Entropie de Shannon, théorie de l'information, et compression de données

En 1948, tandis qu'il travaillait aux Laboratoires Bell, l'ingénieur en génie électrique Claude Shannon formalisa mathématiquement la nature statistique de "l'information manquante" dans les signaux des lignes téléphoniques. Pour ce faire, il développa le concept général d'entropie et d'information. Dans ce tutorat, nous allons étudier cette entropie (qui se trouve être la même, à une constante multiplicative près, que celle de Boltzmann) et ses liens avec la compression de données.

1. Entropie et probabilité : particules dans des boites

Soit N particules et k boites, considéreons l'experience suivante : on prend la premiere particule et on la jette dans l'une des boites, de façon équiprobable. On fait ensuite la même chose avec le seconde, puis la troisième, etc.... A la fin, les particules ont toutes été jetées dans les N boites de façon équiprobable (chacune des boites est associée à une énergie $E_1, E_2, ..., E_k$).

a) Donnez le nombre $\mathcal{N}_N(N_1, N_2, \ldots, N_k)$ d'experiences qui conduisent à une configuration dans laquelle la j-ème boite a N_j particules. Donnez le nombre total d'experiences. En déduire la probabilité de chaque configuration.

b) Considérez la limite "thermodynamique" $N \to \infty$. On note le nombre d'occupation du i-ème état par $n_i = N_i/N$. Montez que l'on peut écrire $\mathcal{N}_N(Nn_1, \ldots, Nn_k) \simeq e^{N\Sigma(n_1,\ldots,n_k)}$, et donnez la fonction $\Sigma(n_1, \ldots, n_k)$. (Utilisez la formule de Stirling). Que vous évoque cette fonction ?

c) Question subsidiaire : En utilisant le formalisme des multiplicateur de Lagrange, calculez les nombres d'occupation n_i les plus probables sachant que (i) le nombre de particules $\sum_{i=1}^{k} N_i = N$ est fixé et (ii) l'énergie $\sum_{i=1}^{k} E_i n_i = \epsilon$ est aussi fixée. On utilisera la fonction Σ.

2. Surprise, incertitude et information

Passons maintenant à la théorie de l'information. Considérez une variable aléatoire X qui peut prendre une valeur parmi N possibles dans un alphabet χ. La probabilité de prendre une valeur $X = x_i \in \chi$ est p_i. Pour cet exercice, considérons par exemple un alphabet de quatre lettres $\chi = \{A, B, C, D\}$ avec les probabilités $p_A = 1/2$, $p_B = 1/4$, $p_C = 1/8$, $p_D = 1/8$.

a) Nous tirons une valeur $X = x_i$. Plus cette valeur est improbable, plus nous sommes surpris de la voir apparaître. Définissons la "surprise" comme $\log_2 1/p_i$. Le fait que nous soyons surpris est clairement lié au fait que nous ne sommes pas certain du résultat du tirage avant qu'il n'ait lieu. Appelons, avec Shannon, la surprise moyenne "l'incertitude". Comment s'écrit l'incertitude en général ? Que vaut-elle dans notre exemple ?

b) Une autre approche est la suivante : je mesure une valeur $X = x_i$, mais je la garde secrète en vous laissant le soin de la deviner. Vous ne connaissez que les probabilités p_i, ce qui peut vous donner quand même une idée sur la valeur en question. Plus x_i est probable, plus petite est l'information que je vous donnerais en vous disant ce qu'est vraiment la valeur mesurée x_i. Si l'on défini l'information qu'il vous manque comme $\log_2 1/p_i$, quelle est la valeur moyenne de cette information manquante en général ? Que vaut-elle dans notre exemple ?

3. Entropie de Shannon et questions OUI/NON

Nous voulons maintenant répondre à la question : à quel point est-il difficile de deviner la valeur d'une variable aléatoire ?

a) J'ai tiré au hasard l'une des lettres A,B,C,D avec les probabilités précédentes. Vous avez le droit de poser des questions auquel je répondrai par oui ou par non, du type "Cette lettre est-elle dans l'ensemble $\{x_1, x_2, \ldots\}$?" Votre but est de trouver le plus vite possible la lettre tirée. Combien de questions devez vous poser en moyenne avant de découvrir la lettre que j'ai tirée ? Comparez ce nombre avec l'entropie de l'alphabet.

b) Considérons un nouvel alphabet de 2^b lettres équiprobables (avec b entier). Combien des questions est-il nécessaire de poser en moyenne ? Comparez avec l'entropie du système.

c) (Théorème de Shannon (1948)) Considérons enfin un alphabet général avec A lettres x_i, $i = 1 \ldots A$ avec probabilités p_i. Comment généraliser la stratégie précédente ? Combien des questions devez vous poser en moyenne pour découvrir la lettre tirée ? [1]

Initialement, il ne semble pas que Shannon ait été au courant de la relation entre sa nouvelle mesure et les travaux précédents en thermodynamique. En 1949, tandis qu'il travaillait à ses équations depuis un moment, il rendit visite au grand mathématicien John von Neuman qui lui dit : "La théorie est excellente mais elle a besoin d'un bon nom pour "information perdue". Pourquoi ne l'appelles-tu pas entropie ? Premièrement, un développement mathématique ressemblant fort au tien existe déjà dans la mécanique statistique de Boltzmann, et deuxièmement, personne ne comprend vraiment bien l'entropie, donc dans une discussion tu te trouverais dans une position avantageuse."

4. Compression de données

Vous avez un fichier avec N valeurs aléatoires $\{x_i\}$, $i = 1 \ldots A$. Ce fichier prend une place de $N \log_2 A$ bits dans votre disque dur.

a) Adoptons la stratégie suivante, en utilisant les questions OUI/NON. On écrit un fichier ou l'on garde les réponses aux question de l'exemple précédent avec OUI= 1 et NON=0. Quelle est la taille de ce nouveau fichier ? Quel est le facteur de compression (rapport entre la taille du fichier avant et après compression) ?

b) Comparez ce résultat avec les performances de votre programme de compression favori (zip, rar ...). Pour cela, utiliser les fichiers ci-joints, qui sont des suites de 10^4 caractères Ascii [2] tirés au hasard avec les alphabet suivant :

(i) *uniform.txt* : caractères tirés au hasard parmi L'ENSEMBLE des 256 caractères possibles.

(ii) *half.txt* : caractères tirés au hasard parmi LA MOITIE (soit 128) caractères possibles.

(iii) *abcd.txt* : caractères tirés au hasard parmi ABCD.

(iv) *abcd2.txt* : caractères tirés au hasard parmi ABCD avec les probabilités de l'exercice 2.

5. L'entropie de l'anglais

a) Les fréquences des lettres dans l'anglais sont disponibles sur : *http ://en.wikipedia.org/wiki/Letter_frequencies*. Calculez le rapport de compression de l'anglais (négligez les espaces, majuscules et autres caractères spéciaux.)

b) Comparez vos résultats aux meilleurs compresseurs disponibles : *http ://en.wikipedia.org/wiki/Hutter_Prize*. D'où vient la différence observée ?

c) Estimez (et commentez vos résultats) l'entropie de l'anglais grace à l'experience numérique : *http ://www.math.psu.edu/dlittle/java/informationtheory/entropy/index.html*

---

[1] On pourra commencer par se demander combien de questions faut-il poser pour découvrir une lettre x_i donnée et ensuite faire la moyenne sur toutes ces lettres.

[2] On rappelle que les caractères Ascii sont codés sur 8 bits et prennent 256 valeurs possibles, voir *http ://fr.wikipedia.org/wiki/ASCII*