant un message. Chaque symbole peut prendre les valeurs s_1, s_2, …, s_k (ce sont par exemple les lettres de l'alphabet), avec des probabilités respectivement p_1, p_2, …, p_k et chaque symbole est supposé être indépendant du suivant (ce qui n'est pas exact pour un texte littéraire, mais cela simplifie les choses). Shannon définit la quantité d'information contenue dans un symbole comme étant $H = p_1 \log \frac{1}{p_1} + p_2 \log \frac{1}{p_2} + \ldots + p_k \log \frac{1}{p_k}$, où log est le logarithme en base 2. Cette quantité d'information est maximale lorsqu'elle est obtenue dans le cas de l'équirépartition et vaut alors $\log(k)$. Cela résulte d'une inégalité de convexité, en utilisant le fait que log est concave :

$$\sum_{i=1}^{k} p_i \log \frac{1}{p_i} \leq \log\left(\sum_{i=1}^{k} p_i \frac{1}{p_i}\right) = \log(k).$$

Enfin, Shannon énonce qu'un message constitué de N symboles peut en moyenne être compressé en un message de longueur aussi proche que l'on veut de $\frac{HN}{\log(k)}$, quantité qui donne donc la longueur moyenne optimale d'un message compressé.

EXEMPLE 1

Reprenons la répartition des 32 lettres et autres symboles, telle qu'elle est donnée dans l'article de Michel Lafond.

s_i	p_i	s_i	p_i	s_i	p_i
E	0,144	D	0,033	*virgule*	0,007
blanc	0,128	C	0,026	*tiret*	0,007
S	0,062	M	0,024	H	0,007
A	0,062	P	0,021	J	0,007
N	0,062	V	0,013	K	0,007
I	0,058	*apostrophe*	0,013	X	0,006
R	0,057	*point*	0,008	Y	0,006
T	0,057	Q	0,008	*interrogation*	0,006
U	0,049	B	0,007	Z	0,005
O	0,048	F	0,007	W	0,005
L	0,043	G	0,007		

Si les 32 symboles s_i étaient équirépartis, alors la quantité d'information donnée par un symbole serait $H = \sum_{i=1}^{32} \frac{1}{32} \log(32) = \log(32) = 5$.

Ce nombre correspond exactement aux 5 chiffres binaires (de 00000 à 11111) qui seraient nécessaires pour coder chacun des 32 symboles, chaque chiffre binaire apportant une quantité d'information d'une unité.

Mais les symboles ne sont pas équirépartis, et si on applique la formule de Shannon à la répartition ci-dessus, on trouvera : $H = \sum_{i=1}^{32} p_i \log \frac{1}{p_i} = 4{,}29$ soit un peu moins de 5. Un message constitué de N symboles comportera une quantité d'information de 4,29N au lieu de 5N. Shannon prévoit qu'on peut compresser le message en utilisant environ $\frac{4{,}29N}{5}$ symboles au lieu de N, soit une réduction de $\frac{0{,}71}{5} = 0{,}142$ proche des 15% annoncés.

EXEMPLE 2

On ne considère plus que deux symboles. Chaque symbole peut prendre deux valeurs s_1 et s_2 avec des probabilités respectivement $p_1 = 0{,}8$ et $p_2 = 0{,}2$. La quantité d'information contenue dans un symbole est $p_1 \log \frac{1}{p_1} + p_2 \log \frac{1}{p_2} \approx 0{,}72$. Un message de N symboles contient en moyenne une quantité d'information égale à 0,72N. Sa quantité d'information est donc inférieure à sa longueur, ce qui est une perte. Il est possible de coder le message de façon à ce qu'il ait en moyenne une longueur aussi proche que l'on veut de 0,72N. En appliquant le codage d'Huffman sur des groupes de symboles, on obtient par exemple :

- ***Codage 1***

On regroupe les symboles deux par deux et l'on code :

s_1s_1	par 0	(avec une probabilité p_1^2)
s_1s_2	par 10	(avec une probabilité p_1p_2)
s_2s_1	par 110	(avec une probabilité p_2p_1)
s_2s_2	par 111	(avec une probabilité p_2^2)

On vérifiera que le décodage est sans ambiguïté. La longueur moyenne du message est :

$\frac{N}{2}(p_1^2 + 2p_1p_2 + 3p_1p_2 + 3p_2^2) = 0{,}78N$ au lieu de N

- ***Codage 2***

On regroupe les symboles trois par trois et on les code comme suit :

$s_1s_1s_1$	0	(avec une probabilité p_1^3)
$s_1s_1s_2$	100	(avec une probabilité $p_1^2p_2$)
$s_1s_2s_1$	101	(avec une probabilité $p_1p_2p_1$)
$s_2s_1s_1$	110	(avec une probabilité $p_2p_1^2$)
$s_1s_2s_2$	11100	(avec une probabilité $p_1p_2^2$)
$s_2s_1s_2$	11101	(avec une probabilité $p_2p_1p_2$)
$s_2s_2s_1$	11110	(avec une probabilité $p_2^2p_1$)
$s_2s_2s_2$	11111	(avec une probabilité p_2^3)

La longueur moyenne du message est : $\frac{N}{3}(p_1^3 + 3 \times 3p_1^2p_2 + 3 \times 5p_1p_2^2 + 5p_2^3) = 0{,}728N$

ce qui est encore meilleur et quasiment optimal.
Mais d'où vient la formule de Shannon ?

Quantité d'information

Considérons N boîtes numérotées de 1 à N. Un individu A a caché au hasard un objet dans une de ces boîtes. Un individu B doit trouver le numéro de la boîte où est caché l'objet. Pour cela, il a le droit de poser des questions à l'individu A auxquelles celui-ci doit répondre sans mentir par OUI ou NON. Mais chaque question posée représente un coût à payer par l'individu B (par exemple un euro). Un individu C sait dans quelle boîte est caché l'objet. Il a la possibilité de vendre cette information à l'individu B. B n'acceptera ce marché que si le prix de C est inférieur ou égal au coût moyen que B devrait dépenser pour trouver la boîte en posant des questions à A. L'information détenue par C a donc un certain prix. Ce prix représente la quantité d'information représentée par la connaissance de la bonne boîte : c'est le nombre moyen de questions à poser pour identifier cette boîte. Comme plus haut, nous la noterons H.

Exemples :
Si $N = 1$, $H = 0$. Il n'y a qu'une seule boîte. Aucune question n'est nécessaire.
Si $N = 2$, $H = 1$. On demande si la bonne boîte est la boîte n°1. La réponse OUI ou NON détermine alors sans ambiguïté quelle est la boîte cherchée.
Si $N = 4$, $H = 2$. On demande si la boîte porte le n°1 ou 2. La réponse permet alors d'éliminer deux des boîtes et il suffit d'une dernière question pour trouver quelle est la bonne boîte par deux.
Si $N = 2^n$, $H = n$. On écrit les numéros des boîtes en base 2. Les numéros ont n chiffres binaires (de 00…0 à 11…1), et pour chaque rang de ces chiffres, on demande si le numéro de la boîte cherchée possède à ce rang le chiffre 0 ou le chiffre 1. En n questions, on a déterminé tous les chiffres binaires de la bonne boîte.
On est donc amené à poser $H = \log(N)$ questions, mais cette configuration ne se produit que dans le cas de N événements équiprobables.

Formule de Shannon

Supposons maintenant que les boîtes soient colorées, et qu'il y ait n_1 boîtes rouges. Supposons également que C sache que la boîte où est caché l'objet est rouge. Quel est le prix de cette information ? Sans cette information, le prix à payer est $\log(N)$. Muni de cette information, le prix à payer n'est plus que $\log(n_1)$. Le prix de l'information "*la boîte cherchée est rouge*" est donc $\log(N) - \log(n_1) = \log \frac{N}{n_1}$.

Supposons maintenant que les boîtes soient de diverses couleurs : n_1 boîtes de couleur C_1, n_2 boîtes de couleur C_2, …, n_k boîtes de couleurs C_k, avec $n_1 + n_2 + \ldots + n_k = N$. La personne C sait de quelle couleur est la boîte cherchée. Quel est le prix de cette information ? L'information "*la boîte est de couleur* C_1" vaut $\log \frac{N}{n_1}$, et cette éventualité a une probabilité $\frac{n_1}{N}$. L'information "*la boîte est de couleur* C_2" vaut $\log \frac{N}{n_2}$, et cette éventualité a une probabilité $\frac{n_2}{N}$, etc. Le prix moyen de l'information est donc $H = \frac{n_1}{N} \log \frac{N}{n_1} + \frac{n_2}{N} \log \frac{N}{n_2} + \ldots + \frac{n_k}{N} \log \frac{N}{n_k}$. Plus généralement, si on considère k événements disjoints de probabilités respectives
$p_1, p_2, \ldots, p_k$ avec $p_1 + p_2 + \ldots + p_k = 1$, alors la quantité d'information correspondant à cette distribution de probabilité est $p_1 \log \frac{1}{p_1} + \ldots + p_k \log \frac{1}{p_k}$.